Revisiting Dense Retrieval with Unaswerable Counterfactuals

Anonymous authors
Paper under double-blind review

Abstract

The retriever-reader framework is popular for open-domain question answering (ODQA), where a retriever samples for the reader a set of relevant candidate passages from a large corpus. A key assumption behind this method is that high relevance score from the retriever likely indicates high answerability from the reader, which implies a high probability that the retrieved passages contain answers to a given question. In this work, we empirically dispel this belief and observe that recent dense retrieval models based on DPR often rank unanswerable counterfactual passages higher than their answerable original passages. To address such answer-unawareness in dense retrievers, we seek to use counterfactual samples as additional training resources to better synchronize the relevance measurement of DPR with the answerability of question-passage pairs. Specifically, we present counterfactually-**Pi**voting **C**ontrastive **L**earning (PiCL), a novel representation learning approach for passage retrieval that leverages counterfactual samples as pivots between positive and negative samples in their learned embedding space. We incorporate PiCL into the retriever training to show the effectiveness of PiCL on ODQA benchmarks and the robustness of the learned models.

1 Introduction

Open-domain question answering (ODQA) (Chen & Yih, 2020) is a task that finds the answers to natural language questions from a large collection of documents. A common approach to ODQA tasks is a two-stage retriever-reader framework (Chen et al., 2017), where a retriever roughly selects relevant candidate passages from which a reader extracts the answers to a question. For the first-stage retrieval, recent studies leverage dense passage retriever (DPR) (Karpukhin et al., 2020), which computes relevance scores based on the similarity between learned representations of questions and passages. Generally, it is assumed that high relevance scores from the retriever naturally lead to high answerability from the reader, which implies that the retrieved passages are more likely to contain the correct answer to a given question. Based on such naive belief, many research efforts for ODQA have focused on solely improving the ranking performance of the retrievers to further enhance the performance of the subsequent reader (Karpukhin et al., 2020; Xiong et al., 2021; Qu et al., 2021).

In this work, our counterfactual simulation dispels this myth. Specifically, we present a simple counterfactual sampling strategy that removes the answer span of a question from its positive passage. We observe that neural retrieval models based on DPR often rank such unanswerable counterfactual passages similarly to or higher than their answerable original passages. We hypothesize that dense retrievers are not answer-aware, which poses the asynchronicity between their relevance scores and answerability as a new challenge of the current retriever-reader framework.

Motivated by this, we seek to repurpose the counterfactual passages as additional training resources to better synchronize the relevance measurement of DPR with the answerability of retrieved passages. A straightforward implementation is to consider these counterfactual samples as hard negatives to supplement in-batch negatives (Zhan et al., 2021). However, the semantic overlap between the counterfactual and source-original passages makes such naive application ineffective, as training a retriever with minimally different negative samples (*i.e.*, answer-removed counterfactuals) potentially hurts the ability to capture semantic relevance of positive question-passage pairs.

To overcome this challenge, we propose *counterfactually-Pivoting Contrastive Learning* (PiCL), a novel contrastive representation learning scheme for improving DPR. By PiCL, our ultimate goal is to make DPR be aware of both relevance and answerability on its dense embedding space, *i.e.*, higher relevance leads to higher answerability, and vice versa. For that, the key idea of PiCL is to use the counterfactual samples as not only hard negatives but also pseudo positives, which can be a pivot between the original positive and negative passages. To effectively learn with such multiple views, PiCL aims to optimize the following three loss functions of learning the dense embeddings of question and original/counterfactual passages: (1) modified DPR loss of mapping the positive passage closer to the question than negatives (including counterfactuals) in a batch, (2) Counterfactuals-as-hard-negatives loss of mapping the original positive passage closer to the question than counterfactuals, and (3) Counterfactuals-as-pseudo-positives loss of mapping the counterfactuals closer to the question than in-batch negative passages.

Our contributions are twofold. First, we demonstrate that the neural retrieval models suffer from the unawareness of answerability, which can be a bottleneck of the retriever-reader pipeline for ODQA. Second, we design a novel training framework named PiCL of learning to synchronize the relevance and answerability. Our extensive experiments validate the effectiveness of PiCL in diverse ODQA scenarios, which are orthogonally applicable to well-studied prior techniques on DPR.

2 Preliminaries and Related work

2.1 Dense Passage Retrieval for ODQA

Open-domain question answering is a knowledge-intensive task that aims at answering factoid questions given a large corpus of texts (Chen & Yih, 2020). We particularly focus on a setting in which the corpus C is a collection of M passages p, *i.e.* $C = \{p_i\}_{i=1}^{M}$. For each question-passage pair (q_i, p_i^+) where q_i is **answerable** from p_i^+, the passage p_i^+ can be viewed as a concatenation of non-answer spans s_l and s_r and an answer span a_i, *i.e.* $p_i^+ = [s_l; a_i; s_r]$. While s_l and s_r provide some relevant contexts to the question, the **answerability** of q_i from p_i is determined by the presence of answer span a_i, which contains not only the exact match for the gold answer but also key evidence to the question. Given a question q_i, the reader finds the corresponding answer span a_i from an answerable passage p_i^+ in the corpus C.

Due to the large search space in the corpus C, a first-stage retriever is commonly used in ODQA to find subsets of relevant passages to questions for the expensive reader. The predominant approach to passage retrieval is DPR (Karpukhin et al., 2020), which leverages the efficient dual-encoder architecture denoted as $[f_q, f_p]$ to encode questions and passages into a learned embedding space. For a question-answer passage pair (q_i, p_i^+) and a set of N negative passages p_j^-, DPR is trained to maximize the similarity between the question q_i and its answer passage p_i^+:

$$\mathcal{L}(q_i, p_i^+, \{p_j^-\}_{j=1}^{N}) = -log(\frac{e^{sim(q_i,p_i^+)}}{e^{sim(q_i,p_i^+)} + \sum_{j=1}^{N} e^{sim(q_i,p_j^-)}}) \tag{1}$$

where $sim(\cdot,\cdot)$ is a function that computes the relevance score between a question and a passage as dot product between the question embedding $f_q(q_i)$ and the passage embedding $f_p(p_i)$:

$$sim(q_i, p_i) = f_q(q_i) \cdot f_p(p_i) \tag{2}$$

At runtime, the retriever indexes all passages based on the similarity metric and performs efficient search using approximate nearest neighbor search libraries (Johnson & Douze, 2019).

To further enhance the discriminative power of DPR, more recent studies on dense retrieval adopt model-centric approaches (Xiong et al., 2021; Qu et al., 2021; Izacard & Grave, 2021a), which involve iterative training pipelines that exploit multiple fine-tuned encoders. For example, state-of-the-art models including ANCE (Xiong et al., 2021) and RocketQA (Qu et al., 2021) focus largely on combining multiple dual- and cross-encoders into sophisticated frameworks for hard negative sampling. While effective, as presented in Table 1, using these complex model-centric approaches requires a significant amount of computing resources, which limits their development in diverse environments (Lindgren et al., 2021; Du et al., 2022; Gao et al., 2022). In this work, we instead focus on a data-centric approach, which aims to maximally leverage the given data sources without complicating the training process, to improve the ranking performance and robustness of DPR. We later show that our data-centric approach based on counterfactual augmentation is both effective as alternatives to DPR training and orthogonal to existing model-centric approaches.

Table 1: Resource comparison of various dense retrieval approaches for Natural Questions dataset. Each column indicates minimal GPU resources to train, reported batch size, whether using knowledge distillation teacher, and the frequency of refresh indices (passage resampling) during training, respectively, reported on (Hofstätter et al., 2021).

Retriever	Min.GPU	Batch Size	KD Teacher	Index Refresh
DPR / DPR+PiCL	1×GTX	32	-	-
ANCE	**4×V100**	32	-	**10K batches**
RocketQA	2×V100	**1024**	Cross Encoder	2×

2.2 Counterfactual Learning for Text Data

Counterfactual inference has been applied to representation learning to obtain fair representations (Kusner et al., 2017) in various domains such as image classification (Goyal et al., 2019) and vision-language tasks (Liang et al., 2020; Niu et al., 2021). The key idea behind counterfactual learning is to train a model that is invariant to specific aspects of the input data (Johansson et al., 2016; Kusner et al., 2017).

Existing work on counterfactual learning on text data expands upon this idea and aims at learning to discriminate causal signals from spurious correlations in learned representations of text data. Specifically, Choi et al. (2020b); Liang et al. (2020); Choi et al. (2022); Tian et al. (2022) show that using minimally different samples is helpful for learning robust representations that are invariant to spurious signals. They generate counterfactual samples via dedicated masking strategies and apply contrastive learning on synthetic samples to discern counterfactual samples from original data.

However, few studies have delved into the effect of counterfactual learning on retrieval tasks (Choi et al., 2020a). In this work, we reinterpret causal signals in existing work as answer spans in ODQA and seek to apply counterfactual contrastive learning for DPR.

3 Relevance-Answerability Asynchronicity on DPR

In this section, we first present the definition of counterfactual samples as unanswerable variants of the given passages. We then conduct a counterfactual analysis on DPR where retrievers compare the relevance scores between an original passage and its counterfactual sample. We show that dense retrievers are incapable of discriminating counterfactual samples from the original data and thus unaware of answerability.

3.1 Unanswerable Counterfactual Samples for Passage Retrieval

Following previous studies (Liang et al., 2020; Choi et al., 2022), we synthesize counterfactual samples by removing causal signals from the original sample. Based on our assumption that answer spans serve as causal signals in ODQA, we remove the answer span to a question in the original passage to create its counterfactual sample. Formally, we define counterfactual samples for passage retrieval as follows:

Definition 1. Let (q, p) denote a question-answer passage pair, where $p = [s_l; a; s_r]$ contains an answer span a to the question q and non-answer spans s_l and s_r. We define a **counterfactual sample** p^* as an unanswerable variant of p with minimal changes such that $p^* = [s_l; s_r]$.

Intuitively, the counterfactual sample is an unanswerable variant of the original passage that loses relevant features to the question (*i.e.*, a). In practice, we set a to be the evidence sentence that contains the exact answer to the question. Figure 1 provides the overview of our counterfactual sampling process.

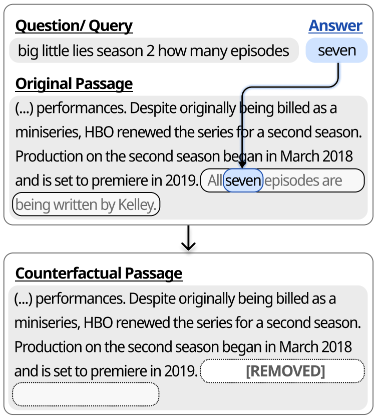
Question/ Query
Answer
big little lies season 2 how many episodes
seven
Original Passage
(...) performances. Despite originally being billed as a miniseries, HBO renewed the series for a second season. Production on the second season began in March 2018 and is set to premiere in 2019. All seven episodes are being written by Kelley.
Counterfactual Passage
(...) performances. Despite originally being billed as a miniseries, HBO renewed the series for a second season. Production on the second season began in March 2018 and is set to premiere in 2019.
[REMOVED]

Figure 1: Illustration of the counterfactual sampling strategy.

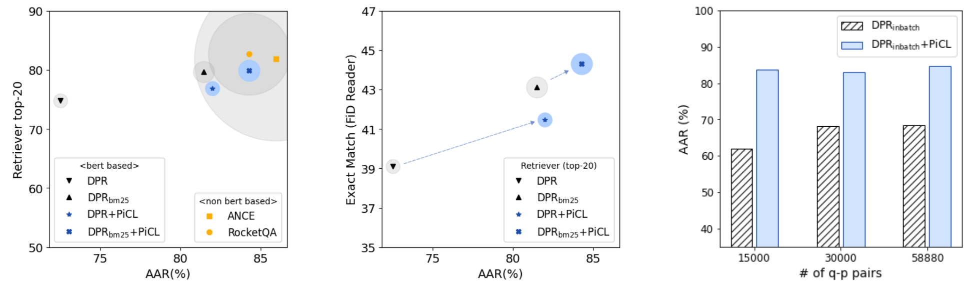

Figure 2: **Left:** AAR and retrieval performance of dense retrievers. **Mid:** AAR and end-to-end QA accuracy (Exact Match). The radius around the mark indicates the minimum GPU memory usage for training (reported in Table 1). **Right:** AAR on varying the number of training samples.

3.2 MEASURING ANSWER-AWARENESS OF PASSAGE RETRIEVAL

Our experiment aims to re-examine the belief that dense retrievers are **answer-aware**, or that the relevance score $sim(q, p)$ between q and p is synchronous with the answerability of q from p. Given that counterfactual samples lack key features relevant to the question, it is generally assumed that dense retrievers rank positive passages higher than their counterfactual counterparts. To assess the answer-awareness of dense retrieval approaches, we check whether such assumption holds true. Specifically, we observe answerability-relevance mismatch, where answerable passages p are given lower relevance scores than their counterfactual counterparts p^*.

Definition 2. (Answerability-Relevance Mismatch) Given a question-passage pair (q, p) where p contains answer a to q, let p^* be the counterfactual sample of p such that $a \notin p^*$, and sim be the scoring function for relevance between q and p. Answerability-relevance mismatch between the question-passage pair (q, p) occurs if the relevance score $sim(q, p)$ between q and p is lower than (or equal to) the score $sim(q, p^*)$ between q and p^*, *i.e.*, $sim(q, p) \leq sim(q, p^*)$

Under the counterfactual simulation setting, we further introduce **Answer-Awareness Rate** of a model to quantify the rate at which the model predictions on relevance scores $sim(q, p)$ are synchronous with the answerability of q to p. Given a set of T question-answer passage pairs $\{(q_i, p_i)\}_{i=1}^{T}$, answer-awareness rate counts the number of cases where answerability-relevance mismatch does not occur.

Definition 3. (Answer-Awareness Rate) Given a corpus C and a set of T triplets $\mathcal{T} = \{(q_i, p_i, p_i^*) | p_i, p_i^* \in C\}_{i=1}^{T}$ where p_i^* is the counterfactual sample of p_i, let $\mathbb{1}_{sim(q_i,p_i) \leq sim(q_i,p_i^*)}$ be a binary indicator of whether answerability-relevance mismatch occurs for a triplet (q_i, p_i, p_i^*), *i.e.*, $sim(q_i, p_i) \leq sim(q_i, p_i^*)$. Answer-Awareness Rate, or AAR, is measured as the proportion of (q_i, p_i, p_i^*) triplets whose relevance scores $sim(q_i, p_i)$ between questions and original passages are higher than scores $sim(q_i, p_i^*)$ between their counterfactual counterparts p^*.

$$\text{AAR} = 1 - \sum_{i=1}^{T} \mathbb{1}_{sim(q_i,p_i) \leq sim(q_i,p_i^*)} / T \tag{3}$$

To validate the assumption on answer-awareness of dense retrievers, we observe how much AARs differ from their theoretical upper bound, which supposedly amounts to 100 percent provided that dense retrievers always measure relevance scores such that $sim(q, p) > sim(q, p^*)$.

3.3 PRELIMINARY EXPERIMENTS

We conduct our preliminary experiment on Natural Questions (NQ) (Kwiatkowski et al., 2019), a commonly used ODQA benchmark. We follow the settings from Karpukhin et al. (2020) and adopt NQ test set used for DPR, which provides 3,610 factoid questions and around 21 million passages segmented from Wikipedia dump for evaluation. To measure AAR, we use the 1,382 questions from NQ test set whose gold (q, p, p^*) triplets are given.[1] We then track whether the relevance score $sim(q, p)$ between each question q and its positive passage p is higher than the score $sim(q, p^*)$ between the counterfactual passage p^*.

[1] Both NQ test set and golden passage information are made available by Karpukhin et al. (2020) at `https://github.com/facebookresearch/DPR`

Figure 2 shows both AAR and the retrieval performance of dense retrievers. We first observe that AAR of a vanilla DPR significantly falls behind the theoretical upper bound, which contradicts the assumption that retrievers constantly rank positive passages higher than counterfactual passages. We also see that AAR of dense retrievers show positive correlations with both their retrieval performance and end-to-end QA accuracy. A considerable increase in QA performances following the increase in AAR further suggests that answer-aware retrievers improve the effectiveness of retriever-reader frameworks even with a small gain in the retrieval performance.

Effect of Hard Negatives. One possible solution to improve AAR of retrievers is to adopt better negative mining strategies. To assess the effect of hard negatives, we compare AAR of $\text{DPR}_{\text{inbatch}}$ and DPR_{bm25}, which are trained with in-batch negatives and hard negatives from BM25 index, respectively. We also compare AAR of DPR with those of ANCE and RocketQA, which adopt sophisticated negative sampling strategies for retriever training. From Figure 2, we observe that DPR_{bm25} show better AAR than $\text{DPR}_{\text{inbatch}}$, and that ANCE and RocketQA achieve better AAR than DPR. However, these approaches require large computational cost since they depend on large PLMs and negative mining strategies which usually involve multi-stage training of different encoders.

Effect of Sample Size. Another possible solution is to increase the number of training samples. To examine the correlation between the size of the train set and AAR of the resulting DPR model, we randomly sample 15,000 and 30,000 question-passage pairs from NQ train set and additionally train two DPR models with randomly sampled data. Figure 2 shows that AAR of DPR increases when trained on more question-passage pairs. Despite such correlation, there is a practical limit to the size of the collected dataset, which makes any such approach infeasible.

Later in this paper we present our data-centric approach based on counterfactual augmentation, PiCL. Note that in Figure 2, DPR+PiCL consistently shows better AAR than DPR trained under the same settings, *i.e.* negative sampling strategies and sample sizes.

4 Synchronizing Relevance and Answerability on DPR

Having seen that dense retrievers are not robust to unanswerable counterfactual passages, in this section, we present a novel contrastive learning approach that repurposes counterfactual samples to train a retriever robust to the counterfactual samples. In Section 4.1, we redefine counterfactual samples as pivots between positive and negative passages to adapt them into a plain DPR setting. In Section 4.2, we propose counterfactually-**Pi**voting **C**ontrastive **L**earning, which learns the relative similarity between questions, positives, negatives, and pivoting samples.

4.1 Redefining Counterfactual Samples as Pivots

We have reinterpreted causal signals in dense retrieval as answer spans in ODQA and synthesized unanswerable counterfactual samples in Section 3.1. Our goal is to incorporate counterfactual contrastive learning into DPR training such that the learned model is invariant to answer-irrelevant features. For that, we aim to design a method where the model learns the relative positions of counterfactual samples on the embedding space from DPR.

Consider a question-passage pair (q_i, p_i^+), its corresponding counterfactual sample p_i^*, and a set of N negative passages $\{p_j^-\}_{j=1}^N$. A DPR-based retriever learns question and passage representations such that the following inequality between positive passages p_i^+ and negative passages p_j^- holds:

$$sim(q_i, p_i^+) > sim(q_i, p_j^-) \tag{4}$$

To specify the relative positions of counterfactual passage on the embedding space, one must take into account their relevance between positive passages (*i.e.* source passages) and negative passages. As shown in Figure 3, we view our counterfactual samples from two different perspectives, as hard negatives and as pseudo-positives.

Counterfactuals as Hard Negatives. Given that counterfactual samples lack relevant features to the question, p_i^* is a counterfactually negative sample to an anchor question q_i while the original passage p_i^+ serves as the positive. Thus the embedding similarity $sim(q_i, p_i^*)$ between q_i and p_i^* is upper bounded by $sim(q_i, p_i^+)$:

$$sim(q_i, p_i^+) > sim(q_i, p_i^*) \tag{5}$$

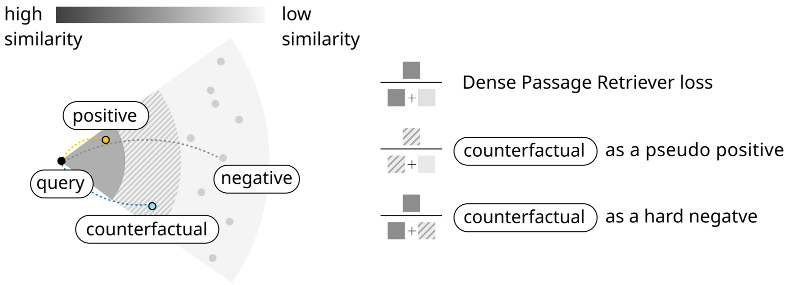
high similarity
low similarity
positive
query
negative
counterfactual
Dense Passage Retriever loss
counterfactual as a pseudo positive
counterfactual as a hard negatve

Figure 3: Illustration of the counterfactual samples as pivots in the embedding space.

Learning to discriminate p_i^* from p_i^+ would minimize the mutual information between representations of questions q_i and answer-irrelevant spans in p_i^+.

Counterfactuals as Pseudo Positives. Since a counterfactual passage p_i^* is a minimally different sample that retains most of semantics in the positive passage p_i^+, p_i^* can be seen as a pseudo-positive example in the passage retrieval. Semantic relevance between q_i and p_i^* distinguishes p_i^* from other negatives p_j^-, which provide noisy contexts with respect to q_i. Thus, the following holds for all p_j^-:

$$sim(q_i, p_i^*) > sim(q_i, p_j^-) \tag{6}$$

Learning to discriminate p_i^* from p_j^- imposes a constraint on the encoders such that embeddings for questions and passages are computed based on their semantic alignment.

From Equation 5 and 6, we can derive that the embedding similarity $sim(q_i, p_i^*)$ between questions and counterfactual samples are bounded by $sim(q_i, p_i^+)$ and $sim(q_i, p_j^-)$. Essentially, they can be re-formulated as pivots between positive and negative samples in the embedding space. Since both inequality constraints in Equation 5 and 6 satisfy Equation 4, our definition of counterfactual samples as pivots is in line with the training objective of DPR.

4.2 Counterfactually-Pivoting Contrastive Learning

Based on the ideas discussed in Section 4.1, we propose counterfactually-**Pi**voting **C**ontrastive **L**earning (**PiCL**) for dense retrieval for a robust, answer-aware dense retriever. Specifically, we introduce additional counterfactual contrastive loss terms into the passage retrieval objective function in Karpukhin et al. (2020) to leverage counterfactual samples as pivots between positives and negatives. Consider a question-positive passage pair $\{q_i, p_i^+\}$, a set of N negatives $\{p_j^-\}_{j\neq i}^N$, and their corresponding counterfactual samples p_i^* and $\{p_j^*\}_{j\neq i}^N$. We define following loss terms as key components of PiCL:

Modified Dense Passage Retrieval Loss. The modified DPR loss $\mathcal{L}_{\texttt{dpr}}$ is a slight modification of the loss term in Karpukhin et al. (2020) in which the counterfactual counterpart p_i^* to the positive passage p_i^+ is added as a negative. To alleviate any interference from masked passages as negatives, we add a balancing coefficient $\lambda < 1$ before the similarity term $e^{sim(q_i, p_i^*)}$.

$$\mathcal{L}_{\texttt{dpr}}(q_i, p_i^+, p_i^*, \{p_j^-\}_{j\neq i}^N) = -\log(\frac{e^{sim(q_i,p_i^+)}}{e^{sim(q_i,p_i^+)} + \sum_{j\neq i}^N e^{sim(q_i,p_j^-)} + \lambda \cdot e^{sim(q_i,p_i^*)}}) \tag{7}$$

Counterfactuals as Hard Negatives. The Counterfactuals-as-Hard-Negatives loss $\mathcal{L}_{\texttt{chn}}$ is optimized to maximize the similarity between q_i and p_i^+ while minimizing the similarity between q_i and p_i^*. It imposes the key constraint on q_i, p_i^+, and p_i^* from Equation 5 to discriminate answer passages from non-answer counterfactual passages:

$$\mathcal{L}_{\texttt{chn}}(q_i, p_i^+, p_i^*) = -\log(\frac{e^{sim(q_i,p_i^+)}}{e^{sim(q_i,p_i^+)} + e^{sim(q_i,p_i^*)}}) \tag{8}$$

Counterfactuals as Pseudo Positives. The Counterfactuals-as-Pseudo-Positives loss $\mathcal{L}_{\texttt{cpp}}$ is optimized to maximize the relative similarity between q_i and p_i^* with respect to negative passages p_j^-

and p_j^* in the given batch. The key difference between $\mathcal{L}_{\texttt{dpr}}$ and $\mathcal{L}_{\texttt{cpp}}$ is that the counterfactual passage is used as a positive in $\mathcal{L}_{\texttt{cpp}}$ to retain semantic relevance in the learned embeddings.

$$\mathcal{L}_{\texttt{cpp}}(q_i, p_i^*, \{p_j^-, p_j^*\}_{j \neq i}^N) = -\log(\frac{e^{sim(q_i, p_i^*)}}{e^{sim(q_i, p_i^*)} + \sum_{j \neq i}^N (e^{sim(q_i, p_j^-)} + e^{sim(q_i, p_j^*)})}) \tag{9}$$

The final loss function $\mathcal{L}$ is a weighted sum of all three loss fuctions $\mathcal{L}_{\texttt{dpr}}$, $\mathcal{L}_{\texttt{chn}}$, and $\mathcal{L}_{\texttt{cpp}}$:

$$\mathcal{L} = \mathcal{L}_{\texttt{dpr}} + \tau_1 \mathcal{L}_{\texttt{chn}} + \tau_2 \mathcal{L}_{\texttt{cpp}} \tag{10}$$

where τ_1, τ_2 are hyperparameters that determine the importance of the terms. Later in Section 5.2 we provide an analysis on the effect of each loss term through an ablation study.

5 Experiments

In this section, we first evaluate the effectiveness of PiCL for passage retrieval task on common ODQA benchmarks. We then apply PiCL on a retriever-reader pipeline and assess its end-to-end QA performance to show that improving answer-awareness of the retriever leads to a more accurate reader. We also provide some analysis on PiCL as possible explanations to our design choice.

5.1 Experimental Settings

Dataset. Following the settings in Sections 3.3, we evaluate our method on Natural Questions (Kwiatkowski et al., 2019). This benchmark is built upon web documents such as Wikipedia articles and human-annotated question-answer pairs, most of which are collected from real-world search queries (Kwiatkowski et al., 2019). We follow prior work on dense retrieval (Xiong et al., 2021; Qu et al., 2021; Ren et al., 2021) and use for training 58,812 factoid questions each paired with a set of positive passages. For inference, we use the preprocessed data from Karpukhin et al. (2020), which include 3,610 factoid questions and a corpus of about 21M Wikipedia passages preprocessed through the pipeline proposed in Chen et al. (2017).

Baselines and Implementation Details. We consider DPR (Karpukhin et al., 2020) as the backbone architecture for all models implemented in this section. Our focus is to assess how applying PiCL affects the downstream performance of the backbone DPR. We also include ANCE (Xiong et al., 2021) and RocketQA (Qu et al., 2021), which use dedicated negative sampling pipelines to improve the performance of vanilla DPR. To show the orthogonality of PiCL to such negative mining approaches, we implement PiCL models under different negative sampling strategies (*e.g.*, sampling from top retrieval results of BM25 and a fine-tuned DPR) and compare them with the baselines. Note that due to limited computation resource, we do not reproduce the performance of ANCE and RocketQA. See Appendix A for more details on the implementation.

5.2 Main Experimental Results

Performance of Passage Retrievers. Table 2 compares the performance of PiCL models with the baselines, *i.e.*, DPR, ANCE, and RocketQA on NQ benchmark. We observe that DPR+PiCL models yield consistent performance gains over vanilla DPR under all tested conditions. Similar to DPR, DPR+PiCL shows particularly strong performance when trained with hard negatives from top retrieval results of a fine-tuned DPR, which suggests that the performance of DPR+PiCL can be further boosted by adding informative negative samples.[2]. Meanwhile, we see a consistent increase in the performance of dense retrievers when trained using model-centric approaches, *i.e.*, ANCE and RocketQA, which rely on larger pretrained language models (*e.g.*, ERNIE (Sun et al., 2020) and RoBERTa (Zhuang et al., 2021)) and compute-intensive negative mining techniques. The performance gain on DPR+PiCL when using additional hard negatives implies that PiCL can be further improved when orthogonally applied to such model-centric approaches.

[2] We obtain hard negative samples from the dataset `nq-adv-hn-train`, which is available on https://github.com/facebookresearch/DPR

Table 2: Main results on Natural Questions benchmark. The best and second results are in **Bold** and underline, respectively. PLMs: pre-trained LMs used for retriever initialization, #N: the number of negative samples, computed as in-batch negatives + hard negatives, * denotes reproduced results in our environment setting where #N=127+128. † indicates reported results from Qu et al. (2021).

Retriever	**Training Resource**			**Natural Questions**			
	PLM	#N	Hard Neg	Top-1	Top-5	Top-20	Top-100
BM25†	-	-	-	-	-	59.1	73.7
DPR†	BERT$_{\text{base}}$	127	-	-	55.8	73.0	83.1
ANCE†	RoBERTa$_{\text{base}}$	31+32	ANN	-	-	81.9	87.5
RocketQA$^\dagger_{\text{step1}}$	BERT$_{\text{base}}$	1023+1024	Cross Batch	-	-	-	86.0
RocketQA†	ERNIE$_{\text{base}}$	1023+1024	Cross Encoder	-	74.0	82.7	88.5
DPR*	BERT$_{\text{base}}$	127	-	31.77	58.12	74.76	84.07
DPR*	BERT$_{\text{base}}$	127+128	BM25	46.62	68.56	79.7	86.34
DPR+PiCL	BERT$_{\text{base}}$	127	-	32.69	60.39	76.73	85.48
DPR+PiCL	BERT$_{\text{base}}$	127+128	BM25	47.81	69.25	79.92	**87.12**
DPR+PiCL	BERT$_{\text{base}}$	127+128	DPR	**52.57**	**71.10**	**80.50**	86.84

Table 3: End-to-end QA performance of retriever-reader pipelines on Natural Questions benchmark. #N : number of negative samples used to train the retriever, computed as in-batch negatives (+ BM25 negatives), Top-k: top-k retrieved passages for reader inference. Note that we reuse the DPR reader and Fusion-in-Decoder base (FiD$_{\text{base}}$) models from Karpukhin et al. (2020) and Izacard & Grave (2021a), switching retrievers to obtain EM score. Best scores are in **Bold**.

#N	Reader	Retriever	Exact Match (EM) score		
			Top-5 passages	Top-20 passages	Top-100 passages
127	DPR reader	DPR	31.83	36.87	37.45
		DPR+PiCL	34.52 (+2.69)	38.31 (+1.44)	38.81 (+1.36)
127	FiD$_{\text{base}}$ (T5)	DPR	31.99	39.11	43.82
		DPR+PiCL	34.27 (+2.28)	41.47 (+2.36)	44.85 (+1.03)
127+128	FiD$_{\text{base}}$ (T5)	DPR	38.31	43.13	45.37
		DPR+PiCL	**40.22** (+1.91)	**44.32** (+1.19)	**47.65** (+2.28)

End-to-End QA Performance of Retriever-Reader. A key assumption underlying our work is that synchronizing a retriever with the reader improves end-to-end QA performance of the retriever-reader pipeline. To validate this assumption, we incorporate DPR+PiCL into a QA system and evaluate the downstream performance of the subsequent reader. Specifically, we re-use the reader models from Karpukhin et al. (2020) and Izacard & Grave (2021b) and switch different retrievers to sample top-k passages for reader inference. We then compute Exact Match (EM) scores for the reader, which measures the proportion of questions whose answer prediction is equivalent to correct answers. Table 3 reports end-to-end QA performance of the retriever-reader pipelines. Overall, applying PiCL on DPR consistently improves the QA performance of the retriever-reader pipeline under various settings. Particularly, we see that using a FiD$_{\text{base}}$ instead of a DPR reader for DPR+PiCL further boosts the performance of the QA system, suggesting that PiCL models can be orthogonally applied with advanced reader models.

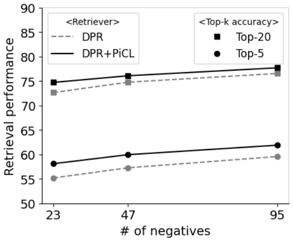

Figure 4: Retrieval performance of DPR and DPR+PiCL on varying the numbers of negatives.

Retriever	**Natural Questions**			
	Top-1	Top-5	Top-20	Top-100
DPR	32.08	59.56	76.54	85.73
(1) Counterfactual Loss				
$\mathcal{L}_{\texttt{dpr}} + \mathcal{L}_{\texttt{cpp}}$	28.14	56.65	75.12	85.54
$\mathcal{L}_{\texttt{dpr}} + \mathcal{L}_{\texttt{chn}}$	30.80	58.59	75.98	84.85
$\mathcal{L}_{\texttt{cpp}} + \mathcal{L}_{\texttt{chn}}$	33.35	61.36	77.87	86.09
$\mathcal{L}_{\mathbf{dpr}} + \mathcal{L}_{\mathbf{cpp}} + \mathcal{L}_{\mathbf{chn}}$	33.99	61.88	77.7	85.84
(2) Counterfactual Sampling				
PiCL$_{\texttt{answer}}$	30.28	57.12	73.99	84.52
PiCL$_{\texttt{window}}$	31.50	60.03	76.40	85.18
PiCL$_{\texttt{sentence}}$	33.99	61.88	77.7	85.84

Table 4: Ablation studies on PiCL.

5.3 Analysis on Dense Passage Retrieval

Negative Samples. Figure 4 shows the retrieval accuracy of DPR and DPR+PiCL trained with different amounts of in-batch and random negative samples [3]. We see a consistent increase in retrieval accuracy for DPR+PiCL, which is in line with the finding that the performance gain from PiCL over a vanilla DPR is independent of the number of negative samples used to train the model. Particularly, DPR+PiCL trained with fewer negatives achieves performance comparable to or better than a vanilla DPR trained with more negatives, as shown in Figure 4. One possible reason for such efficiency is that using counterfactual samples for PiCL leads to the effect of data augmentation as the model learns to discriminate positive passages from hard negative passages.

Counterfactual Loss. To study the efficacy of counterfactual samples as pivots, we conduct an ablation study on counterfactual contrastive learning objective in PiCL (Equation 10). Specifically, we implement three PiCL baselines with the following modifications on the objective function in Equation 10, 1) $\mathcal{L}_{\texttt{dpr}} + \mathcal{L}_{\texttt{cpp}}$, 2) $\mathcal{L}_{\texttt{dpr}} + \mathcal{L}_{\texttt{chn}}$, and 3) $\mathcal{L}_{\texttt{cpp}} + \mathcal{L}_{\texttt{chn}}$. Table 4 compares all baselines with the original PiCL and the vanilla DPR. All three PiCL baselines do not improve much over a vanilla DPR without either $\mathcal{L}_{\texttt{dpr}}$, $\mathcal{L}_{\texttt{cpp}}$ or $\mathcal{L}_{\texttt{chn}}$. In contrast, the original PiCL method outperforms the vanilla DPR, suggesting that using counterfactual samples as pivots is crucial in PiCL training.

Counterfactual Sampling Strategy. To validate our strategy of synthesizing counterfactual samples, we present different augmentation schemes and assess their effect on downstream performance of DPR+PiCL. Specifically, we consider the following four augmentation schemes: 1) removing the exact answer match, PiCL$_{\texttt{answer}}$, 2) removing an extended span around the answer, PiCL$_{\texttt{window}}$, and 3) removing the answer sentence, PiCL$_{\texttt{sentence}}$. Table 4 reports the retrieval performance of PiCL models trained under different masking strategies. Overall, PiCL$_{\texttt{sentence}}$ yields the best performance among all models. This implies that not only answer string but also its context is an important factor in learning answer-awareness.

Answer-awareness. One key assumption underlying our approach is that applying PiCL on DPR results in a more answer-aware retriever. To provide a fine-grained analysis on answer-awareness of PiCL, we use DPR+PiCL trained with in-batch negatives to measure similarity scores $sim(q, p^+)$ and $sim(q, p^*)$ for the gold (q, p^+, p^*) triplets from Section 3.3. Figure 5 shows the average of the scores $sim(q, p^+)$ and $sim(q, p^*)$ from DPR+PiCL compared to those from DPR. We observe a notable difference between the average $sim(q, p^+)$ and $sim(q, p^*)$ for DPR+PiCL and a relatively minor difference for DPR. In DPR+PiCL, the average similarity score $sim(q, p^+)$ is significantly larger than the score $sim(q, p^*)$, which indicates that DPR+PiCL learns answer-awareness to distinguish between the answerable passage and the unanswerable counterfactual one.

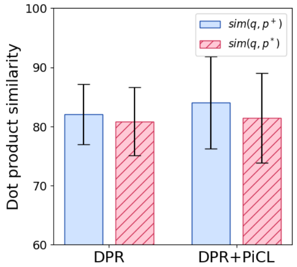

Figure 5: Similarity scores between queries and positive passages or counterfactual passages in DPR and DPR+PiCL.

6 Conclusion and Future Work

This work aims to re-examine the assumption that dense retrievers assign high relevance scores on answerable passages to questions. We first conduct a counterfactual simulation and observe that dense retrievers based on DPR often fail to rank answerable passages higher than their counterfactual counterparts. Based on this observation, we present PiCL, which repurposes counterfactual samples as pivots between positives and negatives for answer-aware DPR. Our experiments show that applying PiCL on DPR not only enhances the model performance on the ODQA benchmark but also improves its robustness to unanswerable, counterfactual passages. We believe that the concept of counterfactually-pivoting samples can be further explored in future work: 1) extending counterfactuals to multiple pivots of incrementally removing causal signals at different levels, 2) refining the definition of answer-awareness or AAR into formal evaluation metrics of ODQA, and 3) shifting a target task to other representation learning tasks such as text/image classification and response retrieval for dialogue systems.

[3] We added one randomly sampled negative passage for each question.

REFERENCES

Danqi Chen and Wen-tau Yih. Open-domain question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, pp. 34–37, 2020.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, 2017.

Seungtaek Choi, Myeongho Jeong, Jinyoung Yeo, and Seung-won Hwang. Label-efficient training for next response selection. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pp. 164–168, 2020a.

Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seung-won Hwang. Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6695–6704, 2020b.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. C2l: Causally contrastive learning for robust text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:10526–10534, Jun. 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Tevatron: An efficient and flexible toolkit for dense retrieval. *arXiv preprint arXiv:2203.05765*, 2022.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2376–2384. PMLR, 09–15 Jun 2019.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pp. 3929–3938. PMLR, 2020.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 113–122, 2021.

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

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021b.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR.

Jeff Johnson and Hervé Douze, Matthijs andJégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3285–3292. Association for Computational Linguistics, November 2020.

Erik Lindgren, Sashank Reddi, Ruiqi Guo, and Sanjiv Kumar. Efficient training of retrieval models using negative cache. *Advances in Neural Information Processing Systems*, 34:4134–4146, 2021.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12700–12710, 2021.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of NAACL*, 2021.

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

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

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L Hamilton, and Bryan Catanzaro. End-to-end training of neural retrievers 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)*, pp. 6648–6662, 2021.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8968–8975, 2020.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. Debiasing nlu models via causal intervention and counterfactual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of ICLR*, 2021.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1503–1512, 2021.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pp. 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL `https://aclanthology.org/2021.ccl-1.108`.

A Implementation Details

In Table 2, we train all implemented models for 40 epochs on a single server with two 16-core Intel(R) Xeon(R) Gold 6226R CPUs, a 264GB RAM, and 8 24GB GPUs. For PiCL training, we set batch size as 16, learning rate as 2e-5, and eps and betas of the adam optimizer as 1e-8 and (0.9, 0.999), respectively. Note that we conduct experiments on the NQ benchmark under the same settings used in Karpukhin et al. (2020). Among the hyperparameters $\{0.1, 0.2, 0.5, 0.9, 1.0\}$, we choose 0.2 for the balancing coefficient λ for counterfactual samples in Equation 7 and 1.0 for the weight hyperparameters τ_1, τ_2 in Equation 10.

For additional analysis in Section 5.3, we move onto a more viable settings with limited computation resources. Specifically, we train all models for 25 epochs on a server with 3 24GB GPUs. We choose 0.2 for the balancing coefficient λ for counterfactual samples in Equation 7 and 0.2 for the weight hyperparameters τ_1, τ_2 in Equation 10. All other hyperparameters including batch size and optimizer remain unchanged.

For reader in Section 5.2, we consider two models: 1) the extractive reader from Karpukhin et al. (2020) implemented on pretrained BERT models (Devlin et al., 2019) and 2) Fusion-in-Decoder reader (Izacard & Grave, 2021b) based on pretrained T5-base (Raffel et al., 2020) models. We conduct inference for the reader on a single 24GB GPU with the batch size of 8.

B Additional Analysis of Answer-awareness

In this section, we extend our preliminary experiment in Section 3.2 and delve into two important aspects of answer-aware retrievers: 1) Score Differences and 2) Effect of Question Types.

Samples	**DPR**	**DPR+PiCL**
Answer-aware	1.99	3.38
Answer-unaware	-0.98	-1.12
All	1.12	2.85

Table 5: Average score differences between $sim(q, p^+)$ and $sim(q, p^*)$, measured using both DPR and DPR+PiCL. Score differences are measured on two sets of (q, p^+, p^*) triplets, one that satisfies $sim(q, p^+) > sim(q, p^*)$ (*i.e.* Answer-aware) and another that doesn't (*i.e.* Answer-unaware).

Score Differences. To provide an overview of how relevance scores differ between original and counterfactual passages, we expand upon the answer-awareness analysis in Section 5.3 and focus on the difference in relevance scores $sim(q, p^+)$ and $sim(q, p^*)$. Specifically, we aim to examine whether DPR+PiCL tends to put larger differences between original and counterfactual passages compared to DPR. For that, we reuse the gold (q, p^+, p^*) triplets from Section 3.3 and compute the score difference $sim(q, p^+) - sim(q, p^*)$ for each (q, p^+, p^*). The relevance scores are measured using both DPR and DPR+PiCL for a comparative analysis.

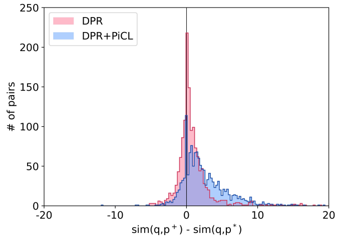

Figure 6: Comparison between distributions of score differences $sim(q, p^+) - sim(q, p^*)$ drawn from DPR and DPR+PiCL.

Table 5 compares the average score differences measured from DPR and DPR+PiCL. For our analysis, we consider two subsets of samples, one that satisfies $sim(q, p^+) > sim(q, p^*)$ (*i.e.* Answer-aware) and another that doesn't (*i.e.* Answer-unaware). We observe that training DPR with PiCL generally leads to larger score differences than a DPR, particularly for samples where the retrievers are answer-aware, i.e. $sim(q, p^+) > sim(q, p^*)$. In line with the previous observations from the AAR experiment in Section 3.3, the results in Table 5 suggest that DPR+PiCL learns to distinguish answerable passages from counterfactual samples. While the differences may seem small, we note that a slight decrease in the relevance score leads to a large drop in the rank of a passage.

We further analyze the distributions of score differences and examine how such distributions affect the answer-awareness of retrievers. For that, we plot and compare the distribution of score differences from DPR and DPR+PiCL. We see from Figure 6 that the distribution of $sim(q, p^+) - sim(q, p^*)$ drawn from DPR+PiCL is skewed more positive than negative, while the distribution from DPR tends to be more symmetrical. Negative skewness of the distribution from DPR+PiCL indicates that 1) more (q, p^+, p^*) samples satisfy $sim(q, p^+) > sim(q, p^*)$ and that 2) the retriever tends to assign higher relevance scores on p^+. This result in turn suggests that DPR+PiCL is more answer-aware, *i.e.*, capable of distinguishing answerable passages from counterfactuals.

Effect of Question Types. To better understand if the answer-awareness of a retriever depends on certain linguistic features, we start by investigating the effect of question types on the answer-awareness of DPR. Specifically, we reuse the gold (q, p^+, p^*) triplets from AAR experiments and classify them into six subsets based on question types. We identify the type of a question simply by looking for exact matches of the question words: how, what, when, where, which, and who.

Question Type	# Questions	AAR(%)	
		DPR	DPR+PiCL
How	114	71.93	85.09
What	228	67.54	77.19
When	295	80.68	87.46
Where	146	65.07	81.51
Which	41	73.17	85.37
Who	547	70.02	80.07

Table 6: AARs of DPR and DPR+PiCL on various question types.

Table 6 shows AARs of DPR and DPR+PiCL on the six subsets of different question types. We first observe that AARs of dense retrievers do vary significantly across different question types, implying that the answer-awareness of DPR depends on the types of questions being asked. Particularly, DPR shows large gaps in AARs between different question types, ranging from 65.07% to 80.68%. On the other hand, DPR+PiCL shows 1) significant improvements in AAR over a vanilla DPR for all question types and 2) smaller variance in AARs across different question types.

Question: Who died in the plane crash greys anatomy

Passage Type	Title	Text
Gold passage	Flight (Grey's Anatomy)	Flight " is the twenty - fourth and final episode of the eighth season of the American television medical drama Grey 's Anatomy ... six doctors from Seattle Grace Mercy West Hospital who are victims of an aviation accident fight to stay alive , but **Dr. Lexie Grey** (Chyler Leigh) ultimately dies. ...
Top-1 DPR	Paul-Louis Halley	Socata TBM 700 aircraft crash on 6 December 2003, during an approach to Oxford Airport. The plane went into an uncontrolled roll, killing Halley, his wife, and the pilot. ...
Top-9 DPR	Flight (Grey's Anatomy)	(Patrick Dempsey), and Dr. Mark Sloan (Eric Dane) desperately fight to stay alive. Meredith is relatively unscathed, while the rest have serious injuries: the pilot, Jerry (James LeGros), has a major spine injury, ... Shepherd is sucked out the side of the plane and awakens alone in the wood; his mangled hand having been pushed through the door of the plane. However, none are in as bad shape as **Lexie**, who is crushed under
Top-1 DPR+PiCL	Comair Flight 5191	... The Flight 5191 Memorial Commission was established shortly after the crash to create an appropriate memorial for the victims, first responders, and community that supported them. The Commission chose the University of Kentucky Arboretum as its memorial site. James Polehinke, the first officer, suffered serious injuries, including multiple broken bones, a collapsed lung, and severe bleeding. ...
Top-2 DPR+PiCL	Flight (Grey's Anatomy)	(Patrick Dempsey), and Dr. Mark Sloan (Eric Dane) desperately fight to stay alive. Meredith is relatively unscathed, while the rest have serious injuries: the pilot, Jerry (James LeGros), has a major spine injury, ... Shepherd is sucked out the side of the plane and awakens alone in the wood; his mangled hand having been pushed through the door of the plane. However, none are in as bad shape as **Lexie**, who is crushed under

Table 7: An example case on ”who” questions, drawn from retrieval results of DPR and DPR+PiCL. Answer is in **Bold**.

Among all question types, we see that DPR shows particularly low AARs on “where”, “what”, and “who” questions. These questions tend to include answers that refer to named entities, i.e. names of people, locations, and objects. Our hypothesis is that DPR fails to identify the presence of target entities, which serve as causal features and clues to answer passages. Table 7 shows an example from the retrieval result of illustrates the problem of named entities for DPR. While DPR is capable of retrieving passages with relevant semantics such as *aircraft crash, plane, and killing*, it fails to identify key named entities in the question such as *Greys Anatomy*. Contrary to a vanilla DPR, DPR+PiCL learns to differentiate answer passages from their counterfactual counterparts in which key entities are not present. In fact, we see from the example in Table 7 that DPR+PiCL is more inclined to identify answer-relevant entities given that the answer passage (*i.e.* top-9 passage retrieved by DPR) is ranked higher. In some sense, our approach is in line with previous methods based on salient span masking (Guu et al., 2020) such as MSS (Sachan et al., 2021), where the retriever is trained by predicting masked salient spans like named entities with the help of a reader.

C Choice of Pretrained Language Models

Previous studies on dense retrieval, including RocketQA (Qu et al., 2021), show that using a better PLM such as ERNIE (Sun et al., 2020) for retriever training leads to better performance. To study the effect of the choice of PLMs on the performance of retrievers trained with PiCL, we have implemented DPR+PiCL models with pretrained ERNIE and measured their performance on the NaturalQuestions benchmark.

Retriever	**PLM**	**# N**	**NaturalQuestions**			
			Top-1	Top-5	Top-20	Top-100
DPR+PiCL	BERT$_{\text{base}}$	47+48	45.38	66.79	78.53	85.04
DPR+PiCL	ERNIE$_{\text{base}}$	47+48	47.29	68.92	79.53	85.98
DPR+PiCL	ERNIE$_{\text{base}}$	127+128	**49.5**	**69.36**	**80.06**	**86.23**

Table 8: Retrieval performance with using a better PLM (i.e., ERNIE). PLMs: pre-trained LMs used for retriever initialization, # N: number of negative samples. Best scores are in **Bold**.

Based on table 8, we observe that the ERNIE implementation of DPR+PiCL outperforms its original BERT implementation, and that its performance increases with the number of negatives (i.e. in-batch and BM25 negative passages) used to train the model. While RocketQA does achieve a substantial performance gain over other baselines, it relies heavily on significantly large batches and a compute-intensive negative sampling pipeline, which makes it infeasible to reproduce the model due to tremendous computational costs.