A Distribution-Aware Approach to Dense Retrieval

Stanford University CS224N Custom Project

Jason Lin
Dept. of Computer Science
jj0@stanford.edu *

Justin Young
Dept. of Economics
justiny@stanford.edu

Simran Arora
Dept. of Computer Science
simarora@stanford.edu

Abstract

Dense information retrieval systems have gained popularity in recent years. Given a user query, similarity-search based systems retrieve the top k most similar document from a background corpus, and performance are largely evaluated over a single distribution, *either* in distribution *or* out of distribution (zero-shot). In this work, we identify that many practical workloads require retrieving from a combination of in-distribution (ID) and out-of-distribution (OOD) candidates, especially as corpora update over time. In response, we ask whether alternate retrieval strategies, i.e., besides simply retrieving the top k most similar documents, and fine-tuning strategies, beyond vanilla fine-tuning, would enable fairer performance in the mixed-distribution setting. We first propose a synthetic setting to evaluate this question, and then discuss our results and opportunities for future work.

1 Introduction

Information retrieval is an important step for open-domain applications such as language modeling [1], question-answering [2], fact-checking, and personal assistants [3]. Such applications can receive user inputs about nearly anything, requiring access to a wide range of knowledge. In literature, retrieval-based systems follow a two-stage approach by first retrieving then reading: explicitly collecting the top k relevant documents from a large background corpus, and providing this to a separate task model, which reasons over the knowledge to generate an output (or answer). Research has shown that improvement on retriever performance transfers readily to downstream tasks. In this work, we focus on dense retrieval system which select relevant passages based on *similarity scores* between dense question and dense passage representations [4, 5]. The performance of retrievers has significantly improved in recent years, though it is well known that they still struggle to retrieve out-of-distribution text domains and retrieval tasks [6, 7, 8]. However, existing work does not consider the further difficulty of retrieving from a *mixture* of in and out of distribution items.

Practical workloads often consist of a mixture of in and out-of-distribution (OOD) data, especially as corpora evolve over time. A key challenge for studying the proposed retrieval setting is the lack of benchmarks that explicitly require retrieving from a mixture of distributions. One option is to combine two existing questions and corpora, however it would be possible for questions from one dataset to be answerable using passages from the other from our preliminary studies, making it difficult to evaluate how retrieval quality changes. We first construct an evaluation testbed by creating synthetic datasets that contain distribution shifts from an existing general domain benchmark, MS MARCO [9]. Sub-distributions in the retrieval setting can arise either due to the types of questions asked or passages retrieved. We compare both styles of distribution shift in our analysis.

The main reasons we hypothesize mixed-distribution retrieval could require alternate strategies are:

1. **Training strategy** Dense retrievers are trained contrastively [4], and there has been limited evaluation of how the fine-tuning strategy impacts retrieval generalization. Prior work shows tries *vanilla* fine-tuning on a distribution of one question type, and shows performance

*Mentor: Eric Mitchell

degrades when the retriever applied to a new question type [10]. Hard negative mining has been reported to improve retrieval performance [4, 11], however these works evaluate in-distribution performance on standard benchmarks. Prior evaluation studies of dense retrievers do not use hard negative mining to fairly compare to methods that do not use hard negatives [6]. In this landscape, we observe that advanced fine-tuning strategies are not well-studied from a generalization perspective and we hypothesize that fine-tuning with hard negatives may mitigate performance tradeoffs between in-distribution and OOD retrieval.

2. **Retrieval strategy** While in single-distributions retrieval all of the top k documents are from the instant distributions, under the mixed retrieval setting, it is possible for zero of the top k passages to be from one of the sub-distributions. We hypothesize a retrieval model that is biased towards one sub-distribution will favor retrieving passages from that sub-distribution over the OOD passages. This might suggest using a *sub-distribution aware* retrieval method, instead of simply choosing the top-k overall documents, may be preferable.

In summary, our contributions are:

1. We design synthetic evaluation datasets to study multi-distribution retrieval.
2. We show that mining hard-negatives provides large gains when fine-tuning on small corpora compared to vanilla fine tuning, and can outperform pretraining on the full corpora.
3. We provide in-depth analysis on fine-tuning and its effect on retrieving in both ID and OOD settings.

2 Related Work

Multiple Distributions in Retrieval Prior work evaluates popular retrievers on out-of-distribution (OOD) data [12, 6], however while in zero-shot retrieval, k of the top k retrieved passages for a question are from the OOD corpus, under mixed-retrieval, it is possible to retrieve zero OOD passages in the top k. [10] considers retrieving over multiple question types, however their question categories are formed by introducing different question templates, while we seek to understand latent sub-distributions in existing benchmarks.

To support retrieval over multiple domains, common approaches in prior work include training a single retriever on a mixture of domains [12, 13], or using a mixture of experts or specialized encoders [14, 10], each tailored to a different domain. While these techniques can improve generalization, we may not be able to access all downstream retrieval distributions during training.

Retrieval Training Approaches Prior work uses hard-negative mining for training dense retrievers [4, 11]. Whiles these methods report end-to-end lifts on in-distribution evaluations, they do not consider the effects on different sub-distributions of the dataset or OOD performance. BEIR [6], which includes a rigorous evaluation of different retrievers' OOD performance, does not use hard negatives in its implementations of dense retrievers to maintain a fair comparison between methods, based on our understanding. In contrast, we explore the how hard-negative mining might impact robustness.

3 Approach

Here we describe the retrieval problem and our process to create synthetic datasets that contain sub-distributions, which is a precursor to studying the multi-distribution setting.

3.1 Preliminaries

Dense Retrieval The goal of dense information retrieval is to train a model to find relevant documents to a user query, from a massive background corpus. The typical approach is to encode questions and documents either with a single neural network, or using a bi-encoder architecture in which one encoder is for queries and the other is for document. Relevance is determined by taking the dot product or cosine similarity between the encoded query and document representations. For efficient similarity search, documents are encoded in optimized data structures [15]. Generally, the top-k most relevant passages are retrieved per query, and provided to a second *task-specific* model such as a reader that produces an answer to the question.

Benchmarks Popular existing question-answering benchmarks require retrieving from a single distribution ([9, 16], inter alia.), so we need to construct an evaluation setting for our retrieval setting. One option to study the multi-distribution setting is to combine the questions of benchmarks spanning multiple domains. In preliminary experiments under this setting, we mix the questions and corpora of the two documents (D_1 and D_2), and use the same pretrained retrieval model checkpoint, that was trained on a third, separate distribution (D_3), to perform retrieval for all the questions. We observe that for the questions coming from D_1 that is *farther* from the D_3 by Jaccard similarity, a large proportion of the top retrieved passages come from the passages associated with D_2, which distributionally *closer* to the training distribution.

While this observation could have important implications for the optimal retrieval strategy — for example, that a retrieval strategy that is aware the underlying data contains two distributions D_1 and D_2 might be preferable to always choosing the overall top-k passages — unfortunately a confounding factor is that it is challenging to decouple whether the alternate passages that are chosen for D_1 questions are *incorrect* or actually, unknowingly, *better* than passages in the D_2 corpus for the question at hand. It is also possible that a question can be answered by passages from multiple datasets with overlapping subdomains. In order to eliminate this confounding factor, we need a new evaluation set — we thus choose to synthetically split *one dataset* into two subsets.

3.2 Synthetic Domain Split

We design synthetic splits from MS MARCO [9], a popular retrieval benchmark. Our objective is to obtain two datasets D_1 and D_2 that reflect a distribution shift. In retrieval, distribution shifts can arise from the set of questions, or the set of passages. We consider generating the shifts from both perspectives. Below, we discuss our protocol for generating synthetics.

Clustering Passages We first use [5] to encode all questions, then uniform manifold approximation and projection (UMAP), a type of non-linear dimensionality reduction [17]. Because UMAP uses a topological approach to capture the interconnectedness of data points, the variance explained by the procedure cannot easily be calculated. As a heuristic, we first apply PCA to our high-dimensional data and iteratively increase the number of components k until the total variance explained first rises above 60% [18] at some $\bar{k}$. We then set $\bar{k}$ as the desired dimension from UMAP.[2]

To set the number of clusters, we use the Gap statistic approach [19]. Intuitively, this statistic aims to capture how tight points are around a cluster. Letting $d_{ii'}$ be the standard $\ell_2(i, i')$ norm, with data separated into k clusters $C_1, \ldots, C_k$, the Gap statistic is given by:

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

where $W_k = \sum_{r=1}^{k}(1/2|C_r|)\sum_{i,i' \in C_r} d_{ii'}$. The term $E_n^*\{\log(W_k)\}$ denotes the expectation of $\log(W_k)$ under a "null" sampling distribution of size n. The larger the Gap statistic, the more confident we are that k is the correct number of clusters. See Appendix A for full details.

3.3 Multi-Distribution Retrieval Approaches

To evaluate our proposed retrieval setting using the constructed D_1 and D_2, we start by using an out of the box retrieval model f_θ trained on the full MS MARCO benchmark. Next, we assume that we only have access to D_1 during training, but questions from both D_1 and D_2 during inference. By training the base model on D_1, we hypothesize this *biases* the model towards D_1. We are ultimately interested in observing whether the biased retriever degrades on D_2 and struggles to select passages corresponding to questions in D_2, and whether it's possible to mitigate performance degradations on D_2 during inference, despite only having access to D_1 during training.

Vanilla Fine-tuning Fine-tuning is a popular transfer learning method in which both the original deep feature extraction model f_θ, and the predictor head are updated via gradient descent. We tune f_θ on D_1 using the Multiple Negatives Reranking Loss function, which expects as input a batch consisting of sentence pairs $(a_1, p_1), (a_2, p_2) \ldots, (a_n, p_n)$ where (a_i, p_i) are a positive pair (i.e., a question and the passage containing its answer), and (a_i, p_j) for $i \neq j$ a negative pair, and minimizes the negative log-likehood for softmax normalized scores over passages [20].

[2] While only a quick-and-easy heuristic, this is largely justifiable given that we have no reason to believe our data is linear; indeed we likely capture more than 60% of the explained variance with this method.

BM25-retrieved Hard Negative Fine-tuning Next, we explore training using *hard negatives*, which intuitively should be more difficult for the retriever to tell apart from the positive passage for the question, thus "encouraging" the model to learn more nuanced reasoning patterns. A popular strategy in the literature is to use a "simple" retriever to mine hard negatives. In our case, we use a popular and powerful retriever that is based on sparse encodings of questions and passages called BM25. We use BM25 to select the top passages for every question in D_1, and select the top passage that is not the gold-labeled positive passage and incorporate during the contrastive fine-tuning.

4 Experiments

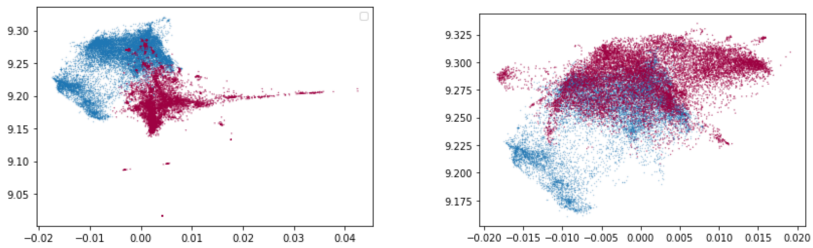
4.1 Data

Our analysis focuses on two synthetic datasets: subsets of the MS MARCO benchmark [9]. MS MARCO is a popular large retrieval dataset containing Bing user queries. The entire corpus contains 8,841,823 million passages. The train set contains 532,761 total (query, passage) pairs. There are 6,980 test queries. For computational feasibility, in whole we pull 50k (query, passage) pairs from the train set, every one of the 6,980 pairs from the test set, and an additional 150k documents from the corpus. In this way, we can simulate the corpus on a smaller scale, while increasing the ratio of gold passages to gain insight on our later fine-tuning procedure.

Using the approach described in Section 3.2, we first apply the UMAP dimensionality reduction to collapse the 768-dimensional passage embeddings (i.e. data) into 40-dimensional embeddings. Then we use K-means to split MS MARCO into 23 clusters based on the Gap statistic. In Figure 3 in the Appendix we plot the cluster sizes to verify all have a reasonable number of samples. In Figure 2 in Appendix A, we show the Gap statistic and the optimal number of clusters.

We denote the two furthest apart clusters as "home" and "away" and pick "home" randomly between these two. In Figure 1, we show a visualization of (a) "home" and "away" as well as (b) "home" and the closest neighbor to it. As the embeddings are 40-dimensional, we select two randomly chosen axes for the figure.[3] Out of concern for small data issues in the resultant clusters, we group these 23 clusters into two groups that partition the passage embeddings: the "home" groups and the "other" groups. For clarity, we illustrate via an example when $random\_seed = 2$. We observe that clusters 12 and 22 are furthest apart. Now we group everything closest to the "home" cluster until we reach roughly half the number of total documents (roughly 100k). We then fine-tune our experiments on the queries tied to the passages in the "home" groups.

$$\text{"home" groups} = [12, 20, 3, 13, 0, 19, 4, 14, 8, 11, 7, 18, 16]$$
$$\text{"other" groups} = [22, 10, 2, 21, 6, 9, 1, 5, 15, 17]$$



(a) "home" & "away" Clusters (b) "home" & "nearest" Clusters

Figure 1: All units are given in Euclidean distance. Here we see a stark contrast between far and near clusters along two axes. Our '"home" cluster is denoted in blue. We observe a clear separation between the cluster furthest away from "home" in (a), while we see significant overlap in (b). In addition, (b) is zoomed in (see axes). This agrees with our intuition and is robust to the chosen axes. Furthermore, we reason that one passage can be close in embeddings to multiple queries, and similarly multiple passages can be close to one query. As an alternate specification, we create our dataset with the same procedure described above, except clustering by queries. Using 200k queries with gold

[3]Note this is in contrast to PCA where it is natural to choose the first two principal axes, as UMAP does not have this same interpretable structure.

passages in the train set, we use the BERT encoder to embed the queries and cluster/chunk them. We then group the associated passages and add in another 300k random passages from the corpus to account for the fact that not all passages are tied to queries. We focus our analysis on the case of passage embeddings as the results were qualitatively similar. For additional details, see Appendix C.

N.B. For the remainder of the paper, note that our "home" chunk is synonymous with in-distribution data, and our "other" chunk is synonymous with out-of-distribution data.

4.2 Small Chunks

In contrast to the main specification given above, we also consider breaking the domain into five chunks. We hypothesize there exists a relationship between OOD relevance scores and (1) the distance between the ID and OOD clusters, and (2) the distribution within each cluster. In order to assess (1), we first form our "home" chunk as before, but we stop at $\sim$40k passages. The other chunks are formed by picking $\sim$40k passages and bucketing them into OOD groups by distance away from our "home" chunk. We then perform vanilla fine-tuning on our "home" chunk and evaluate on the other four OOD chunks to compare the retrieval relevance score degradation. For (2), we examine average Euclidean distance between UMAP embedding points.

4.3 Experimental details

Evaluation As in [6], we use the Normalized Cumulative Discounted Gain (NDCG) as a unified metric that strikes balance between binary and graded relevance while accounting for multiple retrieved documents. In particular, we compare the NDCG@10 scores across ID and OOD domains to evaluate degradation. We provide additional discussion on our choice of metrics in Appendix D.

Model We use Sentence-BERT (SBERT) as the dense retriever baseline, given its simplicity and competitive performance[5]. SBERT is a bi-encoder Siamese architecture that encodes a question and document with a BERT language model to produce sentence embeddings q and d. It is pretrained on SNLI with a classification objective and finetuned with a regression objective — computing the cosine similarity between q and v, it minimizes the mean squared-error loss. During retrieval, query embedding q is used to retrieve the top-k documents $d_1, ..., d_k$ with the highest *retrieval scores* according to maximum inner product search over the dense corpus.

Baselines We start with a distilBERT model pretrained on the full MS MARCO passage ranking dataset in an attempt to eliminate bias towards particular subpopulations, though this a possible limitation of using our syntthic datasets. We compare the following fine-tuning baselines: (1) fine-tuning on questions corresponding to the in-distribution training questions and evaluating on both ID and OOD test questions, (2) fine-tuning on a question set of the same size as (1), but randomly splitting the training questions, (3) no fine-tuning, and (4) fine-tuning with BM25-mined hard negatives. Ablating hyperparameters, we found a learning rate of 2×10^{-6}, batch size of 48, weight decay of 0.1 to work best. On a RTX3090 GPU, training time requires 8 hours with BM25 on a joint home + other corpus for 10 epochs and 1 hour for vanilla finetuning on home chunk.

4.4 Results

In Table 1 we show our results for the main data specification, where our "home" ID chunk and "other" OOD chunk partition the total passages into $\sim$100k halves.

Table 2 we show key results for the small chunk specification described in Section 4.2. For clarity, a denotes the ID small "home" chunk while $b - e$ denote four OOD small "other" chunks in decreasing Euclidean distance from a.

5 Analysis

We begin with an analysis of our fine-tuning results on our main data specification of one "home" and one "other" group, followed by results of fine-tuning on BM25-mined hard negatives. The remainder of this section investigates mechanisms behind degradation in our ID "home" chunk data.

	Evaluation on ID and OOD test queries	
	"home" ID q_{test}	"other" OOD q_{test}
Fine-tuning on ID q_{train}	0.6417	0.6619
Fine-tuning on ID q_{train}, random	0.6292	0.6843
No Fine-tuning	0.7710	0.8334
BM25-retrieved negatives	**0.7767**	**0.8371**

Table 1: Baseline retrieval results on different fine-tuning regimes in our main data specification with two chunks. Fine-tuning ran for 10 epochs with $learning\_rate = 2e-4$. All relevance scores reported use the NDCG@10 metric.

	Evaluation on ID and OOD test queries				
	a q_{test}	b q_{test}	c q_{test}	d q_{test}	e q_{test}
Fine-tuning on q_{train} from a	[0.7165, 0.7277]	0.8751	0.8135	0.8722	0.8147
No Fine-tuning	[0.7121, 0.7202]	0.8742	0.8310	0.8839	0.8062

Table 2: Baseline retrieval results for small chunks (pairwise comparisons of one ID "home chunk" a and four OOD chunks $b-e$). Note the ID q_{test} NDCG@10 score is stable across comparisons, and thus a simple range is given. Fine-tuning ran for 10 epochs with $learning\_rate = 2e-6$. All relevance scores reported use the NDCG@10 metric.

5.1 Vanilla Fine-tuning

As shown in Table 1, we observe some expected and some counter-intuitive results in our vanilla fine-tuning experiments. First, we observed that fine-tuning on in-domain data degraded in-distribution performance less than fine-tuning on random training data of the same size, which is consistent with what we would expect. However, we saw that in-distribution performance significantly dropped in comparison to no fine-tuning at all. We would expect that the retriever learns better representations for the passages and queries when testing on the in-distribution split, so this result is surprising. Because the model is not exposed to a variety of contrastive pairs during training — if new test questions that ask about similar passages are actually quite different from the training questions about that passage cluster, the model may not score the cluster passages highly.

Second, our initial hypothesis was that mixed-distribution retrieval may be more difficult than pure-OOD retrieval because the ID-retriever biases the selection of more ID passages, so OOD questions may receive fewer relevant passages. In our initial analysis, we observed that our synthetic splits *did not* support this hypothesis. Instead, we observe that when the model is trained on either the synthetic splits based on questions clustering, or splits based on passage clustering, the in-distribution questions retrieve an average of 70% in-distribution passages in their top-1 hits, while the OOD questions retrieve an average of 97% in-distribution passages in their top-1 hits. We do not observe any evidence of the degradation for OOD passage retrieval when tested on OOD questions.

5.2 BM25-mined Hard Negative Fine-tuning

BM25 is a TF-IDF variant of sparse retrievers based on exact lexical match. It has been shown to be robust to OOD generalizations and often considered a strong baseline [6]. Remarkably, we note that despite underperforming a non-finetuned SBERT by +10% when trained on the "home" chunk, by hard negative mining with BM25-retrieved passages most similar to groundtruth, fine-tuning on a 200K subset for only 1 epoch outperforms pretraining on the full MSMARCO (8.8M passages) dataset in both ID and OOD settings. We observed consistent gains across different distribution shifts and note that the sample efficiency in improving a pretrained model on the full distribution is a novel observation.

5.3 Cluster Analysis

Qualitative Evaluation From our discussion on vanilla fine-tuning, because it is likely that our synthetic domain splits strongly influence our results, here we analyze the underlying clusters.

We find that the clusters we construct do contain queries or passages which are similar in content beyond syntactical structure. As expected, the BERT encoder proves adept at grouping specific topics (e.g. medicine, finance) together. In Listings 1-6 in Appendix B, we provide further examples of cluster contents based on passage clustering. In Listings 7-9 in Appendix C, we show the same analysis for clustering by queries. Based on the way our chunks were constructed, we would expect to see contents in the "home" chunk to be more similar than those in the "other" chunk, even though the underlying clusters have distinct topics. This is due to the fact that while "home" chunk was aggregated by those closest to the base "home" cluster, the "other" chunk was just everything further away from it. In Listings 1-3 and 4-6 in Appendix B, we show that our resulting chunks do exhibit this desired behavior. For example, our "home" chunk consists of broadly medical topics while our "other" chunk consists of things from movies to climate.

Quantitative Evaluation The discussion above and the large degradation in ID performance motivates more concrete metrics and further investigation into our underlying clusters. In particular, the mechanism behind degradation is not entirely clear; the following analysis provides a first-pass at understanding how degradation behaves relative to the clusters.

As our clustering was based on Euclidean distance, we choose to use pairwise Euclidean distance between embeddings in the UMAP 30-dimensional space to evaluate how diffuse each chunk is. We perform this test on the entire embedding space as well as the "home" and "other" chunks and report our findings in Table 3.

	"home" chunk	"other" chunk	entire domain
Mean $d_{i,j}$	3.66	3.14	4.29
SD $d_{i,j}$	1.74	1.47	1.91

Table 3: Summary statistics for the pairwise Euclidean ℓ_2 distance between a random subsample of 5000 embeddings $i, j \in$ chunk $\subset \mathbb{R}^{30}$ for both "home" and "other" chunks. Standard errors via bootstrapping are very small and suppressed for clarity. In addition, while the mean is higher than the median, it is within $0.1SD$.

Contrary to the qualitative results above, we surprisingly observe that the both the mean and standard deviation of the pairwise distances of our "home" chunk are higher than that of the "other" chunk. Further, the mean pairwise distance is within a standard deviation of that of the entire domain, suggesting that "home" and "other" are very noisy and distributionally similar to the entire domain. We hypothesize that fine-tuning on one chunk may lead to overfitting to the train set queries.

To evaluate this hypothesis further, we turn to our approach from Section 4.2 for a finer breakdown of the space into 5 sub-domain chunks: $a - e$. In Table 4, we see via that a is significantly more concentrated, which is in line with our construction of the chunks. While a has passages that are similar in topic, $b - e$ simply represent chunks in decreasing order of distance from a, and thus are not necessarily grouped by content themselves. If our hypothesis of overfitting is correct, we would expect to see even higher degradation across the board. However, from our results in Table 2, we see no such result, indicating that the spread within chunks is not driving the degradation.

	a ("home")	b	c	d	e
Mean $d_{i,j}$	1.75	3.06	2.87	2.45	2.87
SD $d_{i,j}$	0.97	1.62	1.59	1.15	1.50

Table 4: Summary statistics for the pairwise Euclidean ℓ_2 distance between a random subsample of 5000 embeddings $i, j \in$ small chunk $\subset \mathbb{R}^{30}$ for all small chunks as described in Section 4.2. **Note** a **represents the small "home" ID chunk while** $b - e$ **represent the small "other" OOD chunks.** Standard errors via bootstrapping are very small and suppressed for clarity. In addition, while the mean is higher than the median, it is within $0.1SD$.

5.3.1 Vanilla Fine-tuning on Small Chunks

As per Section 4.2, we break our domain into five smaller chunks with the goal of studying how performance on OOD degrades as a function of distance. From the results in Table 2, we observe that distance seems uncorrelated with OOD relevance degradation. Coupled with Table 4, our preliminary exploration suggests the distribution of the chunks also does not influence results. In particular, we see that differences in mean pair-wise Euclidean distance do not correlate well with the OOD degradation. While these metrics are quite noisy, it provides motivation for further investigation. Note lastly that the in-domain a relevance scores are consistently lower than those of $b - e$. We believe this can be explained by the fact that a is highly concentrated in comparison to the other OOD chunks; test queries on a must retrieve from a much smaller, more precise set of embeddings, which on average reduces retrieval scores when the test queries in the embedding space are not close to a.

Surprisingly, we also see that performance on a and $b - e$ does not degrade from the no fine-tuning regime, contrary to what we saw in the main data specification with two large chunks.[4] This suggests that while cluster size and spread may not be related to relevance scores *after* fine-tuning, they significantly influence the fine-tuning procedure and lead to variable results downstream.

6 Conclusion

Main Findings Dense retrievers are highly sensitive to the training strategy and data selection. We proposed a novel method to construct synthetic multi-distribution retrieval settings, showed that vanilla fine-tuning can degrade performance, and that BM25 fine-tuning is consistently helpful for generalization. Furthermore, our first-pass synthetic split analysis suggests that it is that neither embedding distance nor distributional differences play significant factors in OOD degradation.

For existing retriever training datasets, corpora are often orders of magnitude larger than the training datasets used to train the question and passage BERT encoders. For example, on the MS MARCO benchmark, the number of documents in the corpus is 18x larger than the number of training pairs, so several documents are not incorporated during training process. Our setting, conditioned on further investigation, could implications for how to design question answering benchmarks with good coverage over question and passage types. Overall, we hope this work encourages further attention towards retrieval strategies that account for the sub-distributions in the background corpus.

6.1 Limitations & Future Work

We considered multi-distribution retrieval performance under several regimes, but the primary mechanisms behind degradations are still opaque. Future work is largely focused on understanding the mechanisms behind different fine-tuning regimes and the underlying domains.

While we have made progress in understanding the underlying clusters, the different choices and challenges in clustering and aggregating reflect the vast generality that real ID/OOD data can exhibit. Here we have focused primarily on the case where we aggregate small clusters based on distance away from a designated "home" cluster. The fundamental assumption here was that Euclidean distance in the embedding space had a strong impact on retrieval scores.[5] We analyzed this case for two and five chunks, with both not exhibiting any obvious trend in OOD relevance score degradation.

As suggested by our exploration into the clusters, given that $b - e$ were far more disperse than a (as they were simply grouped by distance away from a), a promising avenue would be to keep our small "home" chunk a the same while finding another small "other" chunk that is also concentrated. This would give us some more insight into whether distributional inequalities impact retrieval scores, as well as a more representative depiction of the ID/OOD setting we aim to mimic. In order to avoid small sample problems, as next steps we will double the sampled corpus size from $\sim$200k to $\sim$400k documents. Lastly, we should perform a set of robustness checks on our results by designating different "home" clusters and seeing if and how our results change.

[4] As a robustness check, we also reverse the fine-tuning procedure and fine-tune on the OOD chunk and evaluate on both. The results are also very similar here.

[5] We also performed the same experiments using cosine similarity as a metric and found qualitatively similar results.

References

[1] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *arXiv:2112.04426v2*, 2021.

[2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*, 2017.

[3] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations (ICLR)*, 2019.

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

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[6] Nandan Thakur, Nils Reimers, Andreas Ruckle, Abhishek Srivastav, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2021.

[7] Mandy Guoa, Yinfei Yanga, Daniel Cera, Qinlan Shenb, and Noah Constant. Multireqa: A cross-domain evaluation for retrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, 2021.

[8] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Towards unsupervised dense information retrieval with contrastive learning, 2021.

[9] 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. Ms marco: A human generated machine reading comprehension dataset. In *30th Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[10] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *arXiv:2109.08535v3*, 2022.

[11] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *arXiv:2007.00808*, 2020.

[12] Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen tau Yih, Barlas Oguz, Veselin Stoyanov, and Gargi Ghosh. Multi-task retrieval for knowledge-intensive tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

[13] Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering. In *arXiv:2012.14610v2*, 2021.

[14] Dan Friedman, Ben Dodge, and Danqi Chen. Single-dataset experts for multi-dataset question answering. In *arXiv:2109.13880v1*, 2021.

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2017.

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

[17] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. In *arXiv:1802.03426*, 2020.

[18] Joseph F. Hair, William Black, Barry Babin, and Rolph Anderson. *Multivariate Analysis*. Pearson, 2012.

[19] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. Royal Statistical Society, 2001.

[20] Matthew Henderson, Rrami Alrfou, Brian Strope, Yun-Hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Milkos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. arXiv:1705.00652v1, 2017.

A Appendix: Cluster Number & Balance

Gap Statistic Letting $d_{ii'}$ be the standard $\ell_2(i, i')$ norm, with data separated into k clusters $C_1, \ldots, C_k$, the Gap statistic is given by:

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

where $W_k = \sum_{r=1}^{k} (1/2|C_r|) \sum_{i,i' \in C_r} d_{ii'}$. The term $E_n^*\{\log(W_k)\}$ denotes the expectation of $\log(W_k)$ under a "null" sampling distribution of size n. To understand this, first note that W_k (and the term $\log(W_k)$) is small when the clusters are compact, as it is the average of the pairwise distances of points in a cluster. In order to make a meaningful statement of what "small" means, the authors propose comparing the W_k values to those obtained under a null distribution of data (the term $E_n^*\{\log(W_k)\}$). In particular, with $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$, we can create this null distribution by going along each dimension and sampling n points according to $\tilde{x}_i \sim U(\min\{x_i^{(1)}, \ldots, x_i^{(n)}\}, \max\{x_i^{(1)}, \ldots, x_i^{(n)}\})$ for $i = 1, \ldots, d$. In other words, we sample based on d different uniform distributions whose endpoints are given by the minimum and maximum values observed in each dimension in the data. We then calculate the W_k under this null distribution; repeating this sampling procedure multiple times yields the expectation $E_n^*\{\log(W_k)\}$ we see above. The larger the gap between these two terms for a given k, the more confident we are that k is the correct cluster size.

However, the authors note the tradeoff between model parismony and a higher Gap statistic. In particular, even if the gap statistic increases from k to $k+1$, we stop at k if Gap_{k+1} is within one standard error of Gap_k.

In Figure 2 we show the Gap statistic and the optimal number of clusters determined by the Gap statistic algorithm. In Figure 3 we show a histogram of cluster sizes to illustrate decent balance among clusters. For different random seeds we see slightly different results; however, qualitatively this picture is robust, in that every cluster has several thousand observations.

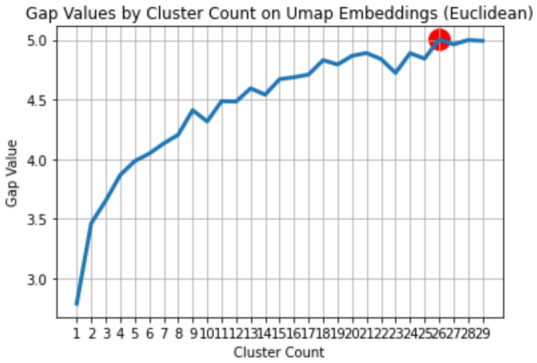


Figure 2: Gap Statistic

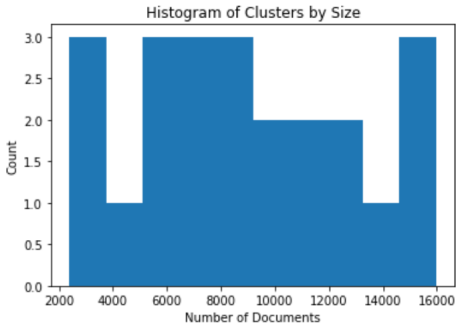


Figure 3: Histogram of the Number of Passages in each Cluster

B Appendix: Clustering by Passages

Here we provide a qualitative look into the clusters when we cluster by passages. Recall these passages are encoded by a BERT model and then projected down into 30-dimensional space via UMAP. We turn our attention to our first experiment, with our "home" and "other" chunks that partition the entire domain. In particular, with 23 clusters in whole and $random\_seed = 2$, we observe that clusters 12 and 22 are furthest apart. As described in the main text, we create our chunks by grouping everything closest to the "home" cluster until we reach roughly half the number of total documents (roughly 100k).

$$\text{"home" chunk} = [12, 20, 3, 13, 0, 19, 4, 14, 8, 11, 7, 18, 16]$$
$$\text{"other" chunk} = [22, 10, 2, 21, 6, 9, 1, 5, 15, 17]$$

We show cluster contents for the first three clusters in each chunk.

```
The National Women's Health Information Center lists these common
    causes of urinary tract infections in women: 1  Wiping from
    back to front after a bowel movement. 2  Having sex. 3
    Holding urine for too long.  Being 1  diabetic. Having a
    kidney stone or other factor that makes it difficult to
    urinate.  Producing less estrogen, such as after menopause.


Causes of insomnia in women: Most causes of insomnia in men and
    women are same but some common causes of insomnia in women
    include: Stress: The demanding lifestyle in today's world (
    responsibility of the family and at work) can cause fatigue
    and stress. Stress is one of the commonest causes of insomnia.


Systemic inflammatory response syndrome (SIRS) is an inflammatory
    state affecting the whole body. It is the body's response to
    an infectious or noninfectious insult. Although the definition
     of SIRS refers to it as an inflammatory response, it actually
     has pro- and anti-inflammatory components.
```

Listing 1: Passage Clustering, Cluster 12 Passages

```


medical Definition of botulinum toxin : a very powerful neurotoxin
     that causes botulism and is produced by the botulinum
    bacterium (Clostridium botulinum); also: botulinum toxin type
    a Note: Botulinum toxin acts primarily on the parasympathetic
    nervous system.


Diarrhea is a common side effect of a detox cleanse. In fact,
    proponents of these types of diet plans claim that the
    diarrhea is a sign your body is ridding itself of toxic
    substances.


But it's never too late to get treatment.. Even for veterans in
    their 70s and 80s, a combination of psychotherapy, medication,
     and marital and family therapy can reduce PTSD symptoms,
    including insomnia, anxiety and irritability, he said. The
    Vietnam War ended in 1975.
```

Listing 2: Passage Clustering, Cluster 20 Passages

```
The ciliary body controls the shape of the lens. The ciliary body
    is composed mainly of smooth muscle and is connected to the
    lens by suspensory ligaments which are not visible in this
    image. Contraction of smooth muscle in the ciliary body makes
    the lens rounder focusing vision on objects which are closer
    to the eye.


Function of the Olfactory Nerve. The olfactory nerve is
    responsible for your sense of smell and partially responsible
    for your sense of taste. It is also known as cranial nerve 1
    because it is the shortest of the cranial nerves and one of
    only two nerves (the other is the optic nerve) that bypass the
     brain stem and connect directly to your brain.


The largest unit within which gene flow can readily occur is a
    species.
```

Listing 3: Passage Clustering, Cluster 3 Passages

```
Climate data for fairbanks intl, Longitude: -147.876, Latitude:
    64.8039. Average weather Fairbanks, AK - 99709 - 1981-2010
    normals. Jan: January, Feb: February, Mar: March, Apr: April,
    May: May, Jun: June, Jul: July, Aug: August, Sep: September,
    Oct: October, Nov: November, Dec: December.


3059 hours. Av. annual snowfall: 47 inch. Climate data for Salt
    Lake City, UT - 84116 - 1981-2010 normals-weather. Jan:
    January, Feb: February, Mar: March, Apr: April, May: May, Jun:
     June, Jul: July, Aug: August, Sep: September, Oct: October,
    Nov: November, Dec: December.


Climate data for Denver, CO - 80201 - 1981-2010 normals - weather
    Jan: January, Feb: February, Mar: March, Apr: April, May: May,
     Jun: June, Jul: July, Aug: August, Sep: September, Oct:
    October, Nov: November, Dec: December
```

Listing 4: Passage Clustering, Cluster 22 Passages

```
Flying time from Chicago, IL to Cairo, Egypt. The total flight
    duration from Chicago, IL to Cairo, Egypt is 12 hours, 47
    minutes. This assumes an average flight speed for a commercial
     airliner of 500 mph, which is equivalent to 805 km/h or 434
    knots. It also adds an extra 30 minutes for take-off and
    landing. Your exact time may vary depending on wind speeds.


Thornton Hotels. Candlewood Suites Denver North - Thornton.
    Welcome to the Candlewood Suites Thornton serving Northglenn,
    Westminster, Brighton, Commerce City, Broomfield, and Thornton
    . We are minutes from downtown Denver, many RTD stations,
    Water World, downtown Louisville, and the Broomfield event
    center.
```

```
4. Hampton Inn Denver-North/Thornton. From Business: Property
    Location With a stay at Hampton Inn Denver Thornton in
    Thornton, you'll be close to Boondocks Food and Fun and
    Thorncreek Golf Course. This hotel is with. Add to
    mybookRemove from mybook.
```

Listing 5: Passage Clustering, Cluster 10 Passages

```
James A. Owen is an American comic book illustrator, publisher and
     writer. He is known for his creator-owned comic book series
    Starchild and as the author of The Chronicles of the
    Imaginarium Geographica novel series, that began with Here,
    There Be Dragons in 2006. 1 Career.


web video star Lil Moco born on 02 11 1994 in . Until now, Lil
    Moco's age is 22 year old and have Scorpio constellation.
    Count down 361 days will come next birthday of Lil Moco !


The Emoji Movie is a 2017 American 3D computer-animated comedy
    film directed by Tony Leondis, and written by Leondis, Eric
    Siegel and Mike White, based on the trend of emojis. It stars
    the voices of T. J. Miller, James Corden, Anna Faris, Maya
    Rudolph, Steven Wright, Rob Riggle, Jennifer Coolidge,
    Christina Aguilera, Sofia Vergara, Sean Hayes and Patrick
    Stewart. The film centers on Gene, a multi-expressional emoji
    who lives in a teenager's phone, and who sets out on a journey
     to become a ...
```

Listing 6: Passage Clustering, Cluster 2 Passages

C Appendix: Clustering by Queries

Now we examine the clusters when clustered by queries rather than passages themselves. The BERT encoding again does a good job of embedding similar topics.

```
what are pssa scores used for
what are requirements for minor in spanish at uta
what are the crimes classified as economic corruption and
    financial crimes
what are the profits
what are the requirements for bethune cookman
what are the salaries of big bang actors
what are the tax benefits of a heloc
can employer pay for individual health policy
case can arrive at the supreme court through each of these ways
    except
what do lenders use as you dti
what do partnerships file tax in michigan
what do quantitative risk analysts do
what do rich people complain about
```

Listing 7: Query Clustering, Cluster 3

```
does schizophrenia cause hallucinations
vasospasms caused by what
cad heart related
what are aneurysm
what are signs of anxiety in your chest
what are the complications of varicose vein
can a pinched nerve cause tooth pain
what are the early signs of colon cancer?
what are the most common causes of paralysis
what are the signs of allergies in the winter time
what are the signs of kidney failure in dogs with dm?
what can be reason of excess urination
what cause dizziness mayo clinic
```

Listing 8: Query Clustering, Cluster 7 Queries

```
the small intestine is small in what
does e coli feed on
what are of common to both aerobic and anaerobic respiration
what are some tissues found in the skin
during which phase does a cell spend the majority of its life
    cycle?
what are the duplicated strands of dna called
what are the five phases of the cell cycle?
what are the mycorrhizae
what are the products and by products of photosynthesis?
what are the two major subdivisions of the nervous system?
what are your upper two teeth called
what body parts does pku affect
what cells don't go through mitosis
```

Listing 9: Query Clustering, Cluster 21 Queries

D Appendix: Additional Experimental Details

NDCG@k Here we provide additional details about the NDCG@k metric. For a specified discount function $\lambda(r)$, where $r \geq 1$ is the rank of a document, the discounted cumulative gain (DCG)

metric of our retriever is a weighted sum of the degree of relevancy of the ranked items, defined by $\sum_{r=1}^{n} \lambda(r) f(r)$ where $f(r) = rel_i$. NDCG has the advantage of rewarding higher degrees of relevancy for documents while downweighting lower ranked relevances. NDCG normalizes DCG by the Ideal-DCG score of the ground truth ranking to be within $[0, 1]$ and we use a cut-off NDCG@$k = 10$, setting the discount $\lambda(r) = 0$ for ranks $k > 10$.