Evidentiality-aware Retrieval for Overcoming Abstractiveness in Open-Domain Question Answering

Yongho Song[1*] **Dahyun Lee**[1*] **Myungha Jang**[1]
Seung-won Hwang[2] **Kyungjae Lee**[3] **Dongha Lee**[1] **Jinyoung Yeo**[1]

Yonsei University[1] Seoul National University[2] LG AI Research[3]
{kopf_yhs, leedhn, donalee, jinyeo}@yonsei.ac.kr
myunghajang@gmail.com seungwonh@snu.ac.kr
kyungjae.lee@lgresearch.ai

Abstract

The long-standing goal of dense retrievers in abtractive open-domain question answering (ODQA) tasks is to learn to capture evidence passages among relevant passages for any given query, such that the reader produce factually correct outputs from evidence passages. One of the key challenge is the insufficient amount of training data with the supervision of the answerability of the passages. Recent studies rely on iterative pipelines to annotate answerability using signals from the reader, but their high computational costs hamper practical applications. In this paper, we instead focus on a data-centric approach and propose Evidentiality-Aware Dense Passage Retrieval (EADPR), which leverages synthetic distractor samples to learn to discriminate evidence passages from distractors. We conduct extensive experiments to validate the effectiveness of our proposed method on multiple abstractive ODQA tasks.

1 Introduction

Information retrieval (IR) has served as a core component in open-domain question answering (ODQA) (Kwiatkowski et al., 2019; Joshi et al., 2017), which require the model to produce factually correct outputs based on a vast amount of knowledge in an unstructured text corpus. The predominant approach to ODQA employs the simple yet effective retriever-reader framework (Chen et al., 2017), where the retriever (*i.e.*, IR system) finds contexts that are relevant to the query from a large collection of texts, and the reader infers the final answer from the retrieved contexts. While augmenting the reader with a retriever is helpful when answerability aligns well with the relevance from the retriever, such an assumption does not always hold in abstractive ODQA tasks, *e.g.*, multi-hop QA (Yang et al., 2018), where target passages do not necessarily include the answer to the question.

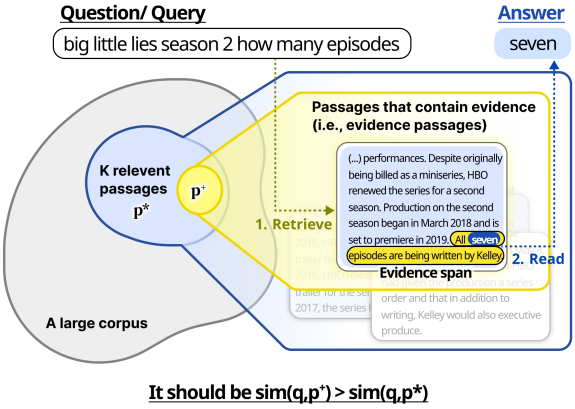


Figure 1: A bird's-eye view of the goal of passage retrieval in abstractive tasks. An ideal retriever (1) retrieves evidence passages such that the reader (2) produces answers based on the evidence span.

The misalignment between relevance and answerability in abstractive tasks poses a significant challenge to IR systems. The standard approach to building an IR system leverages human-annotated pairs of questions and relevant passages (Bajaj et al., 2016), but these IR datasets based on relevance provide only a weak supervision signal to abstractive tasks. This is particularly crucial for state-of-the-art IR systems, which train a dense passage retriever (DPR) (Karpukhin et al., 2020) using the relevance annotations to find relevant passages for a given question based on their learned vector representations. Training a dense retriever with such misaligned supervision leads to suboptimal performance in abstractive tasks, as the retriever fails to capture evidence passages from the corpus based on answerablility (Khattab et al., 2020; Tao et al., 2023).

One straightforward solution is to annotate the answerability of passages for questions. Recent studies (Izacard and Grave, 2021a; Sachan et al., 2021b; Izacard et al., 2022) rely model-centric approaches to obtain strong supervision for abstractive tasks. These methods utilize iterative pipelines

[*] Equal contribution

Findings of the Association for Computational Linguistics: EACL 2024, pages 1930–1943
March 17-22, 2024

that leverage fine-grained supervision signals from the reader to approximately measure the answerability of retrieved passages. However, these methods require exceptionally large computational resources, which hamper their application in practical scenarios.

Instead of pursuing such compute-intensive model-centric approaches, our work takes a step towards a data-centric approach, which aims to convert weak supervision from IR datasets into strong supervision signals for evidentiality-awareness. To this end, we present a data augmentation strategy where we augment strong distractor samples by removing evidence spans from gold evidence passages. Our strategy includes an effective approach that obtains pseudo-evidence using off-the-shelf QA model for datasets without gold annotations. We further propose Evidentiality-Aware Dense Passage Retriever (EADPR), a novel learning approach for dense retrieval that maximally leverages augmented distractor samples to integrate evidentiality-awareness into dense passage retrievers. In EADPR, our distractor passages as both hard negatives and pseudo-positives, as the model learns to discriminate evidence passages from strong distractors (*i.e.*, hard negatives) and distinguish between irrelevant and semantically relevant contexts (*i.e.*, pseudo-positives). Using these distractors as pivots between evidence and irrelevant passages, we aim at training an effective dense retriever that ranks evidence passages higher over distractor passages.

We evaluate EADPR across multiple ODQA tasks to show that our model leads to considerable improvement in retrieval and QA performance, and that our approach can be orthogonally applied with common strategies used to train advanced retrievers such as negative sampling (Xiong et al., 2021a; Qu et al., 2021). We also conduct extensive analysis on EADPR to show that our evidentiality-aware learning shows promise for robust, efficient approach to dense passage retrieval.

2 Preliminaries

A common approach to ODQA tasks usually involves utilizing external knowledge from a large corpus of texts to produce factually correct outputs (Chen and Yih, 2020). Due to the large search space in the corpus, a retriever is used in such settings 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-relevant passage pair (q_i, p_i^+) and a set of N negative passages p_j^-, DPR is trained to maximize the relevance measure (*e.g.*, the vector similarity) between the question q_i and its relevant passage p_i^+:

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

where $\langle q_i, p_i \rangle$ computes the relevance score between q_i and p_i as dot product between the question embedding $f_q(q_i)$ and the passage embedding $f_p(p_i)$ (*i.e.*, $\langle q_i, p_i \rangle = f_q(q_i) \cdot f_p(p_i)$).

Previous studies on dense retrieval have presented some straightforward strategies to further enhance the performance of DPR. One such approach is negative sampling (Xiong et al., 2021a; Qu et al., 2021), which exploits multiple retrievers to collect informative negative samples. While earlier work uses lexical retrievers such as BM25 (Robertson and Walker, 1994) for negative sampling, recent studies find that sampling hard negatives from fine-tuned encoders (Humeau et al., 2020) leads to more informative hard negative (Xiong et al., 2021a).

Despite these efforts, it still remains a challenge to train a dense retriever to the abstractive tasks. The main obstacle arises from the lack of large-scale data with strong annotations of evidentiality (Khattab et al., 2020; Prakash et al., 2021; Tao et al., 2023), *i.e.*, whether each of the passages contains evidence needed to answer the questions. To address this issue, recent studies (Izacard and Grave, 2021a; Sachan et al., 2021b; Izacard et al., 2022) employ an iterative pipeline that annotates evidentiality of the passages using supervision signals from the reader. However, using these complex model-centric approaches requires a significant amount of computing resources, which obstruct their deployment in various scenarios (Lindgren et al., 2021; Du et al., 2022; Gao et al., 2022). In this work, we instead study the validity of a data-centric approach to enhance the quality of IR datasets to obtain strong supervisions for passage retrieval from weak supervision in IR datasets.

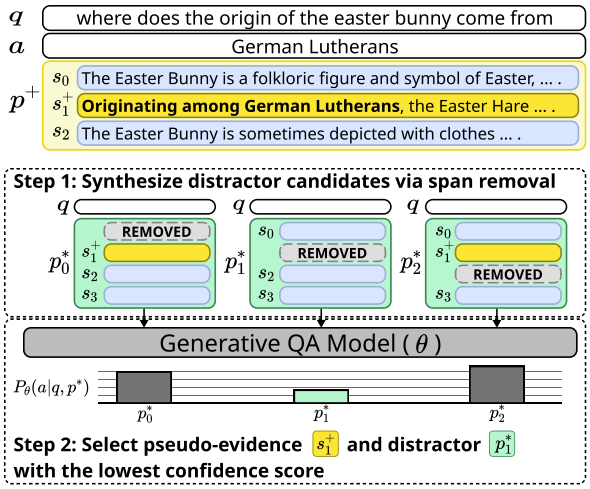


Figure 2: Illustration of pseudo-evidence annotation.

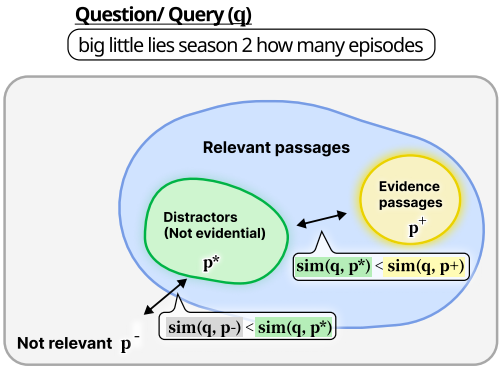


Figure 3: Conceptual overview of EADPR, where distractor samples serve as pivots between positive and negative passages.

3 Methodology

Our goal is to train a dense retriever capable of distinguishing evidence passages from distractor passages within a corpus. In this section, we propose Evidentiality-Aware Dense Passage Retrieval (EADPR), a novel learning approach for dense retrieval where the learned representation is conditioned on evidence spans (*i.e.*, *positive*) and invariant to evidentially-false contexts (*i.e.*, *negative*).

3.1 Augmenting Distractor Samples

An intuitive approach to synthesize distractor samples is to remove evidence spans from the gold evidence passage. Given a question-answer passage pair (q, p^+), where $p^+ = [s_l; s^+; s_r]$ contains an evidence span s^+ to the question q and evidentially-false spans s_l and s_r, we define our *distractor* sample p^* as a variant of p^+ such that $p^* = [s_l; s_r]$. We assume that such distractor samples are less evidential as they retain relevant semantics to the question but lack causal signals for question answering.

One problem in distractor augmentation is that some datasets do not include annotations of evidence spans, which are costly to obtain via human annotations. To address this issue, we follow the approaches from Lee et al. (2021) and incorporate pseudo-evidence annotations for distractor augmentation, as illustrated in Figure 2. Specifically, we employ an off-the-shelf generative question answering (QA) model θ that takes a question and a single evidence passage as inputs to generate the answer to the input question. For a given question q and its gold evidence passage p^+ with n discrete spans, we sample n distractor candidates $\{p_i^*\}_{i=1}^n$ by leaving out each of the n spans from p^+. Each distractor candidate p_i^* is then fed into the QA model with the question q to compute the confidence score $P_\theta(a|q, p_i^*)$. We choose the candidate p_i^* with the lowest confidence score as our distractor sample p^*, as a sharp drop in confidence score indicates that the i-th span is helpful in answering the question.

In practice, we adopt UnifiedQA-T5 (Khashabi et al., 2020) as QA model and select candidates with the highest perplexity, which is commonly used as the indicator of model confidence.

3.2 Evidentiality-aware Learning

We aim to train a retriever to learn representions of questions and passages conditioned on their evidentiality such that the retriever ranks evidence passages higher than other distractor passages. Our design is based on the intuition that our distractor sample, denoted as p^*, serves as both a hard negative and pseudo-positive, as distractor passages are still relevant to the question. Essentially, we model the space that is relevant but not evidential as a middle pivot point between the relevant space and the irrelevant space, as illustrated in Figure 3.

Distractors as Hard Negatives. Our distractor samples are designed to be less evidential, meaning that its content is relevant but doesn't contain the actual information for the question. As our goal is to learn a representation that reflects the evidentiality, we use these distractor samples as hard negatives. Specifically, we consider p_i^* as a hard *negative* sample to an anchor question q_i while the original passage p_i^+ serves as the *positive*. Thus the embedding similarity $\langle q_i, p_i^* \rangle$ between q_i and p_i^* is upper bounded by $\langle q_i, p_i^+ \rangle$:

$$\langle q_i, p_i^+ \rangle > \langle q_i, p_i^* \rangle \quad (2)$$

Following this observation, we define *Distractors-as-Hard-Negative* loss, $\mathcal{L}_{\texttt{HN}}$, to maximize the similarity between q_i and p_i^+ while minimizing the similarity between q_i and p_i^*.

$$\mathcal{L}_{\texttt{HN}}(q_i, p_i^+, p_i^*) = -\log \frac{e^{\langle q_i, p_i^+ \rangle}}{e^{\langle q_i, p_i^+ \rangle} + e^{\langle q_i, p_i^* \rangle}} \quad (3)$$

By learning to discriminate p_i^* from p_i^+, the model learns to minimize the mutual information between representations of questions q_i and evidentially-false spans in p_i^+, strengthening causal effects of evidence spans in the learned embeddings.

Distractors as Pseudo Positives. However, it is not sufficient to solely consider our synthetic distractors as hard negatives. Since these samples still hold relevance, our objective is to rank them lower than evidence passages but higher than irrelevant ones. While distractor samples serve as hard negatives in relation to evidence passages, they can be seen as positive samples in comparison to irrelevant ones. We refer to these samples as pseudo-positives, as 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^-:

$$\langle q_i, p_i^* \rangle > \langle q_i, p_j^- \rangle \quad (4)$$

To incorporate this, we derive *Distractors-as-Pseudo-Positives* loss, $\mathcal{L}_{\texttt{PP}}$, where the model maximizes the relative similarity between q_i and p_i^* with respect to negative passages p_j^- and p_j^* in the given batch.

$$\mathcal{L}_{\texttt{PP}}(q_i, p_i^*, \{p_j^-, p_j^*\}_{j\neq i}^N) = -\log \frac{e^{\langle q_i, p_i^* \rangle}}{e^{\langle q_i, p_i^* \rangle} + \sum_{j\neq i}^N \left(e^{\langle q_i, p_j^- \rangle} + e^{\langle q_i, p_j^* \rangle}\right)} \quad (5)$$

Essentially, the model learns to discriminate three relevancy space check among evidential, evidentially-false, and irrelevant passages, as illustrated in Figure 3.

Evidentiality-aware DPR. From Equation 2 and 4, we can derive that the embedding similarity $\langle q_i, p_i^* \rangle$ between questions and distractor samples are bounded by $\langle q_i, p_i^+ \rangle$ and $\langle q_i, p_j^- \rangle$. Hence, they can be re-formulated as *pivots* between positive and negative samples in the embedding space. Note that our definition of distractor samples as pivots is in line with the objective of DPR, since both inequality constraints in Equation 2 and 4 combined satisfy the below constraint in Equation 1:

$$\langle q_i, p_i^+ \rangle > \langle q_i, p_j^- \rangle \quad (6)$$

Building on top of the above idea, our training objective combines all losses from Equation 3 and 5 with the training loss in Equation 7. To adapt DPR training into our setting, we further define $\mathcal{L}_{\texttt{dpr}}$ as a slight modification of DPR training objective where the distractor p_i^* to the evidence passage p_i^+ is added as a negative:

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

where $\lambda < 1$ is a hyperparameter used to balance the effect from counterfactual passages as negatives in DPR training. The final loss function $\mathcal{L}_{\texttt{eadpr}}$ is a weighted sum of all losses $\mathcal{L}_{\texttt{dpr}}$, $\mathcal{L}_{\texttt{HN}}$, and $\mathcal{L}_{\texttt{PP}}$:

$$\mathcal{L}_{\texttt{eadpr}} = \mathcal{L}_{\texttt{dpr}} + \tau_1 \mathcal{L}_{\texttt{HN}} + \tau_2 \mathcal{L}_{\texttt{PP}} \quad (8)$$

where τ_1, τ_2 are hyperparameters that determine the importance of the terms. See Appendix C for details on hyperparameters.

Dataset	Train	Dev	Test	Corpus
NQ	58,880	8,757	3,610	
TQA	57,369	8,837	11,313	21,015,324
TREC	1,125	133	694	
HotpotQA	180,890	7,405	-	5,233,329

Table 1: Statistics of datasets used in this paper. Train, Dev, and Test represent the size of train sets, dev sets, and test sets, respectively. Corpus indicates the number of passages in the source corpus.

4 Experiments

4.1 Experimental Settings

Dataset. For our experiments we consider two categories of ODQA datasets, single-hop and multi-hop datasets. Single-hop datasets require the model to capture evidence that is not evidently given in the set of retrieved passages. The role of EADPR is to discriminate answer passages from distractor passages such that the answer passages are among the top-k relevant contexts. Following Karpukhin et al. (2020), we choose Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA

Training Strategies	Retriever	NQ			TQA			TREC		
		Top-1	Top-20	MRR	Top-1	Top-20	MRR	Top-1	Top-20	MRR
Vanilla Training	DPR	31.8	74.8	43.1	38.7	74.7	49.3	-	-	-
	EADPR	35.4	76.8	46.4	43.0	74.7	52.4	31.1	79.8	45.5
+ BM25 Negative	DPR	46.6	79.7	56.0	54.3	79.7	62.0	-	79.8†	-
	EADPR	48.6	80.1	57.6	**56.9**	**80.5**	**63.9**	46.8	**83.9**	**58.1**
+ Negative Mining	DPR	52.7	81.4	61.2	54.2	78.2	61.3	-	-	-
	EADPR	**54.0**	**82.6**	**62.4**	54.1	78.0	61.2	-	-	-

Table 2: Passage retrieval results on single-hop QA datasets, *i.e.* NQ, TriviaQA, and TREC. Top-k hit accuracy and MRR scores are reported, and the best results are marked as **bold**. † indicates the performance of the baseline DPR is reported in (Karpukhin et al., 2020).

Reader	Training Strategies	Retriever	Exact Match (EM) score		
			Top-5 passages	Top-20 passages	Top-100 passages
DPR Reader	Vanilla training	DPR	31.83	36.87	37.45
		EADPR	34.27 (+2.44)	38.86 (+1.99)	39.06 (+1.61)
FiD_{base} (T5)	Vanilla Training	DPR	31.99	39.11	43.82
		EADPR	34.27 (+2.28)	41.47 (+2.36)	44.85 (+1.03)
FiD_{base} (T5)	+ Negative Mining	DPR	38.31	43.13	45.37
		EADPR	**40.22** (+1.91)	**44.32** (+1.19)	**47.65** (+2.28)

Table 3: End-to-end QA performance of retriever-reader on Natural Questions. Top-k indicates the number of top retrieved passages used for reader inference. We reuse the checkpoints of DPR reader and FiD_{base} (*i.e.* T5-base implementation of FiD) from Karpukhin et al. (2020) and Izacard and Grave (2021a). Best scores are in **Bold**.

(TQA) (Joshi et al., 2017), and TREC (Baudiš and Šedivý, 2015) for evaluation and use the Wikipedia corpus of 21M passages as source passages.

On the other hand, a multi-hop QA dataset contains questions whose answers cannot be extracted from a single answer passage. We aim to assess whether the retrievers are capable of finding all evidence passages in the corpus such that the reader can derive answers by aggregating evidence from the passage set. Specifically, we evaluate our approach on HotpotQA (Yang et al., 2018) under the *full-wiki* setting, which uses the corpus of 5.2M preprocessed passages from Wikipedia for evaluation. See Appendix B for more details on datasets. Table 1 summarizes the statistics of the datasets used in this paper.

Retriever Training Strategies. We adopt DPR (Karpukhin et al., 2020) as the backbone architecture for all retrievers implemented in this section. Our focus is to assess how applying EADPR affects the performance of the backbone DPR and whether EADPR is orthogonal to conventional approaches for retriever training.

One popular data augmentation approach to enhance DPR involves negative mining (Xiong et al., 2021a; Qu et al., 2021), which adopts additional retrievers to augment the train set with more informative negatives for retriever training. In our experiments, we first consider using BM25 (Robertson and Walker, 1994) to sample hard negatives based on lexical matching. We then follow ANCE (Xiong et al., 2021a) and mine hard negatives from previous retriever checkpoints. For both cases, we mine one negative sample per query from top retrieved results of the retriever. We provide more details on our implementations in Appendix C.

4.2 Single-hop QA Benchmarks

Retrieval Performance. Table 2 compares the performance of EADPR models with the baselines on single-hop QA benchmarks. We observe that EADPR models yield consistent performance gains over vanilla DPR under all tested conditions. Similar to DPR, EADPR shows stronger performance when trained with hard negatives, which suggests that adding informative negative samples further boosts the discriminative power of EADPR. In the case of TriviaQA, we hypothesize that both retrievers trained on TriviaQA fail to deliver high-quality negative samples since the models are trained on the TriviaQA train set that contains false positive annotations (Li et al., 2023). Overall, the performance gain on EADPR implies that EADPR can be

	R@2	R@10	R@20
Single-hop Retrieval			
DPR†	25.2	45.4	52.1
EADPR	29.4	48.5	53.5
Multi-hop Retrieval			
MDR+DPR	47.7	61.0	65.7
MDR+EADPR	**58.5**	**68.1**	**71.8**

Table 4: Retrieval performance of EADPR on HotpotQA. R@k indicates the proportion of questions where all annotated supporting contexts are included in top-k retrieval results. † denotes the reported performance in Xiong et al. (2021b).

Retriever	Answer	Support	Joint
MDR+DPR	61.0	61.5	50.2
MDR+EADPR	**66.1**	**68.2**	**56.4**

Table 5: Reader performance on HotpotQA dev set. We report F1 scores of the ELECTRA reader given 20 supporting contexts, which are much fewer than 100 contexts used in Xiong et al. (2021b).

further improved when orthogonally applied with common training strategies for dense retrieval.

End-to-End QA Performance. To assess the effect of EADPR on QA performance, we pair EADPR into a QA system and evaluate the performance of the subsequent reader. Specifically, we re-use two reader models, an extractive reader from Karpukhin et al. (2020) and a Fusion-in-Decoder (FiD) from Izacard and 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 the QA performance of the retriever-reader pipelines. Overall, EADPR consistently improves the QA performance of different readers over DPR, suggesting that EADPR benefits the subsequent readers.

4.3 Multi-hop QA Benchmark

We evaluate our approach on HotpotQA (Yang et al., 2018) to assess whether EADPR better capture key evidence in multi-hop QA settings, where passages contain implicit evidence rather than answer exact match. Table 4 compares the performance of DPR and EADPR implemented for single-hop and multi-hop retrieval. For our multi-hop retrievers, we follow Xiong et al. (2021b) and implement multi-hop dense retriever (MDR) using EADPR. We use MDR models to produce 20 candidate contexts and feed them into an ELECTRA reader (Clark et al., 2020). Details on multi-hop baselines are included in Appendix C.

Table 4 shows that EADPR shows higher R@k than a vanilla DPR even without applying MDR, suggesting that our evidentiality-aware training improves the model's ability to capture key evidence without attending to exact answer match. We also observe that incorporating EADPR into MDR leads to considerable performance gain over the standard MDR implemented using DPR, and that such gain in retrieval performance leads to improvement in QA performance, as shown in Table 5. Full results are shown in Table 10.

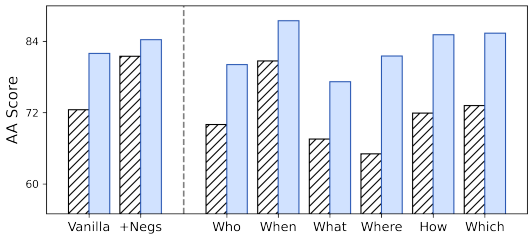
5 Analysis and Discussion

Answer Awareness. To see how EADPR achieves such improvement, we conduct a fine-grained analysis to measure the model's capability of capturing evidence spans. For this purpose, we introduce an additional analytic metric, named Answer-Awareness (AA) score, by measuring how frequently the model deems an answer-masked passage more relevant than its original passage. Formally, given a held-out set of T pairs (q_i, p_i^+) with gold answer annotations, we construct answer-masked passages p_i' by removing exact answer spans from p_i^+. AA score of a retriever is then computed as the proportion of (q_i, p_i^+, p_i') triplets where relevance scores $\langle q_i, p_i^+ \rangle$ are higher than the scores $\langle q_i, p_i' \rangle$ of answer-masked passages:

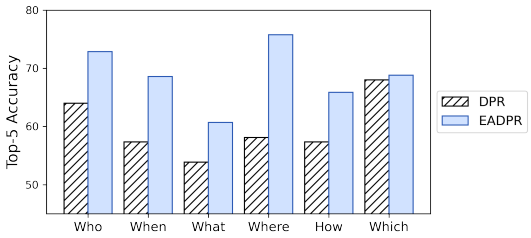
$$\text{AA score} = 1 - \sum_{i=1}^{T} \mathbb{1}_{\langle q_i, p_i^+ \rangle \leq \langle q_i, p_i' \rangle} / T \quad (9)$$

where $\mathbb{1}_{\langle q_i, p_i^+ \rangle \leq \langle q_i, p_i' \rangle}$ is an indicator if $\langle q_i, p_i^+ \rangle$ is smaller than $\langle q_i, p_i' \rangle$. To measure AA score, we reuse the 1,382 gold (q, p^+, p') triplets from NQ test set used in Section 5.

Figure 4 compares the AA score of EADPR with DPR trained under the same conditions (*i.e.*, training strategies). We first observe that AA scores of a vanilla DPR significantly fall behind the theoretical upper bound, which indicates that the relevance measurement learned in DPR may not effectively capture the evidentiality-awareness such that retrievers constantly rank positive passages higher than counterfactual passages. While common training strategies such as negative sampling lead to some increase in AA scores, there is still



(a) Answer Awareness



(b) Retrieval Top-5 Accuracy

Figure 4: AA score and retrieval top-5 hit accuracy on various question types. Note that we feed retrieved passages from DPR or EADPR on the same DPR reader for inference.

Question: Who died in the plane crash greys anatomy	
Passage Type	**Text**
Gold passage	Flight (Grey's Anatomy) ... American television medical drama Grey 's Anatomy ... who are victims of an aviation accident fight to stay alive , but **Dr. Lexie Grey** (Chyler Leigh) ultimately dies. ...
DPR Top-1	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. ...
DPR Top-9	Flight (Grey's Anatomy).. 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 ...
EADPR Top-1	Comair Flight 5191. after the crash to create an appropriate memorial for the victims, first responders, ... suffered serious injuries, including multiple broken bones, a collapsed lung, and severe bleeding. ...
EADPR Top-2	Flight (Grey's Anatomy)... 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 6: An example case on 'who' questions from results of DPR and EADPR. Answers are in **Bold**.

substantial room for improvement towards building an evidentiality-aware retriever. One the other hand, EADPR brings further gain in AA scores, showing that our data-centric approach is effective in enhancing evidentiality-awareness.

In Figure 4, we further break down the the gold (q, p^+, p') triplets with respect to their question types and measure AA scores of DPR and EADPR on subsets of test samples of different question types, *i.e.*, who, when, what, where, how, and which. Overall, we see that AA scores of DPR vary significantly across different question types, ranging from 65.07% to 80.68%. On the other hand, EADPR achieves significant improvements in AA scores for all question types and consistently shows better retrieval performance.

Among all question types, we see that DPR shows particularly low AA scores on who-, what-, and where-questions, whose answers tend to refer to named entities, *i.e.*, names of people, locations, and objects. Our hypothesis is that DPR often fails to identify the presence of target entities, which serve as causal features in evidence passages. Table 6 shows an example of the retrieval results, illustrating the problem of named entities for DPR. While DPR is capable of retrieving passages with relevant semantics such as aircraft crash, it fails to identify key named entities in the question such as Greys Anatomy. In contrast, we observe EADPR ranks evidence passages with key entities higher than DPR (*i.e.*, Top-2 from EADPR compared to Top-9 from DPR), suggesting that EADPR learns to differentiate evidence passages from their distractors in which key entities are absent. In some sense, our approach is in line with previous methods based on salient span masking (Guu et al., 2020; Sachan et al., 2021a), where the retriever is trained to predict masked salient spans with the help of a reader.

Robustness. We have assumed that EADPR learns to discriminate between evidence and distractor passages. To validate this assumption, we perform a simulation test in which we synthesize and add distractor passages into the corpus to measure the robustness of EADPR to these samples. This scenario is often encountered in real-world corpora, where there is a surplus of passages with similar contexts but lack definitive evidence (Spirin and Han, 2012; Pan et al., 2023; Goldstein et al., 2023).

Specifically, we create plausible distractor passages using a large language model (*i.e.*, ChatGPT).

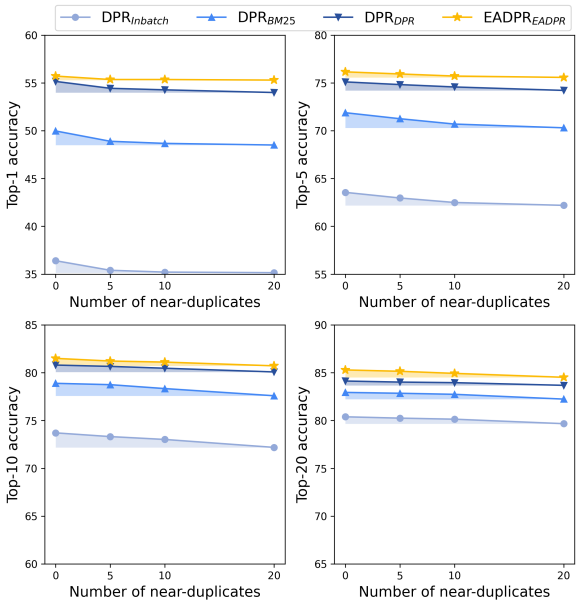
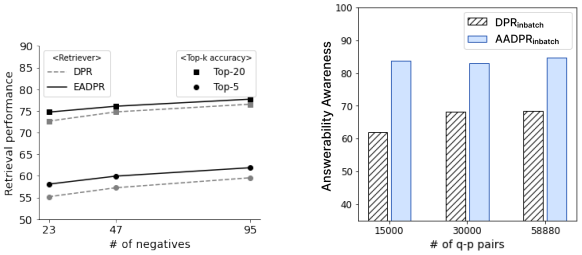


Figure 5: Retrieval accuracy at Top-k with varying number of near-duplicates on Natural Questions dataset. Colored area illustrates the degree of the performance drop.



(a) Number of negatives (b) Training samples

Figure 6: Performance of DPR and EADPR on varying numbers of (a) negatives and (b) training samples.

The model is prompted to generate *near-duplicate* samples for each query that mimic the context of evidence passage but leave out the key evidence. We collect these near-duplicates for 1,382 test queries from NQ with annotated evidence passages and include at most 20 near-duplicates per query.

Figure 5 shows the performance of the dense retrievers on text corpora with varying number of near-duplicates. We observe that the retrieval performance (*i.e.*, Top-k accuracy) decreases substantially when given more near-duplicates, indicating that dense retrievers are vulnerable to the presence of distractor passages. On the other hand, we observe that EADPR is relatively robust against the effect from additional distractor samples, showing promise for robust passage retrieval on a noisy real-world corpus.

Retriever	Natural Questions			
	Top-1	Top-5	Top-20	Top-100
$\mathcal{L}_{\texttt{dpr}}$	31.77	58.12	74.76	84.07
$+\mathcal{L}_{\texttt{PP}}$	32.08	58.64	75.32	83.82
$+\mathcal{L}_{\texttt{HN}}$	31.85	59.53	75.57	84.43
$+\mathcal{L}_{\textbf{PP}} + \mathcal{L}_{\textbf{HN}}$	35.35	61.55	76.81	85.87

Table 7: Ablation studies on the training objective.

Resource and Label Efficiency. We posit that the benefit of EADPR lies in the label efficiency, as counterfactual samples serve as both hard negatives and pseudo-positives in EADPR. To validate this assumption, we train EADPR with fewer (a) negative samples and (b) training instances. Figure 6a shows that EADPR trained with fewer negatives (*e.g.*, 23) yields performance comparable to a vanilla DPR trained with more negatives (*e.g.*, 47). Meanwhile, we see in Figure 6b that EADPR consistently shows higher AA score over DPR when using fewer training samples (*e.g.*, 15k and 30k). These findings support our assumption on the efficiency of EADPR.

Effect of Counterfactual Samples as Pivots. We conduct an ablation studies on the learning objective in Equation 8 to study the effect of using counterfactual samples as pivots (*i.e.*, both pseudo-positives and hard negatives) on DPR training. Specifically, we consider the following modifications to the objective function, 1) $\mathcal{L}_{\texttt{dpr}} + \mathcal{L}_{\texttt{PP}}$, and 2) $\mathcal{L}_{\texttt{dpr}} + \mathcal{L}_{\texttt{HN}}$. Table 7 compares all baselines with EADPR and DPR. We find that all modifications do not bring much improvement to DPR without either $\mathcal{L}_{\texttt{PP}}$ or $\mathcal{L}_{\texttt{HN}}$. In contrast, EADPR consistently outperforms DPR and all its variants, suggesting that using counterfactual samples as pivots is crucial in EADPR.

6 Related Work

Dense Retrieval. Dense retrieval aims at retrieving information based on semantic matching by mapping questions and contexts into a learned embedding space (Karpukhin et al., 2020; Lee et al., 2019). Earlier attempts to enhance dense retrievers have drawn inspiration from studies on learning to rank (Liu, 2009), improving the performance of dual encoders via methods such as negative sampling (*e.g.*, ANCE (Xiong et al., 2021a) and RocketQA (Qu et al., 2021)). More recent approaches are founded upon knowledge distillation (Hinton

et al., 2015), which constructs an iterative pipeline of retrievers and readers such that the retriever learn from the reader's predictions on the evidentiality of passages (*e.g.*, cross-attention in Izacard and Grave (2021a), model confidence in ATLAS (Izacard et al., 2022) and REPLUG (Shi et al., 2023)).

Counterfactual learning in NLP. Counterfactual learning has been a useful tool in enhancing the robustness and fairness in representation learning by attending to causal features (Johansson et al., 2016; Feder et al., 2022). These studies define counterfactual intervention based on causal features and train models using counterfactual samples, which are minimally dissimlar but lead to different (*i.e.*, counterfactual) outcome (Chen et al., 2020; Choi et al., 2020, 2022). By learning from counterfactual samples, these approaches aim to build models that rely more on causal relationship between observations and labels. Our work stems from this line of research, as we introduce assumptions on causal signals in passage retrieval for knowledge-intensive task.

7 Conclusion

In this work, we address the misalignment problem in dense retrievers for abstractive QA tasks, where relevance supervisions from IR datasets are not well-aligned with answerability of passages for questions. To overcome the abstractiveness of ODQA tasks, we present EADPR, which augments distractor samples to train an evidentiality-aware retriever by learning to distinguish between evidence and distractor samples. Our experiments show promising results in many ODQA tasks, indicating that EADPR not only enhances model performance on both retrieval and downstream tasks but also improves robustness to distractors.

Limitations

Below we summarize some limitations of our work and discuss potential directions to improve it: (i) Our definition of causal signals in answerable passages has been limited to answer sentences that contain exact matches of gold answers. While simple and efficient, our counterfactual sampling strategy leaves room for improvement, and more elaborate construction methods would lead to better counterfactual samples and further enhance the performance of EADPR. (ii) We observe that AA scores in Section 5 are not well calibrated with the downstream performance of the retriever, which limits the practical usefulness of AA score as an indicator of the model performance. In future work, we aim to refine the definition of AA score such that it serves as a formal evaluation metrics for dense retrieval.

Broader Impact and Ethics Statement

Our work re-examines the evidentiality-awareness of the dense retrievers and seeks to mitigate undesired model biases to false positives, or contexts in candidate passages with no evidence. While we have focused solely on the effectiveness of our approach on ODQA, we believe that the concept of distractor samples as pivots can be further explored in other representation learning tasks such as response retrieval for dialogue systems.

Meanwhile, our work shares the typical risks towards misinformation from common dense retrieval models (Qu et al., 2021; Santhanam et al., 2022) as our implementation follows the common design based on dual encoders. Our work takes a step towards minimizing such risks from the retriever, but we note that there is still much work needed from the community to ensure the faithfulness of dense retrievers, particularly in specialized domains with insufficient data.

Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)) and (No.2021-0-02068, Artificial Intelligence Innovation Hub) and (No.2022-0-00077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data). Jinyoung Yeo is a corresponding author.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *Proceedings of ICLR*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary,

and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiV preprint*.

Petr Baudiš and Jan Šedivý. 2015. Modeling of the question answering task in the yodaqa system. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of ACL*.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of ACL*.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of CVPR*.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. 2022. C2l: Causally contrastive learning for robust text classification. In *Proceedings of AAAI*.

Seungtaek Choi, Haeju Park, Jinyoung Yeo, and Seung-won Hwang. 2020. Less is more: Attention supervision with counterfactuals for text classification. In *Proceedings of EMNLP*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *Proceedings of ICML*.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*.

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

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiV preprint*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of ICML*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of ICLR*.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *Proceedings of ICLR*.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of EACL*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot Learning with Retrieval Augmented Language Models. *arXiv preprint*.

Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *Proceedings of ICML*.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of ACL*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.

D. Khashabi, S. Min, T. Khot, A. Sabhwaral, O. Tafjord, P. Clark, and H. Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings of EMNLP*.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2020. Relevance-guided supervision for openqa with colbert. *Transactions of the Association for Computational Linguistics*.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, pages 452–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of ACL*.

Kyungjae Lee, Seung-won Hwang, Sang-eun Han, and Dohyeon Lee. 2021. Robustifying multi-hop QA through pseudo-evidentiality training. In *Proceedings of ACL*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of ACL*.

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

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models.

Prafull Prakash, Julian Killingback, and Hamed Zamani. 2021. Learning robust dense retrieval models from incomplete relevance labels. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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 NAACL*.

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*.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94.

Devendra 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 *Proceedings of ACL-IJCNLP*.

Devendra Singh Sachan, Siva Reddy, William 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*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of NAACL*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiV preprint*.

Nikita Spirin and Jiawei Han. 2012. Survey on web spam detection: Principles and algorithms. *SIGKDD Explor. Newsl.*

Chongyang Tao, Jiazhan Feng, Tao Shen, Chang Liu, Juntao Li, Xiubo Geng, and Daxin Jiang. 2023. CORE: Cooperative training of retriever-reranker for effective dialogue response selection. In *Proceedgins of ACL*.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance-level discrimination. In *Proceedings of CVPR*.

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

Wenhan Xiong, Xiang Lorraine Li, Srinivasan Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2021b. Answering complex open-domain questions with multi-hop dense retrieval. In *Proceedings of ICLR*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Retriever	Top-1	Top-20	MRR
DPR			
- 20 epochs	41.5	78.3	52.6
- 40 epochs	46.6	79.7	56.0
- 80 epochs	46.8	79.5	56.5
EADPR (40 epochs)	48.6	80.1	57.6

Table 8: Ablation studies on the number of training epochs. Specifically, we compare EADPR with DPR checkpoints trained over different training epochs. All models are trained using one additional BM25 negative.

Retriever	Top-1	Top-20	MRR
DPR	31.8	74.8	43.1
+ 1 BM25 Neg	46.6	79.7	56.0
+ 2 BM25 Neg	45.5	79.4	55.4
EADPR (40 epochs)	35.4	76.8	46.4
+ 1 BM25 Neg	48.6	80.1	57.6

Table 9: Ablation studies on the number of negatives samples used to train DPR and EADPR.

A Additional Ablation Studies

More training iterations. One possible hypothesis behind the performance gain from EADPR is that the model benefits from more occurrences of positive samples during training, as EADPR uses one additional sample per instance (*i.e.*, p^+ and p^*). To see whether the performance gain indeed comes from more training iterations of positive samples, we additionally train the baseline DPR for more epochs and measure the change in performance on NQ as the training epoch doubles. Table 8 shows that adding more training epochs (from 40 to 80) does not lead to significant performance gain in DPR, suggesting that the performance improvement in EADPR does not come from more training iterations of positive samples.

More negative samples. Another hypothesis is that the model benefits from more negative samples used during training (i.e. p^- and p^*). To test this hypothesis, we compare the performance of EADPR with the baseline DPR trained using the same number of negatives per instance as EADPR. We observe in Table 9 that increasing the number of hard negatives used for DPR training does not increase the model performance on NQ. This is in line with the observation from (Karpukhin et al., 2020) that DPR does not benefit much from additional hard negatives. On the other hand, we see that EADPR trained using one negative (p^-) and one counterfactual sample (p^*) outperforms DPR trained with two negative samples (p^-) per instance, suggesting that the performance gain in EADPR cannot be solely attributed to more negative samples used for training.

B Datasets

Single-hop QA. All of the ODQA datasets used in this paper, *i.e.* NaturalQuestions and TriviaQA, cover Wikipedia articles written in English. Specifically, the Wikipedia corpus used in this paper is collected from English Wikipedia dump from Dec. 20, 2018, as described in Karpukhin et al. (2020). Demographics of the authors do not represent any particular group of interest for both datasets. Details on the data collection can be found in Kwiatkowski et al. (2019) and Joshi et al. (2017). We obtain hard negatives from the dataset provided by Karpukhin et al. (2020), which is available on https://github.com/facebookresearch/DPR.

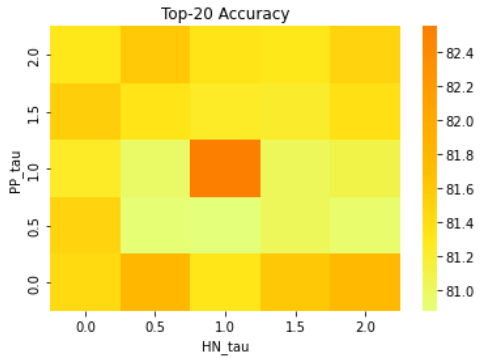
Multi-hop QA. We train our models with the train set from Yang et al. (2018) and evaluate them on the Wikipedia corpus of 523,332 passages. The corpus is constructed from the dump of English Wikipedia of October 1, 2017, and steps to preprocess Wikipedia documents are described in Yang et al. (2018). Similar to single-hop QA datasets, HotpotQA dataset does not include documents where demographics of the authors do not represent any particular group of interest.

C Implementation Details

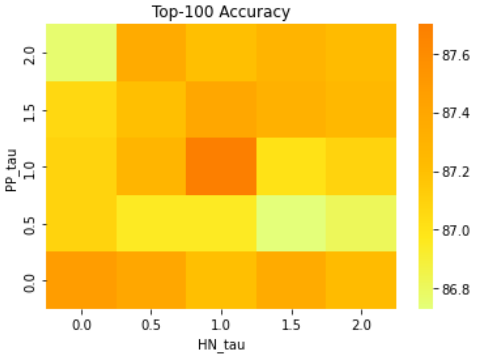
Dense Retrievers. Our implementations of dense retrievers follow the dual encoder framework of DPR (Karpukhin et al., 2020), where each encoder adopts BERT-base (Devlin et al., 2019) (110M parameters) as the base architecture. For experiments on ODQA benchmarks in Section 4.2, we train all implemented 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 EADPR 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 and TriviaQA benchmarks 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 1.0 for the balancing coefficient λ for counterfactual samples in Equation 7. The weight hyperparameters τ_1, τ_2 in Equation 8 are set as 1.0. We find the

Retriever	Answer		Support		Joint	
	EM	F1	EM	F1	EM	F1
MDR+DPR	49.7	61.0	41.1	61.5	30.9	50.2
MDR+EADPR	**54.5**	**66.1**	**47.1**	**68.2**	**35.5**	**56.4**

Table 10: Reader performance on HotpotQA dev set. The QA performance is measure based on Exact Match (EM) and F1 scores of answers (Answer EM/F1), supporting sentences (Support EM/F1), and both (Joint EM/F1).



(a) Top-20 accuracy



(b) Top-100 accuracy

Figure 7: (a) Top-20 and (b) top-100 accuracy EADPR trained on NQ with different τ_1 and τ_2.

best hyperparameters for τ_1, τ_2 using grid search. Figure 7 shows the performance of EADPR trained with different combinations of τ_1, τ_2.

Readers. For reader in single-hop QA experiments, 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 and 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. For all experiments, we conducted a single run of each model tested. Our empirical findings showed little variance in the results over multiple runs.

For reader in multi-hop QA experiments, we use the extractive ELECTRA (Clark et al., 2020) reader provided in Xiong et al. (2021b). Reader inference is conducted on a single 24GB GPU with the number of input contexts limited to 20. For all experiments, we conducted a single run of each model tested. Our empirical findings showed little variance in the results over multiple runs.

Multihop Dense Retrieval. The classic approaches to multi-hop QA usually involve decomposing questions into multiple subquestions, retrieving relevant contexts for each subquestion, and aggregating multiple contexts into a reasoning path (Asai et al., 2020). In line with these studies, Xiong et al. (2021b) train a Multihop Dense Retrieval (MDR) to construct reasoning paths by performing dense retrieval in multiple hops, each time with query representations augmented using the retrieved passages. MDR is paired with a reader that takes reasoning paths as inputs, and the QA performance is measured based on Exact Match (EM) and F1 scores of answers (Answer EM/F1), supporting sentences (Support EM/F1), and both (Joint EM/F1).

We implement MDR using EADPR following Xiong et al. (2021b) but with some constraints due to limited computing resources: (1) we train our models on smaller batch sizes of 120 compared to 150 in the original paper; (2) our MDR implementation is not optimized using the memory bank mechanism (Wu et al., 2018); (3) we generate 20 candidate reasoning paths (*i.e.*, beams) instead of 100 in the original paper. Table 10 reports in detail the QA performance of the reader when paired with different MDR.

Software Packages. We use NLTK (Bird et al., 2009) [1] and SpaCy [2] for text preprocessing. Following DPR, we adopt FAISS (Johnson and Douze, 2019), an approximate nearest neighbor (ANN) indexing library for efficient search, in our implementation of EADPR. DPR also uses an open-sourced

[1] https://www.nltk.org/
[2] https://spacy.io/

library for logging and configuration named Hydra [3], which we use to configure our experiments. No modification has been made to the aforementioned packages.

Terms and License. Our implementation of EADPR is based on the public repository of DPR [4], which is licensed under Creative Commons by CC-BY-NC 4.0. The indexing library FAISS is licensed by MIT license. Both ODQA datasets, NaturalQuestions and TriviaQA, are licensed under Apache License, Version 2.0. We have confirmed that all of the artifacts used in this paper are available for non-commercial, scientific use.

[3] https://github.com/facebookresearch/hydra

[4] https://github.com/facebookresearch/DPR