Multi-stage Training with Improved Negative Contrast for Neural Passage Retrieval

Jing Lu[*], **Gustavo Hernández Ábrego, Ji Ma, Jianmo Ni, Yinfei Yang**
Google Research
{ljwinnie, gustavoha, maji, jianmon, yinfeiy}@google.com

Abstract

In the context of neural passage retrieval, we study three promising techniques: synthetic data generation, negative sampling and fusion. We systematically investigate how these techniques contribute to the performance of the retrieval system and how they complement each other. We propose a multi-stage framework comprising of pre-training with synthetic data, fine-tuning with labeled data and negative sampling at both stages. We study six negative sampling strategies and apply them to the fine-tuning stage and, as a noteworthy novelty, to the synthetic data that we use for pre-training. Also, we explore fusion methods that combine negatives from different strategies. We evaluate our system using two passage retrieval tasks for open-domain QA and using MS MARCO. Our experiments show that augmenting the negative contrast in both stages is effective to improve passage retrieval accuracy and, importantly, they also show that synthetic data generation and negative sampling have additive benefits. Moreover, using fusion of different kinds allows us to reach performance that establishs a new state-of-the-art level in two of the tasks we evaluated.

1 Introduction

Recently, there is a surge of interest in neural first-stage retrieval models (Yang et al., 2020; Guo et al., 2021). These models overcome the lexical gap issue of traditional models based on term matching (Robertson and Zaragoza, 2009) by projecting both query and document to a shared dense space. Finding relevant documents can then be achieved by employing nearest neighbor search. Neural first-stage retrieval models have shown competitive on many benchmark data sets (Karpukhin et al., 2020; Xiong et al., 2021; Qu et al., 2021), and combining them with term matching-based models further boosts their retrieval performance (Bendersky et al., 2020).

Arguably, abundance of training data and negative sampling strategies have been the two most important factors to the success of neural retrieval models. On one hand, deep Neural Networks are data hungry due to their vast volume of model parameters. Ma et al. (2021) has shown that synthetic question generation can be effective to mitigate the data scarcity issue in low-resource settings. In this work we are interested in exploring how synthetic question generation can further improve the neural retrieval models when there is already a decent amount of supervised data available. We propose a two-stage training strategy where we first train the dense retrieval model on synthetic question-passage pairs and then, as illustrated in Fig 1, we fine-tune it on supervised data. We show that such methodology substantially improves upon baseline models.

On the other hand, previous works (Karpukhin et al., 2020; Xiong et al., 2021) found that utilizing extra negatives in addition to in-batch negatives significantly improves the performance of dense retrieval models. Here, we first draw a connection between the cross-entropy loss with in-batch negatives and Noise Contrastive Estimation (Ma and Collins, 2018), and highlight the limitations of in-batch negative sampling. Then, we extensively study the impact of several negative sampling strategies on model accuracy and propose ways to effectively combine them.

In addition to investigating synthetic question generation and negative sampling independently, another research question we explore is whether the benefits of these two techniques are additive. Thus, we apply the proposed negative sampling strategies to the different model stages and study the impact on the final accuracy.

We conduct experiments on three different datasets: **SQuAD** (Chen et al., 2017), **N**atural **Q**uestions (Kwiatkowski et al., 2019) [1], and **MS**

[*] Work done during an internship at Google.

[1] We evaluate on retrieval part of OpenDomainQA tasks.

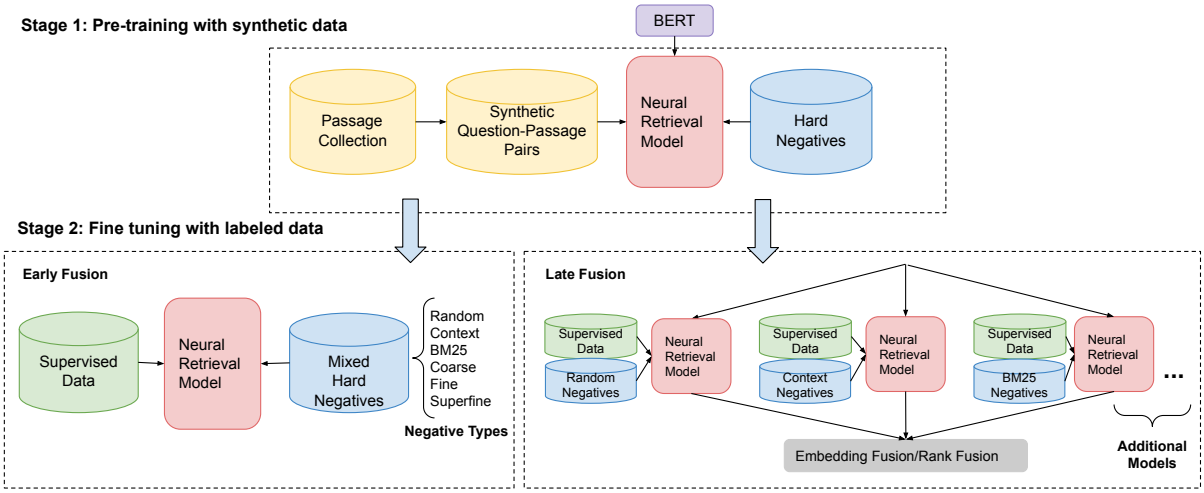

Figure 1: Two-stage neural retrieval model with negative sampling in both stages. In Stage 1, the model is trained using synthetic question-passage pairs. In Stage 2, the model is fine tuned using supervised data. Early and late fusion methods are shown as variations of Stage 2.

MARCO (Nguyen et al., 2016). We show that each of these approaches significantly improves the dual encoder-based retrieval models and combining them together improves the models further. Our final models achieve state-of-the-art performance on NQ and SQuAD improving over the accuracy rates of prior works by 0.8–2.5 points.

The main contributions of this paper are: (1) Systematically explore negative sampling strategies for neural passage retrieval; (2) A novel pre-training approach that integrates synthetic question generation with negative sampling; (3) Fusion approaches that combine models trained with different hard negatives and establish new state-of-the-art performance in the passage-retrieval tasks we tested.

2 Related Work

Previous attempts at improving the quality of dual encoder models can be classified into three types. The first type focuses on finding a good initialization for the model parameters. This is typically achieved by pre-training the model on various tasks (Lee et al., 2019; Chang et al., 2020). Ma et al. (2021) showed that leveraging synthetic question generation is an effective way to improve model accuracy and outperform other variants in zero-shot settings. While the approach was originally proposed for a low-resource scenario, we show that synthetic question pre-training still significantly improves retrieval performance in cases where sufficient amounts of supervised data is available.

The second type focuses on learning better representations using hard negatives. This strategy has proven effective in passage retrieval (Karpukhin et al., 2020), machine translation (Guo et al., 2018) and entity linking (Gillick et al., 2019) tasks. These works mine hard negatives using different strategies. For example, Guo et al. (2018) mine "coarse" negatives with a low-resolution model. Gillick et al. (2019) use a model trained with in-batch negatives and select examples ranked above the correct one as negative examples. Karpukhin et al. (2020)'s dense passage retrieval model (DPR) mines hard negatives using a BM25 model.

Both Xiong et al. (2021) and Zhang and Stratos (2021) proposed to sample negatives from the model itself. While Xiong et al. (2021) analyzed the drawbacks of in-batch negative sampling from the point of view of convergence rate, Zhang and Stratos (2021) argued that contrastive loss is a biased estimator and drawing negative samples from the model itself leads to bias reduction. Moreover, Zhang and Stratos (2021) showed that popular choices of "noisy" distributions such as uniform distribution generally cannot reduce the bias. In this work, we draw a connection between Noise Constrastive Estimation (NCE) and in-batch cross-entropy loss and show that the limited sampling space of in-batch negatives reduces the estimation problem to a much simpler surrogate. Furthermore, we empirically show that combining random sampling with in-batch negatives achieves results competitive with using approximate nearest neighbor negatives, which is typically implemented with asynchronous updates.

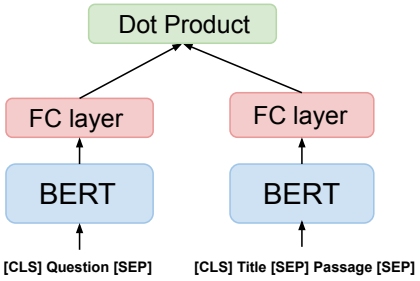

Figure 2: The neural passage retrieval model. The document title and passage are concatenated and fed into the passage encoder.

Another approach focuses on distilling from effective, but less efficient, teacher models such as cross-attention models. Hofstätter et al. (2021) use an ensemble of BERT-based models as teacher and propose a margin mean-squared error that utilize the output margin of the teacher to optimize the student dual encoder model. On the other hand, RocketQA (Qu et al., 2021) applies a different knowledge distillation strategy by using the scores returned by the cross-attention teacher to denoise negative examples and to annotate unlabeled examples. These techniques can be also incorporated in our framework. For example, in a more recent work, Lin et al. (2021) combine knowledge distillation and hard negative sampling in their model.

3 Neural Passage Retrieval Models

Following previous works (Karpukhin et al., 2020; Lee et al., 2019), our dual encoder model is also based on BERT (Devlin et al., 2019). The architecture is shown in Fig 2. To encode a question, we feed the question text to the BERT model and apply a fully-connected (FC) layer of size 768 to the [CLS] token embedding. The output of the FC layer is used as the question embedding. A passage is encoded in a similar way but we prepend to the passage the title of the document where it is found: [CLS] *title* [SEP] *passage* [SEP]. The final question and passage embeddings are then l_2-normalized. The query to passage relevance is computed by the dot-product of their vectors.

The model parameters are initialized from the public uncased $BERT$ checkpoint, and are trained using a listwise loss function (Cao et al., 2007), i.e., cross-entropy loss with in-batch negatives. Let $\mathcal{B}$ denote a batch of question-passage pairs $\{(x_i, y_i)\}^{|\mathcal{B}|}$, we train the model by minimizing the following loss:

$$\mathcal{L}_f = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log\frac{e^{\phi(x_i,y_i)}}{e^{\phi(x_i,y_i)}+\sum_{y_j\in\mathcal{B},j\neq i}e^{\phi(x_i,y_j)}}, \quad (1)$$

where $\phi(x, y)$ denotes the scoring function, in this case the vector dot-product between the question embeddings x and the passages embeddings y. Following Yang et al. (2019), we also add a copy of the above loss in the reverse direction:

$$\mathcal{L}_b = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log\frac{e^{\phi(y_i,x_i)}}{e^{\phi(y_i,x_i)}+\sum_{x_j\in\mathcal{B},j\neq i}e^{\phi(y_i,x_j)}}, \quad (2)$$

and the final loss is the mean of both.

3.1 Pre-training with Synthetic Data

Synthetic data has shown to be as a very effective approach to improve neural passage retrieval models (Ma et al., 2021; Liang et al., 2020). We adopt the approach from Ma et al. (2021) that uses synthetic data for pre-training. In particular, we train our own question generator by fine-tuning a T5-large (Raffel et al., 2020) model which predicts questions given the relevant passage. The model is then used to generate synthetic questions on the passage collection. The generated (synthetic question, passage) pairs are used to train the dense retrieval model.

4 Improved Negative Sampling

In this section, we first draw a connection between the loss function in the previous section and NCE (Ma and Collins, 2018) to shed light on the drawback of in-batch negative sampling. Then we introduce several negative sampling strategies to mitigate the issue.

4.1 Limitation of In-batch Negative Sampling

The training objectives described in section 3, regardless of direction, can be treated as a special case of the ranking-based NCE (Ma and Collins, 2018). To see this, let $p_N(y)$ denote the "noise" distribution from which negative passages are drawn. More importantly, $p_N(y) > 0$ for all $y \in Y$ where Y denotes the set of all passages in the collection. Define $\bar{\phi}(x, y) = \phi(x, y) - \log p_N(y)$ to be the "corrected" scoring function. Then the ranking variant of the NCE loss is defined as:

$$\mathcal{L}^R_{nce} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log\frac{e^{\bar{\phi}(x_i,y_i)}}{e^{\bar{\phi}(x_i,y_i)}+\sum_{y_j\sim p_N(y),y_j\neq y_i}e^{\bar{\phi}(x_i,y_j)}}. \quad (3)$$

Let Y_G denote documents in the annotated relevant (query, document) pairs. We can see that, while $\mathcal{L}^R_{nce}$ draws negatives from the whole document collection Y, in contrast $\mathcal{L}_f$ draws negatives only from Y_G. Although theoretical implications on estimation consistency need further investigation, given the fact that $|Y_G| \ll |Y|$[2], in batch negative sampling reduce the original parameter estimation problem to a much simpler one: given x_i, rank the relevant passage y_i above all others in Y_G rather than Y. There is no guarantee that y_i can be ranked higher than any passage $Y \setminus Y_G$, which harms ranking performance.

4.2 Negative Sampling Strategies

Given the above analysis, this subsection describes several negative sampling strategies to address the drawbacks of in-batch negative sampling.

Random sampling samples negatives passages from Y with equal chance, i.e., treats $P_N(y)$ as a uniform distribution. Despite its simplicity, uniform negative/noise has been shown effective in training language models (Mnih and Teh, 2012).

Context negatives samples negative passages from those occurred in the same document as y_i, assuming these negatives are less relevant to the question than y_i, but more relevant than rest of the passage collection. Documents that contain only one passage are split in half, and the half that does not contain the answer span is picked as negative.

BM25 negatives samples negatives from top passages returned by a BM25 model. Previous work (Karpukhin et al., 2020; Luan et al., 2021) have shown that such negatives are crucial to building high accuracy dense retrieval models.

Neural retrieval negatives employs neural retrieval models to sample negative passages. We do this by running the models on the questions in the training set and then sampling negatives from the top K predictions. As analyzed by Luan et al. (2021), encoding dimension and model size are crucial factors affecting the dense retrieval model accuracy. Varying encoding dimension and model capacity allows us to control the relatedness of the negative passages. In particular, the **coarse** negatives are sampled from a dual encoder model with 3 Transformer layers, and just 25 dimensions in the encoding outputs; the **fine** and **super fine** negatives are sampled from dual encoders with 12 Transformer layers with encoding dimension 512 and 768, respectively.

To illustrate our sampling strategies, Section 7.4 includes examples of all six hard negative types.

5 Hard Negatives in Multi-stage Training

For pre-training and fine-tuning, we use hard negatives in addition to the in-batch negatives. Assuming that there are M hard negatives for each question in the training data, at each training epoch we randomly select N out of M hard negatives. Those N hard negatives are appended to the in-batch negatives as in the standard dual encoder training[3]. Note that the hard negatives for one question are treated as in-batch negatives for the other questions in the batch. Therefore, for a batch of size B, each question is compared during training against $(N+1) \times B$ passages instead of just B passages in the standard way to train a dual encoder.

5.1 Hard Negatives for Pre-training

As the generated question-passage pairs can be noisy, retrieval-based negatives using BM25 or a semantic similarity model could end up generating negative pairs that are better (less noisy) than the synthetic "positive pairs" that result from the question generation process. To avoid this undesirable condition, we use a heuristic-based hard negatives at this stage. Specifically, we use context hard negatives defined in section 4.2. However, this heuristics assumes that there is a mapping between documents and passages. That may not always be the case, as described in Section 7.2 regarding one of our testing tasks.

5.2 Fusion

We study three fusion methods to investigate how the models trained with different negative sampling strategies complement each other.

Mixing. We experiment with mixing all 6 types of negatives in the pool from where to sample N negatives during training. During training, we uniformly sample from the union of different types of negatives for each question. We consider this approach an "early-stage fusion" as opposed to the next two "late-stage fusion" methods.

Embedding fusion. Here, we do a weighted concatenation, as ensemble embeddings, of the question (or passage) embeddings obtained from

[2]Take NQ for example, annotated passages accounts for less than 0.3% of the total number of wiki passages.

[3]The hard negatives are only applied to the question-to-passage loss during training and not in the reverse direction.

the models trained with the different negative strategies. The weights for each embedding type are tuned based on the performance on the development set. Then, we use the ensemble embeddings to retrieve the relevant passages for the questions. The advantage of this fusion is that we only need to perform the retrieval once.

Rank fusion. Following the Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) method, we obtain the final ranking results by considering the ranking positions of each candidate in the rankings generated by the different models.

Notice that for the "early-stage fusion" approach, we train only one single model, while for the "late-stage fusion" approaches, we keep the models trained with different negatives and ensemble during retrieval process.

6 Experimental Setup

We evaluate our proposed approach on two tasks: firstly, we evaluate on the passage retrieval task for open-domain question answering (QA) with the goal of retrieving passages that contain the correct answer spans given a question. Secondly, to understand how our approach performs on large-scale text retrieval datasets, we also evaluate on the MS MARCO passage ranking task.

6.1 Open-Domain QA Retrieval

We evaluate on two open-domain QA datasets: Natural Questions (NQ) and SQuAD. NQ contains questions from actual Google search queries and answers from Wikipedia articles identified by annotators. We follow Lee et al. (2019) and convert the dataset to a format suitable for open-domain QA. Specifically, we only keep questions with short answers (no more than five tokens). On the other hand, SQuAD v1.1 is a commonly used dataset for reading comprehension tasks.[4] In contrast to NQ, the questions in SQuAD are generated by annotators given paragraphs from Wikipedia. The number of questions in each dataset is shown in Table 1.

We use Wikipedia as our collection of documents and knowledge source from where to retrieve passages that answer the questions. Following Lee et al. (2019) and Karpukhin et al. (2020), we use an English Wikipedia dump from Dec.20, 2018. After filtering semi-structured data, such as tables and info-boxes, each document is split into disjoint text passages of 100 words, which yields 21,015,324 passages in total. In order to be able to compare our work with the DPR models from Karpukhin et al. (2020) directly, we use the preprocessed Wikipedia passages as released by the authors.

Dataset	Train	Dev	Test
NQ	58,880	6,515	3,610
SQuAD v1.1	70,096	7,921	10,570
MS MARCO	532,761	6,980	43

Table 1: Number of examples in Train/Dev/Test sets

For open domain QA, we train our question generation model by fine-tuning the T5 large model (Raffel et al., 2020) on NQ, where the model predicts the question conditioned on its long answer. We use the model to sample at most 3 questions for each passage in the collection. This results in 62 million synthetic question-passage pairs in total[5].

We report the results using Top-K accuracy for K = [1, 5, 10, 20, 100], which is the fraction of K retrieved passages that contain a span with the answer to the question.

6.2 MS MARCO Passage Ranking

The MS MARCO passage ranking task consists of two sub-tasks: a full retrieval task and a top-1000 reranking task. In this paper we evaluate on the full retrieval task only, which consists of retrieving passages from a collection of web documents containing about 8.8 million passages. All questions in this dataset are sampled from real and anonymized Bing queries (Nguyen et al., 2016).

Following Xiong et al. (2021), we report results on the MS MARCO dev set and TREC test set from "TREC 2019 DL" track (Craswell et al., 2020). Table 1 shows the number of questions in the train/dev/test sets. We report our results using the MRR@10 and the Recall@1k metrics on the dev set and the Normalized Discounted Cumulative Gain (NDCG@10) on the test set.

We generate synthetic questions in a way similar to as described above but in this case the model is trained on MS MARCO instead of on NQ.

6.3 Implementation Details

We use the public pre-trained uncased BERT[6] as initial checkpoint for our retrieval models. In order to directly compare with prior works, we use

[4] We do not use SQuAD 2.0 because it combines the questions in SQuAD 1.1 with unanswerable questions. It is hard to judge if a question is unanswerable in the open domain setup given that an originally unanswerable question could be answered by one of the passages in the entire passage pool.

[5] See examples of synthetic questions in the Appendix A.

[6] https://github.com/google-research/bert

	Natural Questions					SQuAD				
	Top 1	**Top 5**	**Top 10**	**Top 20**	**Top 100**	**Top 1**	**Top 5**	**Top 10**	**Top 20**	**Top 100**
Baseline models										
DPR (Single)	-	-	-	78.4	85.4	-	-	-	63.2	77.2
DPR *ours*	44.6	68.1	74.5	79.6	86.2	25.3	47.3	56.3	64.4	78.1
BM25	-	-	-	59.1	73.7	-	-	-	68.8	80.0
BM25 + DPR	-	-	-	76.6	83.8	-	-	-	71.5	81.3
ANCE (Single)	-	-	-	81.9	87.5	-	-	-	-	-
Distilled models										
RocketQA	-	74.0	-	82.7	88.5	-	-	-	-	-
Our models										
Synthetic	31.6	59.2	68.1	74.7	84.7	22.3	44.6	54.2	61.9	75.9
+ Gold	40.5	67.2	75.2	80.6	87.4	30.1	53.7	62.0	69.5	81.2
+ Gold + Uniform	40.5	67.7	75.9	81.3	88.2	33.1	56.7	65.4	72.4	83.4
+ Gold + Coarse	42.4	69.3	77.4	81.8	88.1	33.7	57.3	65.8	72.9	83.8
+ Gold + Fine	42.1	69.4	77.4	82.1	88.1	33.4	57.2	65.5	72.8	83.7
+ Gold + BM25	50.0	72.2	78.1	82.2	87.7	30.7	54.4	63.3	70.9	82.7
+ Gold + Context	51.0	72.4	77.8	82.1	88.1	30.6	53.9	62.7	69.7	81.8
+ Gold + Super Fine	51.0	72.6	77.8	82.2	88.2	26.5	49.2	58.4	66.5	80.0
Early fusion (mixing)	50.4	72.5	78.1	82.6	**88.7**	33.9	56.9	64.9	71.7	83.2
Late fusion (embedding)	**62.2**	**74.6**	**79.0**	**82.7**	88.5	**44.4**	**59.4**	**67.0**	**73.5**	**83.8**
Late fusion (rank)	47.9	71.9	78.1	82.5	88.6	34.2	58.0	66.7	73.4	82.7

Table 2: Results on open-domain QA NQ and SQuAD retrieval tasks. **[Our models]** are trained using a two-stage neural retrieval model that uses hard negatives in both stages. The results of baseline models (except "DPR ours") are copied verbatim from the original papers. The missing numbers indicate results that are not reported.

$\text{BERT}_{\text{Base}}$ for open-domain QA retrieval task and $\text{BERT}_{\text{Large}}$ for the MS MARCO passage retrieval task. We encode questions and passages into vectors of size 768. We extract 100 hard negatives for each question and in each training iteration, we randomly pick 2 hard negatives per question to append to the training batch. We train our models for 200 epochs using Adam with learning rate of 5e-6[7]. We use recall@1 on the development set as signal for early stopping. We use Tensorflow version 1.15 and all models are trained on a "4x4" slice of V3 Google Cloud TPU using batches of size 2048.

For question generation, we fine-tune T5 large on a "8x8" slice of V3 Google Cloud TPU. The training data consists of (passage/long-answer, question) pairs, and we truncate passage and question to 256 and 48 sentencepiece (Kudo and Richardson, 2018) tokens, respectively. That batch size is set to 1024 for both NQ and MSMarco. We use the default learning rate, and fine-tune for 15K and 30K steps for NQ and MSMarco, respectively. At inference time we use top-k sampling which is already supported by T5, and K is set 10.

[7]Details of hyperparameter tuning can be found in the Appendix.

7 Results and Discussion

7.1 Results on Open-Domain QA Retrieval

The first rows in Table 2 show the results of the baseline systems starting with **DPR** using the dual encoder model proposed by Karpukhin et al. (2020)[8]. For the sake of reproducibility, we re-implemented the DPR system as described in Section 3. In contrast to ours, the original DPR model does not share the question and passage encoders from the BERT model and instead uses separate encoders for each type of text. Moreover, it does not have an additional fully connected projection layer and the loss function that we use is bidirectional batch-softmax. With these modifications, our implementation **(DPR ours)** outperforms the original DPR on both the NQ and SQuAD evaluations.

The next three rows in the table show the performance of a strong sparse model **BM25**, a hybrid model **BM25+DPR** from Karpukhin et al. (2020) and **ANCE** (Xiong et al., 2021). The second section shows the performance of **RocketQA** (Qu et al., 2021), i.e. the distilled dual encoder model. The subsequent rows in Table 2 show the results of our models starting with the Stage 1 model pre-trained using synthetic data with context hard negatives; no fine tuning. The models in the rest of table are fine-tuned from the model trained in Stage 1

[8]It corresponds to the *Single* version in their paper that trains the model on one dataset only.

	MS MARCO Dev		TREC DL Test
	MRR@10	Recall@1k	NDCG@10
Baseline models			
BM25-Anserini	18.7	85.7	49.7
ANCE	33.0	95.9	64.8
ME-BERT	33.4	-	68.7
ME-HYBRID-E	34.3	-	70.6
$\text{DPR}_{\text{Large}}$ *ours*	27.2	78.6	59.3
Distilled models			
RocketQA	37.0	97.9	-
$\text{BERT-Base}_{\text{DOT}}$	31.5	94.7	66.8
TCT-ColBERT W/ TCT HN+	35.9	97.0	71.9
Our models			
Synthetic	26.5	97.8	58.8
+ Gold	26.6	97.8	58.8
+ Gold + Uniform	33.7	98.2	68.1
+ Gold + Coarse	33.0	98.0	68.1
+ Gold + Fine	33.0	98.3	68.8
+ Gold + BM25	31.9	98.1	66.8
+ Gold + Context	29.6	97.6	65.2
+ Gold + Super Fine	32.1	88.4	66.4
Early fusion (mixing)	**34.2**	98.0	**68.8**
Late fusion (embedding)	33.9	**98.4**	68.1
Late fusion (rank)	32.2	88.4	66.4

Table 3: Results on MS MARCO Dev and TREC DL Test set. Note ANCE uses RoBERTa as the backbone encoder while all others use $\text{BERT}_{\text{Large}}$. ME-BERT and ME-HYBRID-E use multiple vectors.

using the gold data. Our initial approach is a fine-tuned model that uses only in-batch negatives. In this case, it is interesting to notice that the accuracy rates on NQ are already very close to the results of ANCE, and the accuracy rates on NQ and SQuAD outperform both BM25 and DPR. The following six rows show that the models fine-tuned with our different negative sampling strategies outperform the model that does not use hard negatives. They also outperform the baseline models on both NQ and SQuAD. The difference is statistically significant ($p < 0.05$, using the two-tailed t-test). Specifically, when using super fine hard negatives, our model achieves the best Top1 and Top5 accuracy rates on NQ and get a remarkable improvement of 6.4 points and 4.5 points respectively over DPR. The Top 10/20/100 accuracy rates for the six kinds of hard negatives are all very similar. On SQuAD, the model that uses coarse hard negatives achieves the best accuracy rates and outperforms the hybrid BM25+DPR model by 1.4 points on Top20 accuracy and 2.5 points on Top100 accuracy. We reason that the performance difference between NQ and SQuAD is due to the way the datasets were created, and the fact that SQuAD has much larger token overlap between questions and passages compared to NQ. The results illustrate that there is no single best negative sampling strategy across all datasets.

Regarding fusion, we achieve the best Top100 accuracy on NQ by using early-stage fusion in the fine tuning stage. For late-stage fusion, we found that, notably, embedding fusion further improved the Top1 accuracy by 11.2 and 10.7 points on NQ and SQuAD, respectively. Even though not directly comparable with the distilled model, we can see that the embedding fusion model can achieve comparable performance. Rank fusion was helpful to boost the Top 10/20/100 accuracy rates, but not the Top 1/5 cases.

7.2 Results on MS MARCO

Table 3 shows the results on MS MARCO Dev set and TREC DL Test set. The top section of the table shows the results of the baseline models. The dense retrieval models, including **ANCE**, **ME-BERT** and **ME-HYBRID-E** (Luan et al., 2021), significantly outperform **BM25-Anserini** (Yang et al., 2018) with parameters k1=0.82, b=0.68. ME-BERT is a model in which every passage is represented by multiple vectors from BERT. ME-HYBRID-E is a hybrid model of ME-BERT and BM25-Anserini which linearly combines sparse and dense scores using a single trainable weight. Note that ANCE is initialized with $\text{RoBERTa}_{\text{Base}}$ and ME-BERT and ME-HYBRID-E are initialized with $\text{BERT}_{\text{Large}}$. As reference for the performance gains from our improved negative contrast, we also include our implementation of $\text{DPR}_{\text{Large}}$ based on

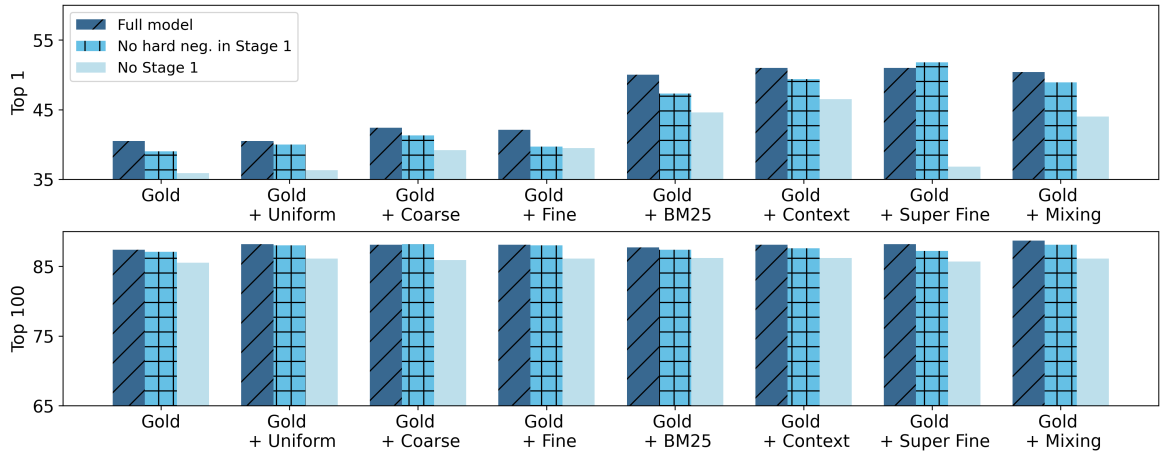

Figure 3: Model Ablation Results on open-domain QA NQ retrieval tasks by removing the hard negatives in stage 1 and removing stage 1 completely.

$\text{BERT}_{\text{Large}}$.

The middle section shows the results of the distilled models. **RocketQA** achieves the state-of-the-art performance on MS MARCO Dev. **$\text{BERT-Base}_{\text{DOT}}$** (Hofstätter et al., 2021) uses an ensemble of three BERT-based cross-attention models to teach a dual encoder student model based on $\text{BERT}_{\text{Base}}$. **TCT-ColBERT**(Lin et al., 2021) uses ColBERT (Khattab and Zaharia, 2020) as teacher with augmented training data containing hard negatives and then distills its knowledge into a student dual encoder model. Note that our results are not directly comparable with those models as they distill additional knowledge from more powerful models and use different training settings.

The bottom sections of the table shows the results of our model. Our Stage 1 model is trained with synthetic data and coarse hard negatives as the mapping between passages to documents is not available in this case. This model outperforms BM25-Anserini and achieves performance close to our DPR baseline. There is not much gain when fine-tuning the Stage 1 model using gold data with in-batch negatives. However, there are considerable gains in all the models that use hard negatives. In particular, the model that uses uniform sampling negatives achieves the best MRR@10 among all six types of hard negatives, and also outperforms ANCE and ME-BERT. We see this as a remarkable confirmation of the benefits of using hard negatives in the fine-tuning stage of this task. The recall@1k for the different types of negatives are very similar except the super fine hard negatives. This may be attributed to the false negatives resulting from the super fine negative sampling given that MS MARCO only annotates one relevant passage for each question. The best NDCG@10 on the test set is achieved when the model is trained with fine hard negatives. Both early and late (embedding) fusion perform similarly on these metrics and they are highly competitive against ME-HYBRID-E, the best performing baseline model, but rank fusion did not help much.

7.3 Model Ablations

We conduct ablation experiments in order to understand the contribution of each component in our models and show the Top1 and Top100 accuracy rates on the open-domain QA NQ dataset in Figure 3. We observe the same trend on other TopK results[9]. The left bars present the performance of the models using the full two-stage training reported in the second part of Table 2. We first remove the hard negatives from Stage 1 but keep them in the fine tuning stage. As shown in the middle bars, the accuracy rates drop across all settings except on the one using super fine hard negatives. This shows that the context hard negatives benefit the training with synthetic data and that using hard negatives in both stages is the best performing option. We go further and remove the Stage 1 training altogether. In this way we are fine tuning directly on the BERT checkpoint. The right bars show that the performance drops significantly and points to the fact that using synthetic data to pre-train the system

[9]See the full table of ablation results in the Appendix B.

Question	Who sings the song Never Be the Same
Answer	Camila Cabello
Gold	Never Be the Same (Camila Cabello song) “Never Be the Same” is a song by Cuban-American singer Camila Cabello from her debut studio album, “Camila” (2018). The song was written by Cabello, Noonie Bao and Sasha Yatchenko .....
In-batch	American Civil War and Cuba’s Ten Years’ War, U.S. businessmen began monopolizing the devalued sugar markets in Cuba. In 1894, 90% of Cuba’s total exports went to the United States...
Uniform	William Robert Brooks (June 11, 1844 – May 3, 1921) was a British-born American astronomer, mainly noted as being one of the most prolific discoverers of new comets of all time.....
Coarse	and his group the Bob-cats. In 2008 Crosby’s rendition of the song appeared as part of the soundtrack of “Fallout 3”. The song made a repeat appearance in “Fallout 4” in 2015. Happy Times (song) “Happy Times” is a jazz ballad written by American lyricist Sylvia Fine .....
Fine	Sisters (song) "Sisters" is a popular song written by Irving Berlin in 1954, best known from the 1954 movie “White Christmas”. Both parts were sung by Rosemary Clooney (who served as Vera-Ellen’s singing vocal dub for this song, while Trudy Stevens dubbed Vera-Ellen’s .....
BM25	release of the album. The song has been certified Gold by the British Phonographic Industry (BPI). An accompanying but unofficial music video for “Never Be the Same” was released on Cabello’s personal YouTube channel on December 29, 2017.....
Context	“Never Be the Same” has been described as a “dark” pop ballad. A “NME” writer described it as “bombastic” electro. The upbeat track features Cabello singing falsetto in the pre-chorus. According to sheet music published by Sony/ATV Music Publishing on Musicnotes.com, ...
Super Fine	wrote in his album review, “Cabello is at her peak on [“Never Be the Same”] which shows off what sets her apart from the pop pack.[...] That she can quickly switch to her full voice for the desperate...”

Table 4: Examples of six types of negative sampling (plus in-batch) for a given question, answer and gold passage.

is highly effective.

7.4 Examples of Hard Negatives

Table 4 shows examples of the six types of negatives plus, for reference, one in-batch negative that was selected from the passages in one of the training batches of NQ[10]. Given a question and its gold passage, the coarse hard negative passage is on topic, about a song, but not about the song mentioned in the question. The fine hard negative passage describes a different song from the one in the question but it mentions the singer of the song discussed. This singer-song relationship is semantically close to the relationship observed in the gold passage. The BM25, context and super fine hard negative passages mention the song in the question and they are semantically closer to the gold passage in comparison to the coarse and fine hard negatives. It is worth noticing the BM25 negative seems to be a plausible answer to the question[11].

8 Conclusions

We presented a multi-stage system for neural passage retrieval based on models that combine the use of synthetic data, negative sampling and fusion. We trained BERT-based dual encoder models using a two-stage system and demonstrated the positive impact of negative sampling in both the pre-training stage, that uses synthetic data, and the fine-tuning stage, that uses supervised data. Results of our pre-training on synthetic data with hard negatives showed the additive benefits of using both methods in combination. We tested our models on passage retrieval tasks and verified that hard negatives in fine-tuning led to considerable gains over previous dense and sparse retrieval models, including on tasks where fine-tuning alone had not shown much improvement. We achieved even greater gains with early- and late-stage fusion. Overall, the combined contributions of synthetic data for pre-training, different negative sampling strategies and late fusion allowed us to achieve state-of-the-art retrieval performance on Natural Questions and SQuAD and highly competitive results on MS MARCO. Our results encourage us to keep exploring this area and investigate similar mechanisms to improve the reranking stage for neural information retrieval and the reading comprehension stage in end-to-end question answering systems.

Acknowledgement

We thank the three anonymous reviewers for their insightful comments, Vladimir Magay, Keith Hall, and Ryan McDonald for valuable discussion, Daniel Cer for reviewing the manuscript and Fangxiaoyu Feng for advice on inference for indexing.

[10] See examples of MS MARCO data in the Appendix C.

[11] *“Never Be the Same”* was released on Cabello’s personal YouTube channel. However, it does not imply that the song is sung by Cabello.

References

Michael Bendersky, Honglei Zhuang, Ji Ma, Shuguang Han, Keith Hall, and Ryan T. McDonald. 2020. RRF102: meeting the TREC-COVID challenge with a 100+ runs ensemble. *CoRR*.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007)*, pages 129–136.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1870–1879.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 758–759.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track.

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

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

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. MultiReQA: A cross-domain evaluation forRetrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 94–104.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2021. Improving efficient neural ranking models with cross-architecture knowledge distillation.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 39–48.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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

Davis Liang, Peng Xu, Siamak Shakeri, Cícero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Embedding-based zero-shot retrieval through query generation. *CoRR*.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9(0):329–345.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the*

European Chapter of the Association for Computational Linguistics: Main Volume, pages 1075–1088.

Zhuang Ma and Michael Collins. 2018. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3698–3707.

Andriy Mnih and Yee Whye Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, page 419–426.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems*.

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

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*, (140):1–67.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using lucene. *J. Data and Information Quality*, 10(4).

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5370–5378.

Wenzheng Zhang and Karl Stratos. 2021. Understanding hard negatives in noise contrastive estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1090–1101.

A Hyperparameter Tuning

We tune the hyperparameters to maximize the top 100 accuracy on open-domain QA retrieval task and MRR@10 on MS MARCO retrieval task. Specifically, we search for (1) the learning rate out of {5e-6, 1e-5, 2e-5, 3e-5}, (2) batch size out of {512, 1024, 2048}, (3) size of hard negatives pool per question out of {1, 20, 100}, (4) number of randomly picked hard negatives per question out of {1, 2, 5}.

B Ablation Results

Table 5 shows the results of ablation experiments on Open Domain QA NQ and SQuAD retrieval tasks by removing the hard negatives in stage 1 and removing stage 1 completely.

C Synthetic Data Examples

Table 6 shows several examples of synthetic questions. The first two are from open-domain QA and the last two are from MS MARCO. As shown in the table, even though they are synthetic questions, they are of high quality. In addition, we can see that questions in these two tasks have different styles.

D MS MARCO Hard Negative Examples

Table 7 shows examples of six types of hard negatives plus an in-batch negative from one of the training batches from MS MARCO dataset.

	Natural Questions					SQuAD				
	Top 1	**Top 5**	**Top 10**	**Top 20**	**Top 100**	**Top 1**	**Top 5**	**Top 10**	**Top 20**	**Top 100**
Full model										
Gold	40.5	67.2	75.2	80.6	87.4	30.1	53.7	62.0	69.5	81.2
Gold + Uniform	40.5	67.7	75.9	81.3	88.2	33.1	56.7	65.4	72.4	83.4
Gold + Coarse	42.4	69.3	77.4	81.8	88.1	33.7	57.3	65.8	72.9	83.8
Gold + Fine	42.1	69.4	77.4	82.1	88.1	33.4	57.2	65.5	72.8	83.7
Gold + BM25	50.0	72.2	78.1	82.2	87.7	30.7	54.4	63.3	70.9	82.7
Gold + Context	51.0	72.4	77.8	82.1	88.1	30.6	53.9	62.7	69.7	81.8
Gold + Super Fine	51.0	72.6	77.8	82.2	88.2	26.5	49.2	58.4	66.5	80.0
Gold + Mixed	50.4	72.5	78.1	82.6	88.7	33.9	56.9	64.9	71.7	83.2
No hard negative in Stage 1										
Gold	39.0	65.9	73.9	79.8	87.1	28.1	51.3	59.9	67.2	80.0
Gold + Uniform	40.0	67.2	75.6	81.6	88.0	30.6	54.7	63.2	70.7	82.8
Gold + Coarse	41.3	68.0	76.0	81.6	88.2	30.7	54.6	63.3	70.7	82.6
Gold + Fine	39.7	67.6	76.0	81.2	88.0	31.0	54.3	63.4	70.5	82.5
Gold + BM25	47.3	70.8	77.0	81.4	87.4	27.0	50.3	59.8	67.9	80.8
Gold + Context	49.4	71.2	77.2	81.4	87.6	28.8	52.4	61.3	68.5	80.9
Gold + Super Fine	51.8	71.1	77.0	81.5	87.2	27.9	50.2	59.6	67.2	79.8
Gold + Mixed	48.9	72.0	78.0	82.2	88.1	32.0	55.8	64.4	71.5	83.3
No Stage 1										
Gold	35.9	62.2	70.3	77.2	85.5	25.0	46.5	54.7	62.4	75.7
Gold + Uniform	36.3	64.1	72.8	78.6	86.1	27.0	48.3	57.0	64.3	77.7
Gold + Coarse	39.2	65.5	73.2	78.8	85.9	30.5	49.7	58.5	66.1	79.8
Gold + Fine	39.5	65.8	73.8	79.3	86.1	28.7	48.1	56.6	64.7	77.7
Gold + BM25	44.6	68.1	74.5	79.6	86.2	25.3	47.3	56.3	64.4	78.1
Gold + Context	46.5	68.5	74.7	79.3	86.5	28.2	47.6	56.3	63.8	77.4
Gold + Super Fine	36.8	63.9	71.9	78.2	85.7	25.3	47.3	55.4	63.6	77.2
Gold + Mixed	44.0	67.1	74.3	79.7	86.1	28.3	50.4	59.3	66.7	80.0

Table 5: Model Ablation Results on Open Domain QA NQ and SQuAD retrieval tasks by removing the hard negatives in stage 1 and removing stage 1 completely.

Passage	Synthetic Questions
North Park Secondary School is a public high school located at the major intersection of Williams Parkway and North Park Drive in Brampton, Ontario, Canada. It was founded in 1978, making it one of the oldest high schools in the area. North Park is best known for being one of three high schools in Brampton to offer the IBT program, a program using business and technology to enrich the learning of its students. Students in the IBT program are often required to bring a device such as a laptop to guide through courses by filing notes	why do students go to north park school what is the type of school north park high school where is north park secondary school in brampton ontario
age 18 and over, there were 93.1 males. The median income for a household in the CDP was $43,125, and the median income for a family was $45,327. Males had a median income of $36,524 versus $29,861 for females. The per capita income for the CDP was $19,670. About 9.7% of families and 9.9% of the population were below the poverty line, including 14.5% of those under age 18 and 8.7% of those age 65 or over. In the state legislature, Valley Springs is in , and . Federally, Valley Springs is in . Valley Springs, California Valley Springs (formerly,	what is the poverty line in valley springs ca what is the median income in valley springs ca what is the largest city in the central valley of california
Start recording at any time during a conference call. Control as you record by pausing and resuming recording. Recording can be initiated by any touch-tone phone. Playback toll-free via phone access, start, stop, rewind and fast forward at your control using touch-tone commands on the phone keypad.	are tap phones recording how to record a conference call can you see what you record on your phone
Updated PANDAS signs and symptoms (1) Pediatric onset. The first symptoms of PANDAS are most likely to occur between 5 and 7 years of age. Symptoms can occur as early as 18 months of age or as late as 10 years of age. If the first clinically recognized episode is detected after the age of 10, it is unlikely true initial episode, but the recurrent one.	child pandas symptoms age of onset of pandas what age can you be affected by pandas

Table 6: Examples of Synthetic Data

Question Gold	Genetic Predispositions definition psychology A genetic predisposition is a genetic effect which influences the phenotype of an organism but which can be modified by the environmental conditions. Genetic testing is able to identify individuals who are genetically predisposed to certain health problems.redisposition is the capacity we are born with to learn things such as language and concept of self. Negative environmental influences may block the predisposition (ability) we have to do some things.
In-batch	They're loaded with nutrients, called antioxidants, that are good for you. Add more fruits and vegetables of any kind to your diet. It'll help your health. Some foods are higher in antioxidants than others, though. The three major antioxidant vitamins are beta-carotene, vitamin C, and vitamin E.
Uniform	How to Deal With a Liar. Do you know someone who can't seem to utter the truth? Some people lie to make themselves look good or to get what they want, and others because they actually believe what the...
Coarse	Prevention of Musculoskeletal Disorders in the Workplace. Musculoskeletal disorders (MSDs) affect the muscles, nerves and tendons. Work related MSDs (including those of the neck, upper extremities and low back) are one of the leading causes of lost workday injury and illness.
Fine	Mycoplasma pneumoniae (M. pneumoniae) is an atypical bacterium (the singular form of bacteria) that causes lung infection. It is a common cause of community-acquired pneumonia (lung infections developed outside of a hospital).M. pneumoniae infections are sometimes referred to as walking pneumonia..n general, M. pneumoniae infection is a mild illness that is most common in young adults and school-aged children. The most common type of illness caused by these bacteria, especially in children, is tracheobronchitis, commonly called a chest cold.
BM25	There is definitely a genetic predisposition to arterial disease and the risk factors that cause it.There have been certain genetic abnormalities that have been identified.here is definitely a genetic predisposition to arterial disease and the risk factors that cause it.
Context	... are at risk for loss of health insurance if they are discovered to have genetic predispositions for health problems. The national center for genome resources found that 85 percent of those polled think employers should not have access to information about their employees genetic conditions risks or predispositions. 2 the us federal government has so far taken only limited measures against discrimination based on genetic testing...
Super Fine	Understanding genetic predisposition to disease and knowledge of lifestyle modifications that either exacerbate the condition or that lessen the potential for diseases (i.e., no smoking or drinking) ...

Table 7: Examples of negative sampling strategies (plus in-batch) for the MS MARCO dataset.