# Towards Robust Neural Retrieval with Source Domain Synthetic Pre-Finetuning

**Revanth Gangi Reddy**[1], **Vikas Yadav**[3]*, **Md Arafat Sultan**[2], **Martin Franz**[2],
**Vittorio Castelli**[3]*, **Heng Ji**[1], **Avirup Sil**[2]
[1]University of Illinois at Urbana-Champaign [2]IBM Research AI [3]Amazon AWS AI
{revanth3,hengji}@illinois.edu, arafat.sultan@ibm.com,
{franzm,avi}@us.ibm.com, vittorca@amazon.com

## Abstract

Research on neural IR has so far been focused primarily on standard supervised learning settings, where it outperforms traditional term matching baselines. Many practical use cases of such models, however, may involve previously unseen target domains. In this paper, we propose to improve the out-of-domain generalization of Dense Passage Retrieval (DPR)—a popular choice for neural IR—through synthetic data augmentation ***only in the source domain***. We empirically show that pre-finetuning DPR with additional synthetic data in its source domain (Wikipedia), which we generate using a fine-tuned sequence-to-sequence generator[1], can be a low-cost yet effective first step towards its generalization. Across five different test sets, our augmented model shows more robust performance than DPR in both in-domain and zero-shot out-of-domain evaluation.

## 1 Introduction

Traditional approaches to information retrieval (IR) such as TF-IDF (Salton and McGill, 1986) and BM25 (Robertson and Zaragoza, 2009) rely on lexical matching for query-passage alignment. In contrast, neural IR encodes passages and questions into continuous vector representations, enabling deeper semantic matching. Modern neural IR systems (Lee et al., 2019; Chang et al., 2019) based on pre-trained masked language models (MLM) (Devlin et al., 2019) typically employ a dual encoder architecture (Bromley et al., 1993), where two separate MLMs encode the question and the passage. Karpukhin et al. (2020) show that useful weak supervision for such systems can be derived from the related task of machine reading comprehension (MRC) (Kwiatkowski et al., 2019; Joshi et al., 2017). Their Dense Passage Retrieval (DPR) model demonstrates state-of-the-art (SOTA) in-domain performance on multiple Wikipedia-based datasets (Kwiatkowski et al., 2019; Joshi et al., 2017; Berant et al., 2013; Baudiš and Šedivỳ, 2015), outperforming both term matching baselines like BM25 and prior neural approaches, e.g., the Inverse Cloze Task (Lee et al., 2019) and latent learning of the retriever during MLM pre-training (Guu et al., 2020).

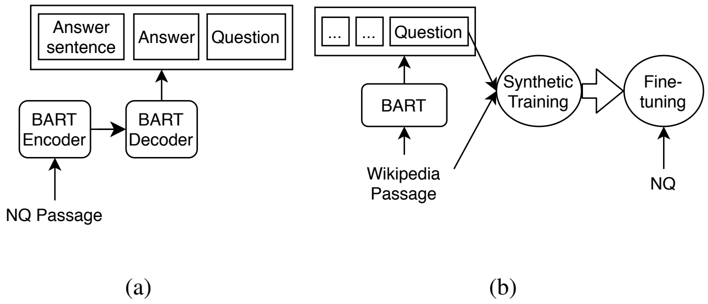
Despite its high in-domain utility, however, Reddy et al. (2021) show that DPR performance can drop significantly in novel test domains. They propose target domain synthetic data augmentation as a solution to this problem, which augments DPR with additional synthetic training data generated from target domain text. While this approach does indeed improve DPR scores in the new test domain, it has a key practical limitation: for every new target domain, it requires generating a new synthetic training corpus and re-training the model. Here we ask if an augmentation approach that only operates once in the source domain, and does not require re-training every time a new test domain is encountered, can also help improve domain generalization.

To better understand DPR's zero-shot out-of-domain (OOD) utility, we first run an empirical evaluation where both BM25 and DPR are applied to several out-of-domain test datasets. We observe that (i) DPR still holds an advantage over BM25 in near domain evaluation on Wikipedia-based datasets, but the difference is considerably lower than in the in-domain case, and (ii) In the far domain of biomedical text, DPR actually underperforms BM25. Our OOD evaluation is more comprehensive than Reddy et al. (2021), demonstrating the zero-shot utility of DPR in a more detailed and fine-grained manner.

Next we investigate if a one-off pre-finetuning of DPR with large amounts of ***source domain*** synthetic IR data can help improve its robustness to domain shift. Utilization of synthetic training data is common in related tasks such as machine reading comprehension (MRC) (Shakeri et al., 2020; Zhang

*Work done while at IBM Research AI

[1] Synthetic question generation code is available at: https://github.com/primeqa/primeqa/tree/main/primeqa/qg



(a) (b)

**History of Tanzania** The African Great Lakes nation of Tanzania dates formally from ***1964***, when it was formed out of the union of the much larger mainland territory of Tanganyika and the coastal archipelago of Zanzibar. The former was a colony and part of German East Africa from the 1880s to 1919, when, under the League of Nations, it became a ***British mandate***. It served as a military outpost ...

*when did tanzania became a country in africa?*
*who owned zanzibar and tanganyika before they were independent?*

(c)

Figure 1: The proposed IR training pipeline and a synthetic example. (a) A BART encoder-decoder LM is fine-tuned on NQ for QA example generation; (b) Synthetic examples generated from Wikipedia passages are used to pre-finetune the neural IR model before fine-tuning on NQ; (c) Two synthetic questions output by our generator from the depicted Wikipedia passage, with corresponding answers highlighted in the text.

et al., 2020; Sultan et al., 2020). Nevertheless, a close examination of synthetic pre-finetuning as an augmentation technique is key for zero-shot neural IR due to the presence of highly effective and domain-agnostic term matching baselines like BM25.

We fine-tune a sequence-to-sequence generator on labeled MRC data and use it to generate synthetic IR examples from source domain passages (§2). Our experiments show that pre-finetuning DPR with these generated examples does indeed improve its accuracy on both in-domain and out-of-domain test sets. Crucially, the gap with BM25 in far domain evaluation is significantly reduced.

The main contributions of this paper are:

- We conduct an empirical evaluation of SOTA neural IR on multiple in-domain and out-of-domain test sets, showing how its utility varies in different test conditions.
- We show that a one-off ***source domain*** synthetic pre-finetuning step can significantly improve the robustness of neural IR, with improvements on five different test sets, including in the practical zero-shot setting.

## 2 Source Domain Synthetic Pre-Finetuning

In this section, we describe the procedure for synthetic pre-finetuning of the DPR model. We first detail how we train the sequence-to-sequence generator and generate source domain syntheic data from it. Next, we describe how this data is used for training the DPR model.

Let $c$ be a text corpus and $d \in c$ be a document. An IR example, more specifically a passage retrieval example, consists of a question $q$ and a passage $p$ in $d$ such that $p$ contains an answer $a$ to $q$. Let $s$ be the sentence in $p$ that contains $a$.

We first train an example generator by fine-tuning BART (Lewis et al., 2020a)—a pre-trained encoder-decoder language model—to generate an ordered triple $(s, a, q)$ from an input passage $p$. This procedure in essence uses generation to first identify a candidate sentence $s$ in $p$, then extract a candidate answer $a$ from $s$, and finally generate a corresponding question $q$. In practice, we approximate the generation of $s$ by generating only its first and last words. Finally, $(q, p)$ is retained as a synthetic IR example. Labeled $(p, s, a, q)$ tuples needed for the supervision of this model are taken from Natural Questions (NQ) (Kwiatkowski et al., 2019), an existing MRC dataset over Wikipedia articles.

With the generator, we produce positive synthetic pre-finetuning examples for DPR from Wikipedia passages. Following Sultan et al. (2020), we use top-$p$ top-$k$ sampling (Holtzman et al., 2020) to promote diversity in the generated examples. Training and inference of the synthetic example generator are depicted in Figures 1a and 1b, respectively. Figure 1c shows two example questions output by the generator from a Wikipedia passage.

To obtain a negative sample for each generated question $q$, we retrieve passages from Wikipedia using BM25 and randomly sample one that does not contain the generated answer $a$. Following Karpukhin et al. (2020), we also use in-batch negative samples for training. After pre-finetuning with synthetic examples, we fine-tune the model with IR examples derived from NQ. We name this synthetically augmented DPR model *AugDPR*. We refer the reader to (Karpukhin et al., 2020) for a more detailed description of the DPR training process.

## 3 Experimental Setup

### 3.1 Datasets

We briefly describe our datasets in this section. Statistics for each dataset are shown in Table 1.

| Dataset | Domain | Passages | Questions |
|---|---|---|---|
| NQ | Wikipedia | 21.0M | 3,610 |
| TriviaQA | Wikipedia | 21.0M | 11,313 |
| WebQuestions | Wikipedia | 21.0M | 2,032 |
| WikiMovies | Wikipedia | 21.0M | 9,952 |
| BioASQ | Biomedical | 37.4M | 1092 |

Table 1: Statistics of the retrieval corpora and the test sets we use to evaluate all IR models.

**Training and In-Domain Evaluation:** We train all systems on Natural Questions (NQ) (Kwiatkowski et al., 2019), a dataset with questions derived from Google's search log and their human-annotated answers from Wikipedia articles. Lewis et al. (2020b) report that 30% of the NQ test set questions have near-duplicate paraphrases in the training set and 60–70% of the test answers are also present in the training set. For this reason, in addition to the entire NQ test set, we also use the non-overlapping subsets released by Lewis et al. (2020b) for in-domain evaluation.

**Near Domain Evaluation:** For zero-shot near domain evaluation, where Wikipedia articles constitute the retrieval corpus, we use the test sets of three existing datasets.

***TriviaQA*** (Joshi et al., 2017) contains questions collected from trivia and quiz league websites, which are created by Trivia enthusiasts.

***WebQuestions (WQ)*** (Berant et al., 2013) consists of questions obtained using the Google Suggest API, and answers selected from entities in Freebase by AMT workers.

***WikiMovies*** (Miller et al., 2016) contains question-answer pairs on movies, built using the OMDb and MovieLens databases. We use the test split adopted in (Chen et al., 2017).

**Far Domain Evaluation.** For zero-shot far domain evaluation, we use a biomedical dataset.

***BioASQ*** (Tsatsaronis et al., 2015) is a competition[2] on large-scale biomedical semantic indexing and QA. We evaluate on all factoid question-answer pairs from the training and test sets of task 8B.

### 3.2 Setup

**Training:** We train the synthetic example generator using the *(question, passage, answer)* triples from NQ. The model is trained for 3 epochs with a learning rate of 3e-5 and batch size of 24. We then randomly sample 2M passages from the 21M-passage Wikipedia corpus and generate around four synthetic questions per passage. For top-$p$ top-$k$ sampling, we use $p = 0.95$ and $k = 10$.

During synthetic pre-finetuning of DPR, for each of the 2M passages, we randomly select one of its synthetic questions at each epoch to create a synthetic example. After six epochs of synthetic pre-finetuning with a learning rate of 1e-5 and batch size of 1024, we fine-tune DPR on NQ for twenty epochs with a learning rate of 1e-5 and batch size of 128 to get the AugDPR model.

**Baselines and Metrics:** We evaluate BM25 as a term matching baseline. Our BM25 baseline is based on Lucene[3] implementation. BM25 parameters $b = 0.75$ (document length normalization) and $k_1 = 1.2$ (term frequency scaling) worked best. As our neural baseline, we use the DPR-single model trained on NQ and made public[4] by Karpukhin et al. (2020). Both DPR and AugDPR use BERT-base-uncased for question and passage encoding. As in (Karpukhin et al., 2020), our evaluation metric is top-$k$ retrieval accuracy, which is the percentage of questions with at least one answer in the top $k$ retrieved passages.

## 4 Results and Discussion

Table 2 shows NQ results on the entire test set as well as on the two subsets released by Lewis et al. (2020b). Synthetic pre-finetuning yields larger gains on the non-overlapping splits, with up to a 4-point improvement in top-1 retrieval accuracy.

To assess the cross-domain utility of AugDPR, we evaluate it zero shot on both near and far domain test sets. Table 3 shows the results. For comparison, we also show results for supervised models reported by Karpukhin et al. (2020) on TriviaQA and WebQuestions where the DPR model was trained directly on the training splits of these datasets. For the near domain datasets, both DPR and AugDPR

[2] http://bioasq.org/participate/challenges

[3] https://lucene.apache.org/

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

| Model | Total | | | No answer overlap | | | No question overlap | | |
|---|---|---|---|---|---|---|---|---|---|
| | Top-1 | Top-10 | Top-20 | Top-1 | Top-10 | Top-20 | Top-1 | Top-10 | Top-20 |
| BM25 | 30.5 | 54.5 | 62.5 | 26.4 | 47.1 | 54.7 | 31.0 | 52.1 | 59.8 |
| DPR | 46.3 | 74.9 | 80.1 | 32.2 | 62.2 | 68.7 | 37.4 | 68.5 | 75.3 |
| AugDPR | **46.8** | **76.0** | **80.8** | **36.0** | **65.0** | **70.8** | **41.4** | **70.8** | **76.6** |

Table 2: NQ top-$k$ retrieval results. Performance improves across the board with synthetic pre-finetuning (AugDPR), but more on the non-overlapping subsets of Lewis et al. (2020b).

| | Near Domains | | | | | | Far Domain | |
|---|---|---|---|---|---|---|---|---|
| Model | TriviaQA | | WebQuestions | | WikiMovies | | BioASQ | |
| | Top-20 | Top-100 | Top-20 | Top-100 | Top-20 | Top-100 | Top-20 | Top-100 |
| BM25 | 66.9 | 76.7 | 55.0 | 71.1 | 54.0 | 69.3 | **42.1** | 50.5 |
| DPR | 69.0 | 78.7 | 63.0 | 78.3 | 69.8 | 78.1 | 34.7 | 46.9 |
| AugDPR | **72.2** | **81.1** | **71.1** | **80.8** | **72.5** | **80.7** | 41.4 | **52.4** |
| Supervised | 79.4 | 85.0 | 73.2 | 81.4 | - | - | - | - |

Table 3: Zero-shot neural retrieval accuracy improves with synthetic pre-finetuning (AugDPR) in all out-of-domain test settings. However, BM25 remains a strong baseline on the far domain dataset of BioASQ. The numbers for the supervised models are taken from (Karpukhin et al., 2020).

outperform BM25 by a sizable margin; additionally, AugDPR consistently outperforms DPR. Furthermore, performance of AugDPR on WebQuestions is comparable to that of the supervised model. On the far domain, however, we observe that BM25 is a rather strong baseline, with clearly better scores than DPR. The synthetic pre-finetuning of AugDPR reduces this gap considerably, resulting in a slightly lower top-20 score but a 2-point gain in top-100 score over BM25.

To investigate the relative underperformance of neural IR on BioASQ, we take a closer look at the vocabularies of the two domains of Wikipedia articles and biomedical literature. Following Gururangan et al. (2020), we compute the overlap between the 10k most frequent tokens (excluding stop words) in the two domains, represented by 3M randomly sampled passages from each. We observe a vocabulary overlap of only 17%, which shows that the two domains are considerably different in terminology, explaining in part the performance drop in our neural models. Based on these results, we also believe that performance of neural IR in distant target domains can be significantly improved via pre-finetuning on synthetic examples that are generated from raw text in the target domain. We plan to explore this idea in future work.

We also examine the lexical overlap between the questions and their passages, since a high overlap would favor term matching methods like BM25. We find that the coverage of the question tokens in the respective gold passages is indeed higher in BioASQ: 72.1%, compared to 58.6% and 63.0% in NQ and TriviaQA, respectively.

To analyze how much synthetic data is required, we experiment with pre-finetuning using 1M and 4M synthetic examples while keeping the number of training updates fixed. As Table 4 shows, we do not see any improvements from using more examples beyond 2M.

Karpukhin et al. (2020) report that DPR finetuning takes around a day on eight 32GB GPUs, which is a notable improvement over more computationally intensive pre-training approaches like (Lee et al., 2019; Guu et al., 2020). Our synthetic pre-finetuning takes around two days on four 32GB GPUs, which is comparable with finetuning in terms of computational overhead.

| Model | Top-10 | Top-20 | Top-100 |
|---|---|---|---|
| DPR | 73.6 | 78.1 | 85.0 |
| AugDPR-1M | 74.4 | 79.2 | 85.5 |
| AugDPR-2M | 74.8 | 79.7 | 85.9 |
| AugDPR-4M | 74.6 | 79.1 | 85.9 |

Table 4: Retrieval accuracy on the Natural Questions development set with varying number of synthetic examples (1M vs 2M vs 4M) during pre-finetuning.

## 5 Conclusion

We have shown that pre-finetuning a SOTA neural IR model using large amounts of *source domain* synthetic data improves its robustness in zero-shot application settings. Our experiments show consistent performance gains on five in-domain and out-of-domain test sets, including a far target domain that has significant vocabulary mismatch with the training domain. Future work will explore incorporating more control into the generation of synthetic data to increase its diversity and also to overcome potential biases in finetuning data.

## References

Petr Baudiš and Jan Šedivỳ. 2015. Modeling of the question answering task in the yodaqa system. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 222–228. Springer.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.

Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688.

Wei-Cheng Chang, X Yu Felix, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2019. 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 (Volume 1: Long Papers)*, pages 1870–1879.

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

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020b. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409.

Revanth Reddy, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avirup Sil, Vittorio Castelli, Radu Florian, and Salim Roukos. 2021. Synthetic target domain supervision for open retrieval qa. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1793–1797.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for qa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avi Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5461–5468, Online. Association for Computational Linguistics.