To Adapt or to Annotate: Challenges and Interventions for Domain Adaptation in Open-Domain Question Answering

Dheeru Dua[1]* **Emma Strubell**[2]* **Sameer Singh**[1] **Pat Verga**[3]
[1]University of California, Irvine [2] Carnegie Mellon University [3] Google DeepMind

Abstract

Recent advances in open-domain question answering (ODQA) have demonstrated impressive accuracy on general-purpose domains like Wikipedia. While some work has been investigating how well ODQA models perform when tested for out-of-domain (OOD) generalization, these studies have been conducted only under conservative shifts in data distribution and typically focus on a single component (i.e., retriever or reader) rather than an end-to-end system. This work proposes a more realistic end-to-end domain shift evaluation setting covering five diverse domains. We not only find that end-to-end models fail to generalize but that high retrieval scores often still yield poor answer prediction accuracy. To address these failures, we investigate several interventions, in the form of data augmentations, for improving model adaption and use our evaluation set to elucidate the relationship between the efficacy of an intervention scheme and the particular type of dataset shifts we consider. We propose a generalizability test that estimates the type of shift in a target dataset without training a model in the target domain and that the type of shift is predictive of which data augmentation schemes will be effective for domain adaption. Overall, we find that these interventions increase end-to-end performance by up to ~24 points.

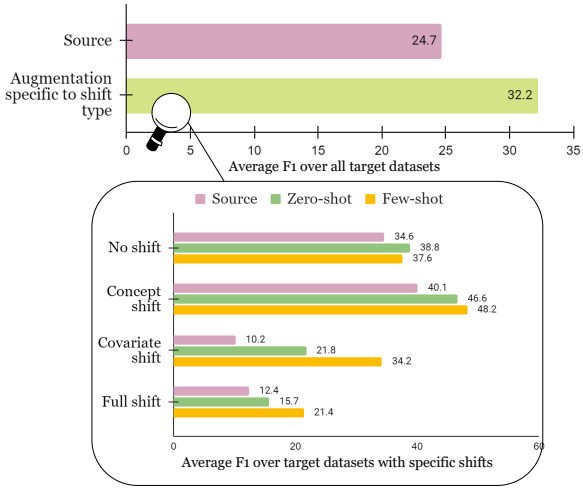


Figure 1: **Effect of interventions on dataset shifts**. *Top:* Average end-to-end performance of source domain model is quite poor when applied to OOD datasets. Source model (trained on general-purpose domain) performance improves when adapted to unseen target domain with interventions. *Bottom:* Drill-down of performance into zero and few-shot data augmentations averaged over target datasets exhibiting these shifts shows covariate and concept shifts respond to zero and few-shot data augmentations. Target datasets with No shift do not improve much with any intervention while full shift benefits most from Few-shot.

1 Introduction

General-purpose open-domain question answering (ODQA; Chen et al. (2017); Lee et al. (2019); Izacard et al. (2022)) is an important task that automates reading and understanding a large corpus of documents to answer a given question succinctly. It is especially crucial in fields such as biomedicine, legal, news, etc., where more documents are added daily, outpacing the speed at which a user can process the information. Current state-of-the-art ODQA systems perform a two-stage pipeline process (Izacard et al., 2022): 1) Given a question and document corpus, a *retriever* (Karpukhin et al., 2020; Izacard et al., 2021; Raffel et al., 2020) selects relevant passages and 2) a question answering model, also known as a *reader* (Izacard and Grave, 2021) answers the given question based on the retrieved passages. This decoupling allows for independent advances in domain adaptation of general-purpose retrievers (Thakur et al., 2021) and question-answering (Fisch et al., 2019) models.

To enable practical application, an ODQA system should assist humans in keeping up with new knowledge without requiring annotations for every new domain or concept. For this, the system should be resilient to changes in the document, question,

*This work was done while authors were at Google.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14429–14446

and answer distributions. Unfortunately, the current work in ODQA focus solely on Wikipedia corpus and do not study effectiveness of a model trained on such a general-purpose domain when applied to an unseen domain. To gauge how likely it is for a source domain model to succeed on an unseen domain we need to understand its ability to work out-of-the-box or even adapt to a new target domain, under varying types and degrees of dataset shifts. (Quinonero-Candela et al., 2008).

In this work, we study the challenges and interventions for generalizing ODQA models to new domains via four contributions. First, to understand how well the state-of-the-art ODQA system (trained on the general-purpose domain) performs on a variety of target distributions, we define a collection of datasets for evaluating domain generalization. We aggregate a set of seven ODQA datasets spanning five different domains (§2). We observe that the source ODQA model does not generalize well (Fig.1, Top) on this collection (§4). Second, to automatically determine the type of data shift with only a small number of labeled target domain examples, we propose a *generalizability test*. This test assesses the type and degree of shift a new domain suffers with respect to the source domain (§3). Third, to understand the adaptability of the source model to a target domain, we analyze the performance of various intervention schemes, including existing zero-shot in-domain question generation and a novel few-shot language model-aided generation. These schemes create data akin to the target domain which is augmented with the source domain to learn an adapted version of the source model. Overall, we observe improvement in performance across all the target datasets (Fig. 1). The degree of improvement depends on the intervention scheme and underlying dataset shift (§5). Finally, we propose a simple and effective few-shot method that improves the performance by up to 24% in F1. This method prompts a large language model with 8 examples to generate examples for further adaptation.

Putting it all together, we use the generalizability test to gauge the type and degree of dataset shift in a target dataset. Then, we empirically show that certain types of dataset shifts respond well to specific intervention schemes (§5, Fig. 1). This helps ascertain whether we can adapt a source model to unseen domain with minimal supervision. The resources used in this work are released at https://github.com/dDua/adapt-or-annotate

2 Background and Evaluation Setup

An ODQA model learns interactions among three random variables: Question ($\mathbb{Q}$), answer ($\mathbb{A}$) and context corpus ($\mathbb{C}$). For a given $q \in \mathbb{Q}$, first the retriever $\mathcal{R}$ returns a set of passages, $C_q = \mathcal{R}(q, \mathbb{C})$. These passages are then sent to a reader model $\mathcal{M}$ to obtain the final answer, $\hat{a} \leftarrow \mathcal{M}(a|q, C_q)$.

Following prior work, we evaluate retriever performance with the Acc@k metric, which computes if the oracle answer is found in the top-k retrieved passages[1]. We set k=100 in all of our experiments. For reader performance, we compute token-level F1 between the oracle and predicted answer.

2.1 Datasets

We test the generalization capabilities of a model trained on a *source domain* when applied to seven datasets in five very different *target domains*.

Source Domain: For source domain we use documents from English Wikipedia and QA pairs for supervision from NaturalQuestions (NQ) (Kwiatkowski et al., 2019) and BoolQ (Clark et al., 2019). We treat this domain as our source as it is used for the vast majority of current work in ODQA (and many other areas of language research). In addition to the supervised training data from NQ and BoolQ, we also consider cloze-style questions derived from the QA pairs in NQ. For each QA pair, we retrieve a sentence from Wikipedia with the highest BM25 similarity score. We convert the retrieved sentence into a cloze-style question by replacing the answer string in the sentence with sentinel markers (Raffel et al., 2020)[2].

Target Domains: For our target corpora, we re-purpose seven open-domain QA and/or reading comprehension datasets spanning five different domains (Stack Overflow, Reddit, Pubmed, Japanese Statute Law codes, CNN/DailyMail, and Wikipedia). The datasets Quasar-S (Dhingra et al., 2017), Quasar-T (Dhingra et al., 2017), SearchQA (Dunn et al., 2017) and BioASQ (Balikas et al., 2015) were introduced as ODQA

[1]The only exception is COLIEE dataset which primarily contains boolean (yes/no) answers so we instead use oracle passage to compute Acc@100.

[2]We use cloze augmentation for training reader models because some target datasets contain cloze-style questions, keeping the question distribution consistent across different experimental setups. We do not perform this augmentation for retrievers because we observed a performance drop in initial experiments.

datasets over Stackoverflow, Reddit, Wikipedia, and Pubmed corpus respectively. Additionally, we re-purpose reading comprehension datasets, NewsQA (Trischler et al., 2017) and CliCR (Šuster and Daelemans, 2018) as ODQA datasets, by retrieving a set of passages for each QA pair from Pubmed and CNN/Dailymail corpus. For COLIEE (Rabelo et al., 2022), we convert the original entailment questions into boolean questions and retrieve passages from legal code statutes provided with the task. We confirm that these reading comprehension datasets can be reasonably re-purposed for our ODQA setup by achieving a reasonable end-to-end performance of ODQA models trained on gold target domain QA pairs with BM25 retrievals from the target corpus (UB-Ret, Fig. 3).

2.2 Models

We compare four **retrievers**: (1) BM25 (Robertson and Spärck Jones, 1994) (sparse and unsupervised), (2) Contriever, semi-supervised with MS-MARCO (Izacard et al., 2021), (3) Dense Passage Retriever (DPR) (Karpukhin et al., 2020), and (4) the state-of-the-art source domain model Spider (Ram et al., 2022). DPR and Spider are dense and supervised. As for **reader**, we use the state-of-art fusion-in-decoder (FiD) model (Izacard and Grave, 2021) that uses the top 100 documents to generate the final answer.

3 Generalizability Test

There are many aspects that determine in what ways and to what extent one data distribution differs from another. It is often challenging to quantify the degree of *generalizability* or diverseness for a new domain without collecting enough samples to train a model in the new domain. To address this issue, we propose a method to assess the type and degree of diversity by utilizing only a few examples from the target domain as an evaluation set.

3.1 Types of dataset shift

Different types of dataset shifts (Quinonero-Candela et al., 2008) have been proposed in the literature but they are often studied in a classification setup. For our application, we consider *concept* and *covariate* shifts which are more amenable to our pipelined ODQA setup — with input as a joint distribution over question and contexts and output as a distribution over answers given question and contexts as input.

No shift occurs when the input and output distributions match across the source and target domains.

Concept shift (Widmer and Kubát, 2004) occurs when the input distribution of the source and target domains match, i.e., $p_s(x) = p_t(x)$ while the output distribution between source and target domain does not match, $p_s(y|x) \neq p_t(y|x)$.

Covariate shift (Zadrozny, 2004) occurs when the source and target input distributions do not match, i.e. $p_s(x) \neq p_t(x)$ while the output distributions match $p_s(y|x) = p_t(y|x)$.

Full shift occurs when both the source and target input and output distributions do not match.

3.2 Calculating shift for ODQA

We characterize the shift in ODQA as a two-step process. First, we compute the input distribution, i.e, the joint question and context distribution using un-normalized (energy) scores from a dense retriever (Karpukhin et al., 2020) that quantifies the compatibility between a given question, q and a context, c via $\mathcal{R}(q, c)$. Then, we normalize the scores from the retriever over a set of contexts. Ideally, the set of contexts should be the entire target domain document corpus, however, that can be prohibitively computationally expensive and also results in a high entropy distribution. Instead, we use a subset of contexts, C, from the entire corpus $\mathbb{C}$. We ignore the prior over questions since it remains constant when calculating the context distribution for a specific question. Instead, we approximate the joint with conditional distribution over contexts given question.

$$p(q, c) \propto \frac{\mathcal{R}(q, c)}{\sum_{c_k \in C} \mathcal{R}(q, c_k)} \tag{1}$$

In the second step, we test whether the output distributions match by computing the likelihood of generating the oracle answer given a question, q, and the relevant contexts, C_q. In an ideal scenario, we can do this by performing global normalization (Goyal et al., 2019) over all possible answer spans in the corpus which is intractable. Instead, we use a sub-sample of answers, $\mathcal{A}$, to compute the output distribution as shown below.

$$p(a|q, C_q) = \frac{\prod_t \mathcal{M}(a^t|a^{<t}, q, C_q)}{\sum_{a_k \in \mathcal{A}} \prod_t \mathcal{M}(a_k^t|a_k^{<t}, q, C_q)} \tag{2}$$

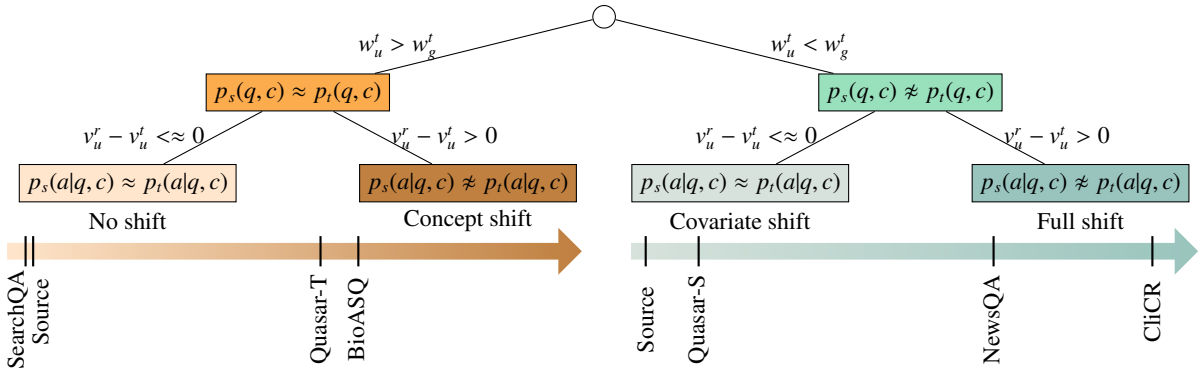


Figure 2: **Generalizability Test:** At the first level, we decide whether the input distribution is closer to the uniform distribution or gold. At the second level, the gradual increase from left to right in the leaf nodes depicts decrease in distance of output distribution from uniform. The target datasets at the bottom are placed based on distances in Table 1. The nodes represent if the source model $p_s(a|q,c)$ is compatible or not with the target dataset $p_t(a|q,c)$

Dataset	**Retriever** $(w_u^t - w_u^g)$	**Reader** $(v_u^r - v_u^t)$	Shift
BioASQ	0.30	0.17	Concept
CliCR	-0.88	0.23	Full
Quasar-S	-0.66	0.07	Covariate
Quasar-T	0.20	0.16	Concept
NewsQA	-0.19	0.18	Full
SearchQA	0.61	0.00	No

Table 1: **Wasserstein distance** computed over 100 labeled examples from the target set. The negative retriever value implies that the target dataset falls on the right side of decision tree at first level (Fig. 2).

3.3 Predicting type of dataset shift

To compute the type of shift (§3.1), we need a model trained on the target domain (p_t) which requires a large number of examples. However, our goal is to determine if a source model can be adapted to the target dataset with only a few examples for target evaluation. To do this, we conceptualize adapting or fine-tuning a pre-trained source model as a Bayesian framework. In this framework, the source model acts as a prior which when exposed to interventional data (for adapting) and target data (for fine-tuning), results in an adapted or fine-tuned posterior distribution. If the prior (source model) contains an informative signal with respect to the target dataset then we do not require much supervision to learn an effective posterior. However, if the prior is non-informative we need a lot of supervision to learn the posterior. Towards this end, we devise a *generalizability test*, where we use a small set of evaluation examples sampled from each target dataset to compute input and output distribution using the source domain model. Then, we compare these distribution with the a non-informative prior like uniform distribution and informative prior like the oracle distribution to gauge if the source model is closer to uniform or oracle distribution. This helps us assess the effectiveness of the source model towards the target dataset without having to train a model in the target domain.

Input/Retriever Distribution: To determine if the input distribution contains informative signal with respect to target evaluation set, we need to compute the distance of the input distribution from uniform and oracle distribution. To do this, we follow Eq. 1 and compute the input distribution, with passages from across examples in the entire target evaluation set as the subset for normalizer computation. Then, for a each question, we compute the Wasserstein distance, w_u^t, (Kantorovich, 1960) between the input distribution and the uniform distribution and average these values over all the examples in the target evaluation set. Similarly, we also compute the distance between the gold or oracle distribution and the input distribution as w_g^t. If $w_u^t > w_u^g$, we conclude that the target distribution is far from the uniform distribution and closer to the gold distribution, indicating that the source model is compatible with the target distribution (Fig. 2).

Output/Reader Distribution: In similar vein as input distribution, we need to compare the output distribution with corresponding uniform and oracle distribution over answers. To do this, we follow Eq. 2 and compute the output distribution, with set of answer spans from across all the examples in the target evaluation set for normalizer computation. Then, we compute the Wasserstein distance between the uniform and output distribution aver-

aged over the target evaluation set as v_u^t.

In an ideal scenario, we would compare the distance between and oracle and output distribution with v_u^t, similar to input distribution. However, empirically we find that output distribution is always closer to uniform than oracle, even when evaluated on source domain. We believe this is because of two reasons. First, the conditional answer generation model ($\mathcal{M}$) is not trained with a contrastive loss like the retriever, resulting in a high entropy answer likelihood distribution. Second, the support set of answers used for normalization contains only grammatically correct answer spans making the likelihood scores attenuated. To address these issues, we use a reference answer conditional distribution to de-bias the likelihood scores with a threshold. To obtain this threshold, we consider the source distribution as a reference and compute the distance between output distribution evaluated on examples from source evaluation set and the uniform distribution as v_u^r. Since the reference based output distribution is in-domain, it should be far from the uniform distribution and closer to oracle distribution. As a result, if $v_u^r - v_u^t$ is close to 0, we assess that the target is far from uniform and that source model is compatible with the target dataset.

In Figure 2, we put this altogether as a decision tree to identify the type of dataset shift. We observe that SearchQA falls under the *No shift* category as it is close to the source domain, hence, we conjecture that it will observe minimal improvements under most data intervention schemes as the source model already captures the target distribution (§5). We also conjecture that datasets falling under *Concept shift* and *Covariate shift* are more amenable to zero-shot data interventions, while, *Full shift* would benefit more from few-shot or in-domain annotations from the target domain. We consider few shot augmentations as a proxy for annotating examples in the target domain because they are generated with supervision from target dataset.

4 How Well do Models Generalize?

We test the OOD performance of the source model on target datasets and analyze the failures.

4.1 Reader Generalization

In Fig. 3, we test the end-to-end performance of three model variants:

Source: a reader trained with source dataset and contexts retrieved by BM25, demonstrating zero-shot generalization performance.

Upperbound-Reader (UB-READ): a reader trained on the target dataset with contexts retrieved by BM25 – the overall strongest retriever.

Upperbound-Retriever (UB-RET): a reader trained on the target dataset with gold contexts to approximate upper-bound performance.

We observe large performance drops when evaluating the source model on target domains (Fig. 3), especially when the target corpus differs from Wikipedia, such as in Quasar-S (Stack Overflow) and CliCR (PubMed), even though the model requires similar reading capabilities to those needed in the source domain. Interestingly, even though BM25 retriever accuracy is relatively high on the target datasets (Fig. 4, ~83% Acc@100 on Quasar-S), that accuracy does not translate to strong reader performance (Fig. 3, ~11% F1 on Quasar-S).

To understand this performance gap, we manually sample 50 predictions from each target dataset where retrieved passages contain the oracle answer but the reader produced an incorrect prediction. We observe that in **~65% cases, the Acc@100 metric yields a false positive**, where the passage contains an exact string match of the correct answer, but the context does not actually answer the given question. In other cases, the reader is unable to understand the context. For example, for the question: `What is the name of the office used by the president in the white house?` and answer: `oval`, the retrieved passage: `A tunnel was dug into the White House connecting the Oval Office to a location in the East Wing....` is credited (incorrectly) as context answering the question.

4.2 Retriever Generalization

We compare the zero-shot generalization of four retrieval models in Fig. 4. Spider, which is the best performing model on the source domain, exhibits improvement on SearchQA (~1%) (which is similar to source distribution), but shows large drops in performance when applied to the target datasets: ~40% on NewsQA, ~28% on Quasar-T and, Quasar-S. To understand the drop, we manually analyze 50 random incorrect predictions from Spider. We observe two major failure modes. First, we find that dense models are sensitive to changes in the length of contexts. When exposed to documents with heterogeneous lengths, models tend to over-retrieve shorter contexts. To

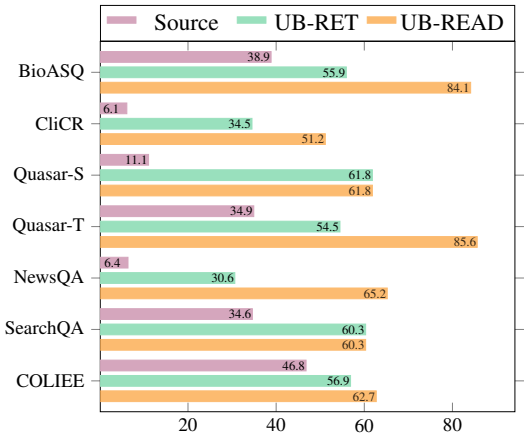


Figure 3: Reader performance on the target set without any interventions. SearchQA, Quasar-S and Quasar-T do not have gold passage annotations and so UB-READ does not improve over UB-RET. The majority voting baseline on COLIEE is 50.95.

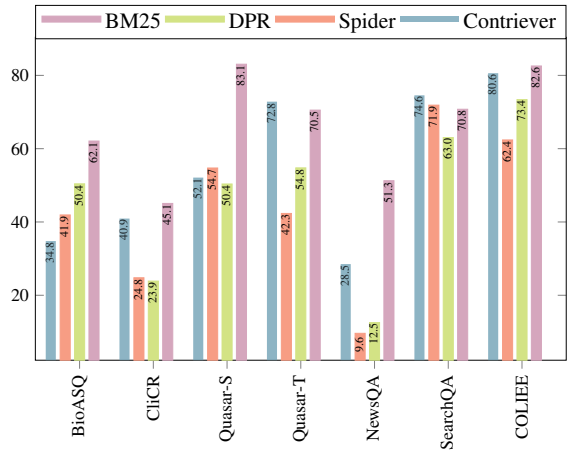


Figure 4: Retriever performance (Acc@100) without any interventions on target domain corpus

quantify the sensitivity to changes in lengths on source domains itself, we pool passages from all target corpus into a combined index. We observe that the performance of Spider when exposed to this combined index reduces by ~15% and restricting the minimum length of contexts to 50 words alleviates the problem and recovers the original performance. The second common failure mode occurs due to changes in distribution of entity types from source to target. For example, words like `plant` in `Which is produced in plants of narora kakrapar tarapur` refers to `power plant` in Wikipedia, while in case of PubMed it often refers to living organic matter (Sciavolino et al., 2021). Overall, BM25, being an unsupervised method, has the best performance across all domains.

5 Interventions for Improving Adaptation

Domain Adaptation is shown to be a causal intervention (Jin et al., 2021) mechanism to effectively understand impact of an augmentation technique without much concern about spurious correlations.

5.1 Zero-shot adaptation methods

We perform a set of zero-shot data intervention methods, where we consider the effect of change in distribution of each random variable: Question, answer and context one at a time, while keeping the other two fixed.

Varying context distribution To test the effect of change in context distribution, we pool passages from all corpora into a combined index. We observe that supervised models like Spider are sensitive to out-of-domain distractors, unlike BM25, especially when the target dataset uses same corpus as source (Wikipedia). For example, SearchQA suffers a performance drop of ~15%. On average we see a performance improvement of ~2% (w/o COLIEE) when the target index is changed to the combined index. BM25 still out-performs Spider on average by 19.1% with the combined index. However, we observe a drop in performance of the FiD reader of up to ~5% in F1 for NewsQA with the combined index. More details are in the appendix (Figs. 5 and 6.)

Augmentations	Retriever	Reader
Random	45.35	33.50
Uniform	50.02	39.07
Most frequent	39.33	38.18
BioASQ train answers	47.48	41.33

Table 2: Answer distribution: Retriever (DPR) and Reader (FiD with BM25 retrievals) F1 on BioASQ.

Varying answer distribution Many works (Gururangan et al., 2018; Dua et al., 2020; Jiang and Bansal, 2019) have shown that bias in the answer prior distribution can introduce spurious correlations in model learning. This effectively improves the model performance at the cost pf OOD generation. To test whether we can improve the performance of adapted source model by varying the answer distribution, we experiment with a variety of answer distributions over plausible set of answer spans. To obtain the set of answer spans, we extract and annotate coarse-grained entity types from the

Dataset	Retriever			Reader		
	Source	ClozeQA	QGen	Source	ClozeQA	QGen
BioASQ	50.41	48.0	45.4	45.3	49.4	46.4
CliCR	23.8	24.9	23.9	6.12	7.34	10.5
Quasar-S	50.3	66.8	68.2	10.2	21.7	17.4
Quasar-T	54.7	53.9	55.5	34.9	41.9	44.7
NewsQA	12.5	18.7	15.2	18.5	21.2	12.7
SearchQA	63.0	52.9	54.7	34.6	38.8	37.2
COLIEE	61.4	60.5	57.8	46.7	54.1	62.3

Table 3: Zero-shot: Comparing retriever (DPR) and reader (FiD with BM25 retrievals) performance on two types of question formats for augmentation.

target corpus using spaCy[3]. We use this coarse-grained entity type information as a set of classes from which to choose 50k entities with four different sampling strategies: Most frequent, uniform, randomly sampled based on entity type categories, and sampling in proportion to entity type distribution of answers in the target training set.

The source model has reasonable end-to-end performance on BioASQ, even with passages from the source corpus (Wikipedia), suggesting that it contains sufficient information for answering many BioASQ questions. Consequently, we select BioASQ for these controlled experiments (Appendix Fig. 6). This allows us to use the Wikipedia corpus alone for retrieval, which makes it easier to fix the passage distribution. In Table 2, we show that uniform sampling boosts retriever performance compared to random sampling, allowing the model to learn from all types of answers and generalize better to unseen answer distributions. On the other hand, the best reader model performance is when we know the correct answer distribution of the target dataset up front, showing that the answer priors influence reader performance. However, in a zero-shot setup, we do not have access to this distribution, so we adopt the second-best technique, uniform sampling from across the entity type categories, in the following experiments.

Varying question distribution We vary the question distribution by augmenting the source domain with QA pairs generated using two different methods. Our first approach (QGen) uses a question generation model (Subramanian et al., 2017) trained on the source domain to generate a question given a passage and an answer. This question generation model is applied to a new target passage and a plausible answer span from the passage (Shakeri et al., 2020; Krishna and Iyyer, 2019; Song et al., 2018; Klein and Nabi, 2019). The second approach (Cloze QA), which has been less explored previously, converts a sentence in the target corpus to a fill-in-the-blank style cloze question (Taylor, 1953) by masking a plausible answer span (entity mention) in the sentence. We sample answer spans uniformly based on an entity type distribution from the target corpus and then query our combined index to create a dataset containing cloze style questions aligned with relevant documents. We use these same sampled answers to generate standard QGen QA pairs as well. We combine these data interventions with our initial source domain data to train a DPR retriever and a FiD reader (Table 3). We observe similar average performance across both intervention types in retriever and reader models. However, cloze QA pairs are computationally much more efficient to generate as they do not require additional question generation models.

Discussion on generalizability test In §3, we hypothesized that datasets with less severe shift (Quasar-S, Quasar-T, and BioASQ) would show more performance improvements with zero-shot adaptation as compared to datasets with severe shift (CliCR and NewsQA). Indeed, we observe an avg. improvement of about 8.5% F1 on datasets having Concept and Covariate shift while only 3.5% F1 on datasets with Full shift in Table 3. Moreover, in Fig. 1, we see that target datasets with *No shift*, do not show much improvement with any intervention as the source model already captures the distribution. Datasets with *Full shift* need few-shot examples for better adaptation while datasets with *Concept* and *Covariate* shift are able to adapt with zero-shot data interventions.

5.2 Few-shot Generalizability and Adapatability

Zero-shot adaptation does not work well when the target distribution is far from the source. For these cases, we experiment with few-shot adaptation.

Few-shot data generation Zero-shot interventions like QGen are trained on the source and do not produce generations that are fully compatible with the target domain and thereby do not provide much useful signal. An alternative approach would be to train a question generation model with a few examples from the target domain. However, it is difficult to adapt or fine-tune a question genera-

[3]https://spacy.io/

tion and answering model (for validating QA pair correctness) with very few examples.

Dataset	Retriever		Reader		Closed Book
	Baseline	DataGen	Baseline	DataGen	(F1)
BioASQ	50.4	51.3	45.3	50.6	32.0
CliCR	23.8	29.0	6.12	19.4	10.8
Quasar-S	50.3	71.9	10.2	34.2	23.7
Quasar-T	54.7	55.4	34.9	45.8	**55.3**
NewsQA	12.5	22.7	18.5	23.3	8.67
SearchQA	63.0	63.3	34.6	37.6	**61.5**
COLIEE	73.3	82.2	46.8	61.1	53.0

Table 4: Both Closed Book and DataGen use eight few-shot examples from the target domain. Closed Book LLM contains 540B params while the Retriever and Reader contain 110M and 770M params respectively. Closed-book performance for NQ is 36.71.

To capture target distribution without a lot of supervision, we propose a few-shot technique (DataGen) that prompts a large language model (LLM; Chowdhery et al. (2022)) to generate a sentence given a passage. We use eight seed examples from the target domain to generate additional training data to help bootstrap adaptation in the target domain. We observe that it is easier for large language models to condition on a single variable (context) and compress (Goyal et al., 2022) multiple facts from the passage into a single sentence, as compared to conditioning on a context and answer span together. Moreover, in section 5.1 we observed that augmentation with cloze-style QA pairs yields similar performance to using question-formatted QA pairs, offering evidence that the precise format is not as important as the content itself.

We prompt the model in the following format: `After reading the article, «context» the doctor said «sentence»` for PubMed articles. For other target corpora we replace `doctor` with `engineer`, `journalist`, and `poster` for Stack Overflow, DailyMail, and Reddit respectively. To filter out invalid sentences, we remove any generation that: 1) includes a number, 2) does not repeat part of the passage verbatim, and 3) has less than 75% word set overlap with the passage (after removing stopwords). To gauge the precision of our generations, we sampled 20 generated sentences for each dataset and found that they are correct more than 70% of the time. To test retriever performance, we train a DPR model with source domain data and ~8k examples containing pairs of original passage and generated sentence for each target dataset. We observe performance improvements of ~18% on NewsQA, ~13% on CliCR, and ~24% on Quasar-S (Table 4). Moreover, LLMs contain substantial factual knowledge in their parameters and we observe that they do particularly well in a closed-book setting on datasets with trivia-based factual questions, like SearchQA and Quasar-T, but do not perform well in other cases. Following our conjecture in §3, datasets with *Full shift* on average show an improvement of 12.1% with few-shot interventions, compared to 3.5% with zero-shot, which is also evident in Fig. 1. We show qualitative examples in Appendix (Fig. 8).

6 Related Work

Domain generalization in readers The most popular work in generalization in reading comprehension was introduced as part of the MRQA (Fisch et al., 2019) challenge, which focuses on transfer learning from multiple source datasets to unseen target datasets (Gottumukkala et al., 2020). This multi-task learning setup requires the model to perform complex reasoning at test time that may be unseen at training. However, in this work, we focus on the generalization capabilities of an end-to-end ODQA setup that is able to understand passages in the new domain and not the ability to perform unseen reasoning.

Domain generalization in retrievers A recent line of work that tests domain generalization of retrievers (Petroni et al., 2021; Ram et al., 2022; Izacard et al., 2022) focuses on conservative changes to the source domain, for instance, testing generalization of a model trained on Natural Questions applied to WebQuestions (Berant et al., 2013) or TriviaQA (Joshi et al., 2017), all of which use the same Wikipedia corpus. BEIR is a recent retrieval benchmark, (Thakur et al., 2021) tests the generalizasbility of only the retriever in isolation and not end-to-end ODQA performance, which is a brittle metric.

Domain adaptation work in retrievers (Dai et al., 2022) generate passages using few shots but do not require the answer to be correct. Ma et al. (2021) performs a zero-shot adaptation using noisy labels for retrievers. Siriwardhana et al. (2022) utilizes examples from the target domain in a transfer learning setup while we work in a zero to a few shot setting.

Domain generalization in other tasks Incidental supervision signals in (He et al., 2021) deter-

mine which dataset has a useful signal for a target classification task. Similar to (Fisch et al., 2019), in machine translation, various works (Dua et al., 2022; Fedus et al., 2022) learn to balance positive and negative feature transfer from multiple source domains to a target domain.

7 Conclusion

We investigate failures of ODQA models under non-conservative dataset shift. We also propose a way to test compatibility of source model with new domains without much supervision. We establish how different dataset shift behave under a variety of intervention schemes. We hope future research will adopt these target datasets for evaluation.

8 Limitations

This work focuses on English QA datasets only. Similar techniques should apply in other languages as well; however, we did not evaluate them. The augmentations generated are difficult to validate for yes/no questions for the few-shot method. Moreover, it can be challenging to generate these augmentations if access to large LM is unavailable. However, under those scenarios, data in the target domain should be annotated, which ideally would perform better than the few-shot setting. Our models also suffer from similar problems as LLMs, like hallucinations, misinformation, etc.

9 Ethics Statement

This works focuses on testing the generalization and adaptability of general-purpose models to various domains. It uses existing training data and conventional methods of testing model performance. This work does not deal with any social impacts or biases in natural language processing systems.

Acknowledgements

We would like to thank William Cohen, Haitian Sun, Tom Kwiatkowski and the anonymous reviewers for their feedback. This work was partly supported by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research.

References

Georgios Balikas, Anastasia Krithara, Ioannis Partalas, and George Paliouras. 2015. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *International Workshop on Multimodal Retrieval in the Medical Domain*, pages 26–39. 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, Seattle, Washington, USA. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *ArXiv preprint*, abs/2209.11755.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017. Quasar: Datasets for question answering by search and reading. *ArXiv preprint*, abs/1707.03904.

Dheeru Dua, Shruti Bhosale, Vedanuj Goswami, James Cross, Mike Lewis, and Angela Fan. 2022. Tricks for training sparse translation models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3340–3345, Seattle, United States. Association for Computational Linguistics.

Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Benefits of intermediate annotations in reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5627–5634, Online. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *ArXiv preprint*, abs/1704.05179.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Ananth Gottumukkala, Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Dynamic sampling strategies for multi-task reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 920–924, Online. Association for Computational Linguistics.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2019. An empirical investigation of global and local normalization for recurrent neural sequence models using a continuous relaxation to beam search. 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 1724–1733, Minneapolis, Minnesota. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv preprint*, abs/2209.12356.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hangfeng He, Mingyuan Zhang, Qiang Ning, and Dan Roth. 2021. Foreseeing the benefits of incidental supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1782–1800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *ArXiv preprint*, abs/2112.09118.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *ArXiv preprint*, abs/2208.03299.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel 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, Vancouver, Canada. Association for Computational Linguistics.

Leonid V Kantorovich. 1960. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422.

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, Online. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds. *ArXiv preprint*, abs/1911.02365.

Kalpesh Krishna and Mohit Iyyer. 2019. Generating question-answer hierarchies. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 2321–2334, Florence, Italy. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2022. Learning to retrieve passages without supervision. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2687–2700, Seattle, United States. Association for Computational Linguistics.

Stephen E Robertson and Karen Spärck Jones. 1994. Simple, proven approaches to text retrieval. Technical report, University of Cambridge, Computer Laboratory.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2022. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *ArXiv preprint*, abs/2210.02627.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 569–574, New Orleans, Louisiana. Association for Computational Linguistics.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Yoshua Bengio, and Adam Trischler. 2017. Neural models for key phrase detection and question generation. *ArXiv preprint*, abs/1706.04560.

Simon Šuster and Walter Daelemans. 2018. CliCR: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1551–1563, New Orleans, Louisiana. Association for Computational Linguistics.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv preprint*, abs/2104.08663.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Gerhard Widmer and Miroslav Kubát. 2004. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23:69–101.

Bianca Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM.

A Experimental details

We used JAX on TPUs for training reader models and PyTorch on GPU for training retriever models. We used open-source github implementations for DPR [4], Contriever [5] and Spider [6]. For retrieving top-100 passages for reader input, we used ScaNN[7] library. We use T5-base model for reader and BERT-base for retriever. We fine-tune the retriever and reader with learning rate 1e-5 and 5e-5 respectively.

B How are evaluation sets curated?

We consider validation sets from each of the target dataset, BioASQ, CliCR, Quasar-S, Quasar-T, NewsQA, SearchQA, COLIEE as part of our evaluation set. SearchQA, Quasar-S and Quasar-T were already published as ODQA datasets so we used them as it is while we had re-purpose some of the other datasets that were not originally ODQA dataset by processing them as described below.

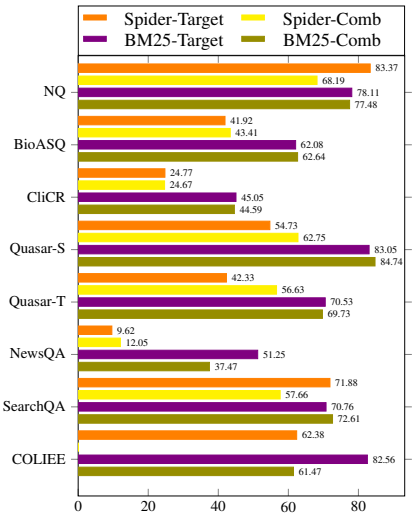
COLIEE: The COLIEE Shared Task 4[8] provides a list of Japanese legal codes in English language. To convert these legal codes from a flat list into paragraphs, first we split them into specific article sections with regex string "Article [0-9]+ ". We further split each article into passages containing a maximum of 256 words.

NewsQA: We created an index on CNN/Dailymail documents by splitting them into passages of 256 words and pooled them together to create a corpus.

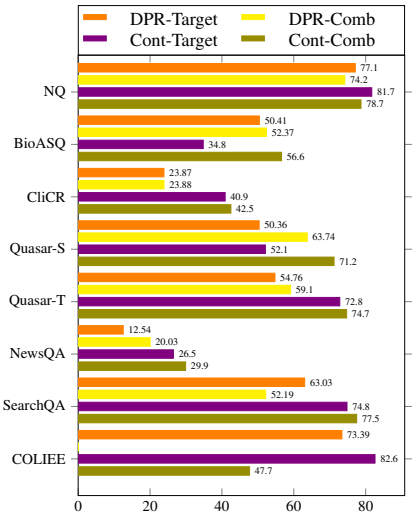
CliCR and BioASQ: We used PubMed corpus published as part of BEIR (Thakur et al., 2021) benchmark. We split the pubmed abstracts in this corpus into passages of size 256 words.

C Varying context distribution

As described in §5.1, we test retriever (Fig. 5) and reader performance (Fig. 6) when exposed to different set of passage. Fig. 6 shows reader performance with passages retrieved with BM25 on source (i.e. wikipedia), target (i.e. respective target corpus) and combined (i.e. all corpora pooled together). Fig. 5 compared performance of Spider and BM25 with Target (i.e. dataset specific target corpus) and Comb (i.e. all corpora pooled together)



(a) Spider-Comb has 0% Acc@k on this COLIEE due to a large numbre of distractors.



(b) DPR-Comb has 0% Acc@k on this dataset due to a large numbre of distractors.

Figure 5: Retriever Performance (Acc@100): Varying context distribution by creating a combined document index. For COLIEE, we use oracle passages for performance computation.

D Varying answer distribution and pre-training corpus

Following §5.1 we try to understand the impact of pre-training and fine-tuning corpus on answer distribution. We do this by comparing the performance of the FiD reader initialized from T5 pre-trained on common-crawl dataset(C4) compared to one that was pre-trained on PubMed articles (Table 5). After pre-training, both models are then fine-tuned on our source domain data. In this case,

[4]https://github.com/facebookresearch/DPR.git
[5]https://github.com/facebookresearch/contriever.git
[6]https://github.com/oriram/spider
[7]https://ai.googleblog.com/2020/07/announcing-scann-efficient-vector.html
[8]https://sites.ualberta.ca/ rabelo/COLIEE2022/

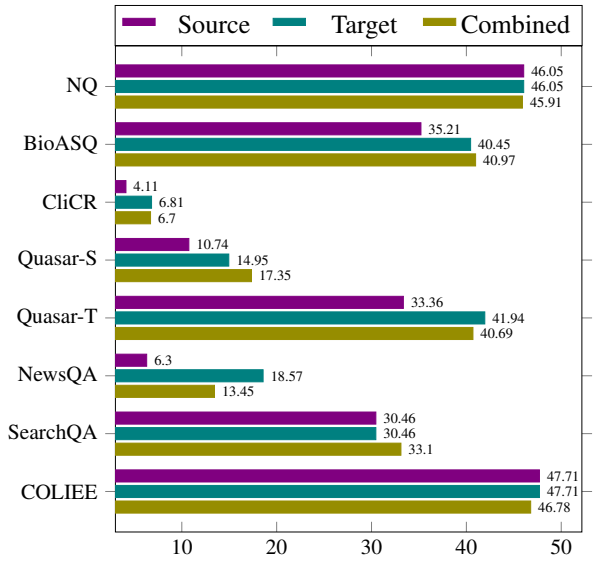


Figure 6: Reader Performance (F1): Effect of change in context distribution with BM25 retrievals from the combined index.

we observe that fine-tuning on a domain that differs from that used in pre-training results in deterioration of model performance.

Augmentations	C4	Pubmed
Random	33.50	33.51
Uniform	**39.07**	35.97
Most frequent	**38.18**	34.90
BioASQ train answers	**41.33**	36.71

Table 5: Answer distribution: Reader performance on BioASQ with C4 and Pubmed pre-trained T5

E Degree of domain shift

In Table 1, we showed only differences that governed which side of decision tree the shift types were categorized into, while in Table 6 we show all the raw distance values.

F Statistical Significance

The number of examples in all datasets except COLIEE are in the order of thousands, making the performance improvements significant. In the case of COLIEE, which has a boolean output space (i.e. answers are yes/no), we performed a binomial test to test the significance of few-shot reader performance in Table 4. The number of samples n = 116 (number of test examples), p_0=0.468 and p_t=0.616. We will reject the null hypothesis that baseline and few-shot distribution are equivalent, when $P(X >= p_t * n) <= 0.05$, where X is drawn

Dataset	Retriever		Reader	
	w_u^t	w_g^t	v_u^t	$v_u^r - v_u^t$
BioASQ	0.6477	0.3450	0.1160	0.1765
CliCR	0.0602	0.9448	0.0573	0.2352
Quasar-S	0.1658	0.8355	0.2158	0.0767
Quasar-T	0.5978	0.3962	0.1231	0.1694
NewsQA	0.3992	0.5959	0.1125	0.1800
SearchQA	0.80350	0.1870	0.2988	-0.0063

Table 6: Wasserstein distance computed over 100 target domain examples. The distance between reference (source) and uniform over 100 validation set source domain examples is v_u^r=0.2925

from a binomial distribution, i.e., $X \sim B(n, p_0)$ (Berg-Kirkpatrick et al., 2012) and we can compute the L.H.S to be, $P(X >= 0.616 * 116) = 0.00006$ making it significant.

Dataset, Corpus	#ques, #docs	Passage	Question-Answer
BioASQ, Pubmed	5k, 30M	Parkinson's disease (PD) is one of the most common degenerative disorders of the central nervous system that produces motor and non-motor symptoms. The majority of cases are idiopathic and characterized by the presence of Lewy bodies.	Q: Which disease of the central nervous system is characterized by the presence of Lewy bodies? A: Parkinson's disease
CliCR, Pubmed	90k, 30M	Detailed history and examination ruled out the above causes except the exposure to high altitude as a cause for koilonychia in our patient. Exposure to high altitude is a known aetiology for koilonychias, also described by some authors as "Ladakhi koilonychia".	Q: __ is a known cause of koilonychia, described by some as Ladakhi koilonychia. A: High altitude exposure
Quasar-S, Stackoverflow	30k, 1.5M	I have a mixed integer quadratic program MIQP which I would like to solve using SCIP. The program is in the form such that on fixing the integer variables the problem turns out to be a linear program.	Q: scip – an software package for solving mixed integer __ problems A: linear-programming
Quasar-T, Reddit	30k, 2M	Because of widespread immunization , tetanus is now rare. Another name for tetanus is lockjaw.	Q: Lockjaw is another name for which disease A: tetanus
NewsQA, Dailymail	70k, 0.5M	Former boxing champion Vernon Forrest, 38, was shot and killed in southwest Atlanta, Georgia, on July 25.	Q: Where was Forrest killed ? A: in southwest Atlanta , Georgia
SearchQA, WIkipedia	70k, 20M	The Dangerous Summer and The Garden of Eden. Written in 1959 while Hemingway was in Spain on commission for Life...	Q: While he was in Spain in 1959, he wrote "The Dangerous Summer", a story about rival bullfighters A: Hemingway
COLIEE, Japanese Legal Codes	886, 1k	A manifestation of intention based on fraud or duress is voidable. If a third party commits a fraud inducing a first party to make a manifestation of intention to a second party, that manifestation of intention is voidable only if the second party knew or could have known that fact. The rescission of a manifestation of intention induced by fraud under the provisions of the preceding two paragraphs may not be duly asserted against a third party in good faith acting without negligence.	Q: Is it true: A person who made a manifestation of intention which was induced by duress emanated from a third party may rescind such manifestation of intention on the basis of duress, only if the other party knew or was negligent of such fact. A: No

Figure 7: Examples from datasets with context and question-answer pairs from different domains.

Dataset, Corpus	Passage	Generated Sentence
BioASQ, Pubmed	Herceptin is widely used in treating Her2-overexpressing breast cancer. However, the application of Herceptin in prostate cancer is still controversial.... This implies that targeting Her2 by both radio- and immunotherapy might be a potential strategy for treating patients with androgen-independent prostate cancer...	Herceptin is a breast cancer drug that has been used in treating prostate cancer.
CliCR, Pubmed	An infant was admitted with symptoms of diarrhoea and vomiting. After initial improvement she unexpectedly died. Postmortem confirmed a diagnosis of cytomegalovirus (CMV) enterocolitis. The authors report this case and review other published cases of immunocompetent infants who presented with this infection. Clinicians should consider stool CMV PCR test or referral for endoscopy and biopsy in young babies who present with profuse and prolonged episodes of diarrhoea.	Immunocompetent infants can present with CMV enterocolitis.
Quasar-S, Stackoverflow	I've recently found scala-bindgen from a Gitter room on Scala Native. Seems like at the present point in time they are developing a tool for generating Scala bindings for C header-files. Are there plans for generating Scala bindings for Objective-C and C++ too...	scala-bindgen is a tool that generates scala bindings for C header files.
Quasar-T, Reddit	Interview With Gary James' Interview With Marshall Lytle of Bill Haley's Comets It can be safely said that "Rock Around The Clock" was the song by the group Bill Haley And His Comets that started the Rock 'n Roll movement. Still performing today, he spoke about those early days of Rock 'n Roll and his appreciation for what it meant to him.	Bill Haley and his comets made rock and roll music
NewsQA, CNN/ Dailymail	The Kardashians are already a staple on E! Network . But they've chosen the month of November to assert their dominance on the book world. Kourtney, Kim, and Khloe's first novel," Dollhouse ," hits shelves today . "Dollhouse," the first fiction endeavor from the Kardashians, follows sisters Kamille, Kassidy, ...	The Kardashians released a new book called 'Dollhouse'.
SearchQA, Wikipedia	Charles Henry Dow was an American journalist who co-founded Dow Jones and Company with Edward Jones and Charles Bergstresser. Dow also founded The Wall Street Journal, which has become one of the most respected financial publications in the world... In 1877, he published a History of Steam Navigation between New York and...	Charles Henry Dow, an American journalist, founded The Wall Street Journal in 1882.

Figure 8: Examples of data generated from few-shot prompting.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☒ A2. Did you discuss any potential risks of your work?
This is a study of existing work, we do not use a lot of compute resources for running these experiments so this point is not applicable.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1(introduction)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

Section 4 and 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 and 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4, 5 and Appendix

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.