The Effect of Natural Distribution Shift on Question Answering Models

John Miller [1] **Karl Krauth** [1] **Benjamin Recht** [1] **Ludwig Schmidt** [1]

Abstract

We build four new test sets for the Stanford Question Answering Dataset (SQuAD) and evaluate the ability of question-answering systems to generalize to new data. Our first test set is from the original Wikipedia domain and measures the extent to which existing systems overfit the original test set. Despite several years of heavy test set re-use, we find no evidence of adaptive overfitting. The remaining three test sets are constructed from New York Times articles, Reddit posts, and Amazon product reviews and measure robustness to natural distribution shifts. Across a broad range of models, we observe average performance drops of 3.8, 14.0, and 17.4 F1 points, respectively. In contrast, a strong human baseline matches or exceeds the performance of SQuAD models on the original domain and exhibits little to no drop in new domains. Taken together, our results confirm the surprising resilience of the holdout method and emphasize the need to move towards evaluation metrics that incorporate robustness to natural distribution shifts.

1. Introduction

Since its release in 2016, the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) has generated intense interest from the natural language processing community. At first glance, this intense interest has lead to impressive results. The best performing models in 2020 (Devlin et al., 2019; Yang et al., 2019) have F1 scores more than 40 points higher than the baseline presented by Rajpurkar et al. (2016). At the same time, it remains unclear to what extent progress on these benchmark numbers is a reliable indicator of progress more broadly.

The goal of building question answering systems is not merely to obtain high scores on the SQuAD leaderboard, but rather to *generalize* to new examples beyond the SQuAD test set. However, the competition format of SQuAD puts pressure on the validity of leaderboard scores. It is well-known that repeatedly evaluating models on a held-out test set can give overly optimistic estimates of model performance, a phenomenon known as *adaptive overfitting* (Dwork et al., 2015). Moreover, the standard SQuAD evaluation only measures model performance on new examples *from a single distribution*, i.e., paragraphs derived from Wikipedia articles. Nevertheless, we often use models in settings different from the one in which they were trained. While Jia & Liang (2017) demonstrated that SQuAD models are not robust to *adversarial* distribution shifts, one might still hope that the models are more robust to *natural* distribution shifts, for instance changing from Wikipedia to newspaper articles.

This state of affairs raises two important questions:

Are SQuAD models overfit to the SQuAD test set?
Are SQuAD models robust to natural distribution shifts?

In this work, we address both questions by replicating the SQuAD dataset creation process and generating four new SQuAD test sets on both the original Wikipedia domain, as well as three new domains: New York Times articles, Reddit posts, and Amazon product reviews.

We first show that there is no evidence of adaptive overfitting on SQuAD. Across a large collection of SQuAD models, there is little to no difference between the F1 scores from the original SQuAD test set and our replication. This even holds when comparing scores from the SQuAD *development* set (which was publicly released with answers) to our new test set. The lack of adaptive overfitting is consistent with recent replication studies in the context of image classification (Recht et al., 2019; Yadav & Bottou, 2019). These studies leave open the possibility that this phenomenon is specific to the data or models typical in computer vision research. Our result demonstrates this same phenomenon also holds for natural language processing.

Beyond adaptive overfitting, we also demonstrate that SQuAD models exhibit robustness to some of our natural distribution shifts, though they still suffer substantial performance degradation on others. On the New York Times dataset, models in our testbed on average drop 3.8 F1 points.

[1]Department of Computer Science, University of California, Berkeley, Berkeley, California, USA. Correspondence to: John Miller <miller_john@berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Figure 1. Model and human F1 scores on the original SQuAD v1.1 test set compared to our new test sets. Each point corresponds to a model evaluation, shown with 95% Student's t-confidence intervals (mostly covered by the point markers). The plots reveal three main phenomena: (i) There is no evidence of adaptive overfitting on SQuAD, (ii) all of the models suffer F1 drops on the new datasets, with the magnitude of the drop strongly depending on the corpus, and (iii) humans are substantially more robust to natural distribution shifts than the models. The slopes of the linear fits are 0.92, 1.02, 1.19, and 1.36, respectively, and the R^2 statistics for the linear fits are 0.99, 0.97, 0.9, and 0.89, respectively. This means that every point of F1 improvement on the original dataset translates into roughly 1 point of improvement on our new datasets.

On the Reddit and Amazon datasets, the drop is on average 14.0 and 17.4 F1 points, respectively. All of our datasets were collected using the same data generation pipeline, so this degradation can be attributed purely to changes in the source text rather than differences in the annotation procedures across datasets.

We complement each of these experiments with a strong human baseline comprised of the authors of this paper. On the original SQuAD data, our human accuracy numbers are on par with the best SQuAD models (Yang et al., 2019) and significantly better than the Mechanical Turk baseline reported by Rajpurkar et al. (2016). On our new test sets, average human F1 scores decrease by 0.1 F1 on New York Times, 2.9 on Reddit, and 3.0 on Amazon. All of the resulting F1 scores are substantially higher than the best SQuAD models on the respective test sets.

Figure 1 summarizes the main results of our experiments. Humans show consistent behavior on all four test sets, while models are substantially less robust against two of the distribution shifts. Although there has been steady progress on the SQuAD leaderboard, there has been markedly less progress in this robustness dimension.

To enable future research, all of our new tests sets are freely available online.[1]

[1]https://modestyachts.github.io/squadshifts-website/

2. Background

In this section, we briefly introduce the SQuAD dataset and present a formal model for reasoning about performance drops between our test sets.

2.1. Stanford Question Answering Dataset

SQuAD is an extractive question answering dataset introduced by Rajpurkar et al. (2016). An example in SQuAD consists of a passage of text, a question, and one or more spans of text within the passage that answer the question. An example is given in Figure 2.

Model performance is evaluated using one of two metrics: exact match (EM) or F1. Exact match measures the percentage of predictions that exactly match at least one of the ground truth answers. F1 measures the maximum overlap between the tokens in the predicted span and any of the ground truth answers, treating both the prediction and each answer as a bag of words. Both metrics are described formally in Appendix A.

After releasing the SQuAD v1.1 dataset, Rajpurkar et al. (2018) introduced a new variant of the dataset, SQuAD 2.0, that includes unanswerable questions. Since SQuAD v1.1 has been public longer and potentially subject to more adaptivity, we focus on SQuAD v1.1 and refer to it as the SQuAD dataset. The SQuAD test set is not publically available. Therefore, while we use public test set evaluation numbers, we use the public development set for analysis.

Passage: "In our neighborhood, we were the small family, at least among the Irish and Italians… We could almost field a full baseball team. But the Flynns, they could put an entire football lineup… We loved Robert F. Kennedy's family: 11 kids, and Ethel looks great. Bobby himself was the seventh of nine."

Question: How many kids did Robert F. Kennedy have?
Answer: 11

Question: The author believes his family could fill a team of which sport?
Answer: baseball

Figure 2. Question and answer pairs from a sample passage in our New York Times SQuAD test set. Answers are text spans from the passage that answer the question.

2.2. A Model for Generalization

Although progress on SQuAD is measured through performance on a held-out test set, the implicit goal is not to achieve high F1 scores on the test set, but rather to *generalize* to unseen examples. Our experiments test the extent to which this assumption holds—if models with high leaderboard scores on the test set continue to perform well on new examples, whether from the same or different distributions.

To be more formal, suppose the original test set S is sampled from some underlying distribution $\mathcal{D}$, and consider a model f submitted to the SQuAD leaderboard. Let $L_S(f)$ denote the empirical loss of model f on the sample S, and let $L_{\mathcal{D}}(f)$ denote the corresponding population loss. In our experiment, we gather a new dataset of examples S' from a distribution $\mathcal{D}'$, potentially different from $\mathcal{D}$. We wish for the loss on the new sample, $L_{S'}(f)$ to be close to the original, $L_S(f)$. Omitting f, we can decompose this gap into three terms (Recht et al., 2019).

$$L_S - L_{S'} = \underbrace{(L_S - L_{\mathcal{D}})}_{\text{Adaptivity gap}} + \underbrace{(L_{\mathcal{D}} - L_{\mathcal{D}'})}_{\text{Distribution gap}} + \underbrace{(L_{\mathcal{D}'} - L_{S'})}_{\text{Generalization gap}}$$

The *adaptivity gap* $L_S - L_{\mathcal{D}}$ measures how much adapting the model to the held-out test set S biases the estimate of the population loss. Since recent models are in part chosen on the basis of past test set information, the model f is not independent of S. Hence $L_S(f)$ can underestimate $L_{\mathcal{D}}(f)$, a phenomenon called *adaptive overfitting*. The *distribution gap* measures how much changing the distribution from $\mathcal{D}$ to $\mathcal{D}'$ affects the model's performance. Finally, the *generalization gap* $L_{S'} - L_{\mathcal{D}'}$ captures the difference between the sample and the population losses due to random sampling of S'. Since S' is sampled independently of the model f, this gap is typically small and well-controlled by standard concentration results. For example, on the new Wikipedia test set, the average size of Student's t-confidence intervals for models in our testbed is ± 0.6 F1.

In the sequel, we empirically measure both the adaptivity gap and the distribution gap for a wide range of SQuAD models by collecting new test sets from a variety of distributions $\mathcal{D}'$. We first review related work that motivates our choice of SQuAD and natural distribution shifts.

3. Related Work

Adaptive data analysis. Although repeated test-set reuse puts pressure on the statistical guarantees of the holdout method (Dwork et al., 2015), a series of replication studies established there is no adaptive overfitting on popular classification benchmarks like MNIST (Yadav & Bottou, 2019), CIFAR-10 (Recht et al., 2019), and ImageNet (Recht et al., 2019). Furthermore, Roelofs et al. (2019) also found little to no evidence of adaptive overfitting in a host of classification competitions on the Kaggle platform. These investigations either concern image classification or smaller competitions that have not been subject to intense, multi-year community scrutiny. Our work establishes similar results for natural language processing on a heavily studied benchmark.

A number of works have proffered explanations for why adaptive overfitting does not occur in machine learning (Blum & Hardt, 2015; Mania et al., 2019; Feldman et al., 2019; Zrnic & Hardt, 2019). Complementary to these results, our work provides a new data point with which to validate and deepen our understanding of overfitting.

Datasets for question answering. Beyond SQuAD, a number of works have proposed datasets for question answering (Richardson et al., 2013; Berant et al., 2014; Joshi et al., 2017; Trischler et al., 2017; Dunn et al., 2017; Yang et al., 2018; Kwiatkowski et al., 2019). We focus our analysis on SQuAD for two reasons. First, SQuAD has been the focus of intense research for almost four years, and the competitive nature of the leaderboard format makes it an excellent example to study adaptive overfitting in natural language processing. Second, SQuAD requires all submissions to be uploaded to CodaLab[2], which ensures reproducibility and makes it possible to evaluate every submission on our new datasets using the same configuration and environment as the original evaluation.

Generalization in question answering. Given the plethora of question-answering datasets, Yogatama et al. (2019), Talmor & Berant (2019), and Sen & Saffari (2020)

[2] https://worksheets.codalab.org/

evaluate the extent to which models trained on SQuAD generalize to other question-answering datasets. Hendrycks et al. (2020) evaluates robustness to distribution shift for question answering, among other tasks, by carefully splitting subsets of the ReCoRD dataset (Zhang et al., 2018). In a similar vein, Fisch et al. (2019) conduct a shared task competition that evaluates how well models trained on multiple datasets generalize to unseen datasets at test time. In these cases, the datasets encountered at test time vary across a number of dimensions: the question collection procedure, the origin of the input text, the question answering interface, the crowd worker population, etc. These differences are *confounding factors* that make it difficult to interpret performance differences across datasets. For example, human performance differs by 10 F1 points between SQuAD v1.1 and NewsQA (Trischler et al., 2017). In contrast, our datasets focus on a single factor of variation—the input text corpus. In this controlled setting, we observe non-trivial F1 drops across a large collection of models, while human F1 scores are essentially constant.

From a different perspective, Jia & Liang (2017) and Ribeiro et al. (2018) consider robustness to *adversarial* dataset corruptions. Kaushik et al. (2019) and Gardner et al. (2020) evaluate model performance when individual examples are perturbed in small, but semantically meaningful ways. While we instead focus on *naturally occurring* distribution shifts, we also evaluate our model testbed on adversarial distribution shifts in Appendix B.

4. Collecting New Test Sets

In this section, we describe our data collection methodology. Data collection primarily proceeds in two stages: curating passages from a text corpus and crowdsourcing question-answer pairs over the passages. In both of these stages, we take great care to replicate the original SQuAD data generation process. Where possible, we obtained and used the original SQuAD generation code kindly provided by Rajpurkar et al. (2016). We ran our dataset creation pipeline on four different corpora: Wikipedia articles, New York Times articles, Reddit posts, and Amazon product reviews. Table 1 summarizes the statistics of our new datasets.

4.1. Passage Curation

The first step in the dataset generation process is selecting the articles from which the passages or contexts are drawn.

Wikipedia. We sampled 48 articles uniformly at random from the same list of 10,000 Wikipedia articles as Rajpurkar et al. (2016), ensuring there is no overlap between our articles and those in the SQuAD v1.1 training or development sets. To minimize distribution shift due to temporal language variation, we extracted the text of the Wikipedia articles from around the publication date of the SQuAD v1.0 dataset (June 16, 2016). For each article, we extracted individual paragraphs and stripped out images, figures, and tables using the same data processing code as Rajpurkar et al. (2016). Then, we subsampled the resulting paragraphs to match the passage length statistics of the original SQuAD dataset.[3] See Appendix D.1 for a detailed comparison of the paragraph distribution of the original SQuAD dev set and our new SQuAD test set.

Table 1. Dataset statistics of our four new test sets compared to the original SQuAD 1.1 development and test sets.

Dataset	Total Articles	Total Examples
SQuAD v1.1 Dev	48	10,570
SQuAD v1.1 Test	46	9,533
New Wikipedia	48	7,938
New York Times	797	10,065
Reddit	1969	9,803
Amazon	1909	9,885

New York Times. We sampled New York Times articles from the set of all articles published in 2015 using the NYTimes Archive API. We scraped each article with the Wayback Machine[4], using the same snapshot timestamp as our Wikipedia dataset and removed foreign language articles. Since the average paragraph length for NYT articles is significantly shorter than the average paragraph length for Wikipedia articles, we randomly merged each NYT paragraph with its successor, and then we subsampled the merged paragraphs to match the passage length statistics of the original SQuAD dataset.

Reddit Posts. We sampled Reddit posts from the set of all posts across all subreddits during the month of January 2016 in the Pushshift Reddit Corpus (Baumgartner et al., 2020). We restricted the set of posts to those marked as "safe for work" and manually inspected and removed inappropriate posts. We concatenated each post's title with its body, removed Markdown, and replaced all links with a single token, `LINKREMOVED`. We then subsampled the posts to match the passage length statistics of the original SQuAD dataset.

Amazon Product Reviews. We sampled Amazon product reviews belonging to the "Home and Kitchen" category from the dataset released by McAuley et al. (2015). As in the previous datasets, we then subsampled the reviews to match the passage length statistics of SQuAD.

[3] The minimum 500 character per paragraph rule mentioned in Rajpurkar et al. (2016) was adopted midway through their data collection, and hence the original dataset also includes shorter paragraphs (Rajpurkar, 2019).

[4] https://archive.org/web/

Table 2. Comparison of model F1 scores on the original SQuAD test set and our new Wikipedia test set. Rank refers to the relative ordering of the models in our testbed using the original SQuAD v1.1 F1 scores, new rank refers to the ordering using the new Wikipedia test set scores, and Δ rank is the relative difference in ranking from the original test set to the new test set. The confidence intervals are 95% Student's t-intervals. No confidence intervals are provided for the SQuAD v1.1 dataset since the dataset is not public and only the average scores are available. A complete table with data for the entire model testbed and analogous data for EM scores is in Appendix E.

New-Wiki F1 Score Summary

Rank	Name	SQuAD	New-Wiki	Gap	New Rank	Δ Rank
-	Human Average (this study)	95.1	92.4	2.7	-	-
1	XLNet	95.1	92.3 [91.9, 92.8]	2.7	1	0
8	BERT-Large Baseline	92.7	90.8 [90.3, 91.3]	1.9	9	-1
42	BiDAF+SelfAttention+ELMo	85.9	83.8 [83.1, 84.5]	2.1	45	-3
83	RaSoR	78.7	77.2 [76.4, 78.1]	1.5	84	-1
85	AllenNLP BiDAF	77.2	76.5 [75.7, 77.3]	0.7	88	-3

4.2. Crowdsourcing Question-Answer Pairs

We employed crowdworkers on Amazon Mechanical Turk (MTurk) to ask and answer questions on the passages in each dataset. We followed a nearly identical protocol to the original SQuAD dataset creation process. We used the same MTurk user interface, task instructions, MTurk worker qualifications, time per task, and hourly rate (adjusted for inflation) as Rajpurkar et al. (2016). For full details and examples of the user interface, refer to Appendix D.2.

For each paragraph, one crowdworker first asked and answered up to five questions on the content of the paragraph. Then we obtained at least two additional answers for each question using separate crowdworkers. There are two points of discrepancy between our crowdsourcing protocol and the one used to create the original SQuAD dataset. First, we interfaced directly with MTurk rather than via the Daemo platform because the Daemo platform has been discontinued. Second, in our MTurk tasks, workers asked and answered questions for at most five paragraphs rather than for the entire article because MTurk workers preferred smaller units of work. Although each difference is a potential source of distribution shift, in Section 5 we show that the effect of these changes is negligible—models achieve roughly the same scores on both the original and new Wikipedia datasets. On average, the difference in F1 scores is 1.5 F1, and 95% of models in our testbed are within 2.7 F1.

After gathering question and answer pairs for each paragraph, we apply the same post-processing and data cleaning as SQuAD v1.1. We adjusted answer whitespace for consistency, filtered malformed answers, and removed all documents that had less than an average of two questions per paragraph after filtering. In Appendix C.7, we show that further manual filtering of incorrect, ungrammatical, or otherwise malformed questions and answers has negligible impact on our results.

4.3. Human Evaluation

Although both SQuAD and our new test sets have answers from MTurk workers, it is not clear whether these answers represent a compelling human baseline. At minimum, workers are not familiar with the typical style of answers in SQuAD (e.g., how much detail to include), and they receive no feedback on their performance. To obtain a stronger human baseline, the graduate student and postdoc authors of this paper also answered approximately 1,000 questions on each of the four new test sets and the original SQuAD development set, following the same procedure and using the same UI as the MTurk workers. To take feedback into account, each participant first labelled 500 practice examples from the training set and compared their answers with the ground truth.

5. Main Results

We use the four new datasets generated in the previous part to test for adaptive overfitting on SQuAD and probe the robustness of SQuAD models to natural distribution shifts.

We evaluated a broad set of over 100 models submitted to the SQuAD leaderboard, including state-of-the-art models like XLNet (Yang et al., 2019) and BERT (Devlin et al., 2019), as well as older, but popular models like BiDAF (Seo et al., 2016). All of the models were submitted to the CodaLab platform, and we evaluate every model using the exact same configuration (model weights, hyperparameters, command-line arguments, execution environment) as the original submission. Tables 2 and 3 contain a brief summary of the results for key models on the new Wikipedia and Amazon datasets. Detailed results table and citations for the models, where available, are given in Appendix E.

5.1. Adaptive Overfitting

The SQuAD models in our testbed come from a long sequence of papers that incrementally improve F1 and EM scores over a period of several years. Consequently, if there is adaptive overfitting, we should expect the later models to have larger drops in F1 scores because they are the result of more interaction with the test set. In this case, the higher F1 scores are partially the result of a larger adaptivity gap, and we would expect that, as the observed scores L_S continue to rise, the population scores L_D would begin to plateau.

To check for adaptive overfitting on the existing test set, we plot the SQuAD v1.1 test F1 scores against F1 scores on our new Wikipedia test set. Figure 1 in Section 1 provides strong evidence against the adaptive overfitting hypothesis. Across the entire model collection, the F1 scores on the new test set closely replicate the original F1 scores. The observed linear fit is in contrast to the concave curve one would expect from adaptive overfitting. We use 95% Student's t-confidence intervals, which make a large-sample Gaussian assumption, to capture the error in the new F1 scores due to random variation. No such confidence intervals are available for the original test set scores since the test set is not publicly available. A similar plot for EM scores is provided in Appendix C.1.

Not only is there little evidence for adaptive overfitting on the test set, there is also little evidence of adaptive overfitting on the SQuAD development set. In Figure 3, we plot F1 scores on the SQuAD v1.1 development set against F1 scores on the SQuAD v1.1 test set. With the exception of three models, the F1 scores on the dev set closely match the scores on the test set, despite the fact that the development set is aggressively used during model selection. Moreover, the models that do not lie on the linear trend line—`Common-sense Governed BERT-123 (April 21)`, `Common-sense Governed BERT-123 (May 9)`, and `XLNet-123++`—are directly trained on the development set (Qiu, 2020).

5.2. Robustness to Natural Distribution Shifts

Given the correspondence between the old and new Wikipedia test set F1 scores, the adaptivity gap and the distribution gap are small or non-existent. Consequently, the distribution shift stemming from our data generation pipeline affects the models only minimally. This allows us to probe the sensitivity of the SQuAD models to a set of controlled distribution shifts, namely the choice of text corpus. Since all of the datasets are constructed with the same preprocessing pipeline, crowd-worker population, and post-processing, the datasets are free of confounding factors that would otherwise arise when comparing model performance across different datasets.

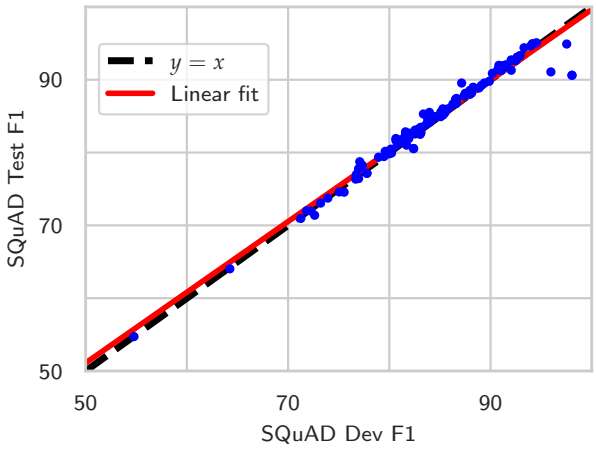

Figure 3. Comparison of F1 scores between the SQuAD v1.1 dev set and the SQuAD v1.1 test set. Despite heavy use of the dev set during model development, the dev set and test set scores closely match, with the exception of three models that were explicitly trained on the dev set, `Common-sense Governed BERT-123 (April 21)`, `Common-sense Governed BERT-123 (May 9)`, and `XLNet-123++`. (Qiu, 2020). The slope of the linear fit is 0.97.

Figure 1 in Section 1 shows F1 scores on the SQuAD v1.1 test set versus the F1 scores on each of our new test sets for all the models in our testbed. All models experience an F1 drop on the new test sets, though the magnitude strongly depends on the specific test set. On New York Times, for instance, BERT only drops around 2.1 F1 points, whereas it drops around 11.9 F1 points on Amazon and 11.5 F1 points on Reddit. The top performing XLNet model (Yang et al., 2019) is a clear outlier. Despite generalizing well to the new Wikipedia dataset, XLNet drops nearly 10 F1 and 40 EM points on New York Times, substantially more than models with similar performance on SQuAD v1.1 as well as other XLNet variants, e.g., XLNet-123[5].

In general, F1 scores on the original SQuAD test set are highly predictive of F1 scores on the new test sets. Interestingly, the relationship is well-captured by a linear fit even under distribution shifts. Similar to Recht et al. (2019), in Figure 11, we observe the linear fits are better under a probit scaling of F1 scores. See Appendix C.2 for more details. Moreover, the gap between perfect robustness ($y = x$) and the observed linear fits varies with the dataset: 3.8 F1 points for New York Times, 14.0 points for Reddit, and 17.4 F1 for Amazon. In each case, however, higher performance on SQuAD v1.1 translates into higher performance on these

[5]This large drop persists even when normalizing Unicode characers and replacing Unicode punctuation with Ascii approximations.

Table 3. Comparison of model F1 scores on the original SQuAD test set and our new Amazon test set. Rank refers to the relative ordering of the models in our testbed using the original SQuAD v1.1 F1 scores, new rank refers to the ordering using the Amazon test set scores, and Δ rank is the relative difference in ranking from the original test set to the new test set. The confidence intervals are 95% Student's t-intervals. See Appendix E for EM scores and evaluation results for the entire model testbed.

Amazon F1 Score Summary

Rank	Name	SQuAD	Amazon	Gap	New Rank	Δ Rank
-	Human Average (this study)	95.1	92.1	3.0	-	-
1	XLNet	95.1	81.7 [81.1, 82.2]	13.4	5	-4
8	BERT-Large Baseline	92.7	80.8 [80.2, 81.5]	11.9	8	0
45	BiDAF+SelfAttention+ELMo	85.9	69.2 [68.3, 70.0]	16.7	43	2
90	RaSoR	78.7	57.6 [56.8, 58.5]	21.1	91	-1
93	AllenNLP BiDAF	77.2	56.2 [55.3, 57.0]	21.0	95	-2

natural distribution shift instances.

Despite the robustness demonstrated by the models, on all of the test sets with distribution shift, human performance is substantially higher than model performance and well above the linear fits shown in Figure 1 and Figure 11. This rules out the possibility that the shift in F1 scores are entirely by a change in the Bayes error rate. Moreover, it points towards substantial room for improvement for models on our new test sets.

6. Further Analysis

In this section, we further explore the properties of our new test sets. We first study the extent to which common measures of dataset difficulty can explain the performance drops on our new test sets. Then, we evaluate whether training models with more data or more diverse data improves robustness to our distribution shifts.

6.1. Are The New Test Sets Harder Than The Original?

One hypothesis for the performance drops observed in Section 5.2 is that our new dataset are harder in some sense. For instance, the diversity of answers may be greater among Reddit comments than Wikipedia articles. To better understand this question, we compare the original SQuAD development set to our four new test sets using the three difficulty measures introduced in Rajpurkar et al. (2016).

Answer diversity. Following Rajpurkar et al. (2016), we automatically categorize each answer into numerical and non-numerical answers, named entities, and constituents using spaCy (Honnibal & Montani, 2017) and the constituency parser from Kitaev & Klein (2018). Histograms of answer types for each data are shown in Figure 4. Since the original pipeline is not available, our implementation differs slightly from Rajpurkar et al. (2016) and we include results on the SQuAD v1.1 development set for comparison. Both the original and our new Wikipedia test set have very similar answer type histograms. The distribution shift datasets have slight variations in the answer distributions. For instance, NYT has more person answers, whereas Amazon has more adjective phrases. However, changes in the answer type distribution between datasets are not sufficient to explain

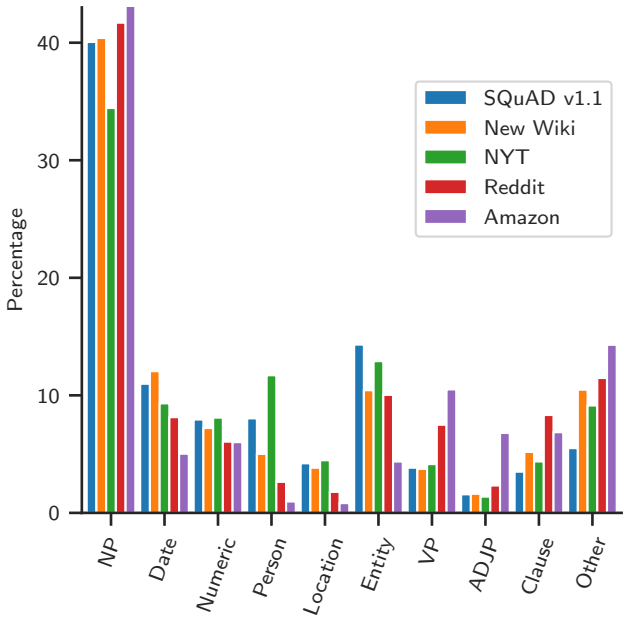

Figure 4. Comparison of answers types in the original and new datasets. We automatically partition our answers into the same categories as Rajpurkar et al. (2016). Although there are differences between the datasets, e.g., New York Times has more person answers, the four datasets are very similar. Moreover, we show in Appendix C.4 that differences in answer categorization across datasets do not explain the performance drops we observe.

the performance differences between the datasets. In Appendix C.4, we consider a simple model that predicts F1 scores on our new test sets by stratifying the dataset by answer type, computing model F1 scores for each type, and then reweighing these scores by the relative frequency of each answer type in our new test set. This model explains only a small fraction of the performance differences across test sets.

Syntactic divergence. We also stratify our datasets using the automatic syntactic divergence measure of Rajpurkar et al. (2016). Syntactic divergence measures the similarity between the syntactic dependency tree structure of both the question and answer sentences and provides another metric of example difficulty. In Figure 5, we compare the histograms of syntactic divergence for the SQuAD v1.1 development set and our new test sets. All of the datasets have similar histograms, though both the Reddit and Amazon test sets have slightly more examples with small syntactic divergence. As in the previous part, in Appendix C.5, we consider a simple model that predicts F1 scores on the new test sets by stratifying the dataset according to syntactic divergence and reweighting based on the relative frequence of examples with a given syntactic divergence measure. As before, this model explains only a small fraction of the performance differences across test sets.

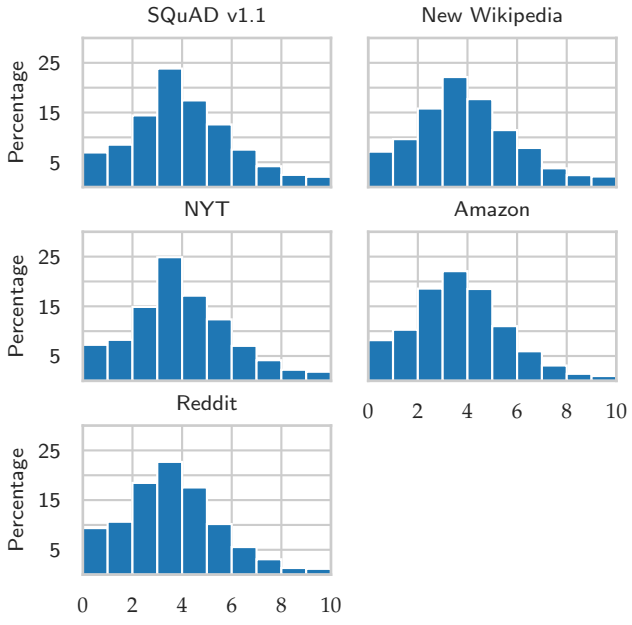

Figure 5. Histograms of syntactic divergence between question and answer sentences for both the original and new datasets. All of the datasets have a similar distribution of syntactic divergence, though the Reddit and Amazon datasets have more question-answers pairs with small (1-2) syntactic divergence.

Reasoning required. Finally, we compare our new test sets in terms of the reasoning required to answer each question-answer pair, using the same non-mutually exclusive categories as Rajpurkar et al. (2016). For each test set, as well as the SQuAD development set, we randomly sampled and manually labeled 192 examples. The results for each dataset are presented in Table 4. Both the Amazon and Reddit dataset have more examples requiring world knowledge to resolve lexical variation, while the New York Times dataset has more examples requiring multi-sentence reasoning. Differences in reasoning required between test sets do not explain the observed performance drops. In Appendix C.6, we present a another model that predicts F1 scores on our new test sets by computing model F1 scores in each reasoning category and then reweighing these scores based on the relative frequency of each category on new test sets. This model explains virtually none of the observed changes in F1 scores.

6.2. Are Models Trained with More Data More Robust to Natural Distribution Shifts?

High performance on our new datasets requires models to generalize to data distributions that may be different from those on which they were trained. Our primary evaluation only concerns the robustness of SQuAD models, and a natural follow-up question is whether models trained on more data, or explicitly trained for out-of-distribution question-answering, perform better on our new test sets.

To test this claim, we evaluated a collection of models from the Machine Reading for Question Answering (MRQA) 2019 Shared Task on Generalization (Fisch et al., 2019). In the shared task, models were trained on 6 question-answering datasets, including SQuAD v1.1, and then evaluated on 12 held-out datasets. The datasets simultaneously differed not just in the passage distribution, as in our experiments, but also in confounders like the data collection procedure, the question distribution, and the relationship between questions and passages.

In Figure 6, we plot the F1 scores of MRQA models on the SQuAD v1.1 dataset against the F1 scores on each of our new test sets, along with the linear fits from Figure 1. On the Reddit and Amazon test sets, the best MRQA model in our testbed, Delphi (Longpre et al., 2019), achieves higher F1 scores than any SQuAD model and is substantially above the linear fit. However, many of the models trained on more data exhibit little to no improved robustness. In addition, all of the models are still substantially below the human F1 scores and robustness. See Appendix E.2 for the full results table.

Table 4. Manual comparison of the reasoning required to answer each question-answer pair on a random sample of 192 examples from each dataset using the categories from Rajpurkar et al. (2016). The Reddit and Amazon datasets have more examples requiring world knowledge to resolve lexical variation, whereas the New York Times and Amazon datasets require more multi-sentence reasoning. We show in Appendix C.6 that these differences in reasoning required do not explain the performance drops we observe.

Reasoning Type	SQuAD v1.1	New Wiki	NYT	Reddit	Amazon
Lexical Variation (Synonomy)	39.1	39.1	31.8	35.9	36.5
Lexical Variation (World Knowledge)	8.3	4.7	9.9	20.3	18.8
Syntactic Variation	62.5	53.6	50.5	53.1	46.4
Multiple Sentence Reasoning	8.9	8.3	16.7	12.0	16.7
Ambiguous	1.6	3.6	1.6	1.6	1.0

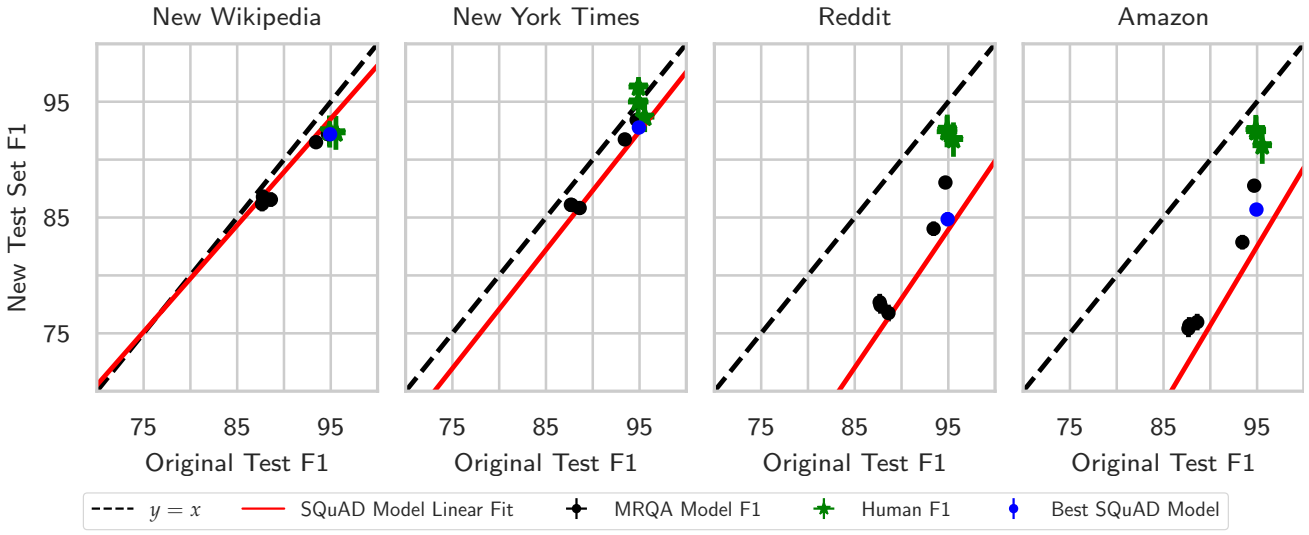

Figure 6. Model from the MRQA Shared Task 2019, trained on 5 datasets beyond SQuAD, and human F1 scores on the original SQuAD test set and each of our new test sets. The error bars are 95% Student's t-confidence intervals. Although the MRQA models still lag human performance and robustness across datasets, these models, particularly those with high F1 scores on the original SQuAD, exhibit increased robustness and generalization across each of the datasets compared to models that are only trained on SQuAD.

7. Discussion

Despite years of test set reuse, we find no evidence of adaptive overfitting on SQuAD. Our findings demonstrate that natural language processing benchmarks like SQuAD continue to support progress much longer than than reasoning from first principles might have suggested.

While SQuAD models generalize well to new examples from the same distribution, results on our new test sets show that robustness to distribution shift remains a challenge. On each of our test sets, a strong human baseline is largely unchanged, but SQuAD models suffer non-trivial and nearly uniform performance drops. While question answering models have made substantial progress on SQuAD, there has been less progress towards closing the robustness gap under natural distribution shifts. This highlights the need to move beyond model evaluation in the standard, i.i.d. setting, and to explicitly incorporate distribution shifts into evaluation. We hope our new test sets offer a helpful starting point.

There are multiple promising avenues for future work. One direction is constructing metrics for comparing datasets that can explain the performance differences we observe. Why do models perform so well on New York Times, but experience much larger drops on Reddit and Amazon? Stratifying our datasets using common criteria like answer type or reasoning required appears insufficient to answer this question. Another important direction is to better understand the interplay between additional data and model robustness. Some of the models from the MRQA challenge, e.g., Delphi (Longpre et al., 2019), benefit substantially from training with additional data, while other models remain near the same linear trend line as the SQuAD models. From both empirical and theoretical perspectives, it would be interesting to better understand when and why training with additional data improves robustness, and to offer concrete guidance on how to collect and use additional data to improve robustness to distribution shifts.

Acknowledgments

We thank Pranav Rajpurkar, Robin Jia, and Percy Liang for providing us with the original SQuAD data generation pipeline and answering our many questions about the SQuAD dataset. We thank Nelson Liu for generously providing many of the SQuAD models we evaluated, substantially increasing the size of our testbed. We also thank the Codalab team for supporting our model evaluation efforts. This research was generously supported in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1752814 ABC, an Amazon AWS AI Research Award, and a gift from Microsoft Research.

References

Bahdanau, D., Bosc, T., Jastrzebski, S., Grefenstette, E., Vincent, P., and Bengio, Y. Learning to compute word embeddings on the fly. *arXiv preprint arXiv:1706.00286*, 2017.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pp. 830–839, 2020.

Berant, J., Srikumar, V., Chen, P.-C., Vander Linden, A., Harding, B., Huang, B., Clark, P., and Manning, C. D. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1499–1510, 2014.

Blum, A. and Hardt, M. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pp. 1006–1014, 2015.

Chen, Z., Yang, R., Cao, B., Zhao, Z., Cai, D., and He, X. Smarnet: Teaching machines to read and comprehend like human. *arXiv preprint arXiv:1710.02772*, 2017.

Clark, C. and Gardner, M. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 845–855, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Dunn, M., Sagun, L., Higgins, M., Guney, V. U., Cirik, V., and Cho, K. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 117–126, 2015.

Feldman, V., Frostig, R., and Hardt, M. The advantages of multiple classes for reducing overfitting from test set reuse. In *Proceedings of the 36th International Conference on Machine Learning (ICML) 2019*, 2019.

Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., and Chen, D. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 1–13, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5801.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pp. 1–6, 2018.

Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., et al. Evaluating nlp models via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

Gong, Y. and Bowman, S. Ruminating reader: Reasoning with gated multi-hop attention. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pp. 1–11, 2018.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Hu, M., Peng, Y., Huang, Z., Qiu, X., Wei, F., and Zhou, M. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 4099–4106, 2018.

Huang, H.-Y., Zhu, C., Shen, Y., and Chen, W. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In *International Conference on Learning Representations*, 2018.

Jia, R. and Liang, P. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, 2017.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. 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)*, pp. 1601–1611, 2017.

Joshi, M., Choi, E., Levy, O., Weld, D. S., and Zettlemoyer, L. pair2vec: Compositional word-pair embeddings for cross-sentence inference. 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)*, pp. 3597–3608, 2019.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Kaushik, D., Hovy, E., and Lipton, Z. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2019.

Kitaev, N. and Klein, D. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Kong, L., de Masson d'Autume, C., Yu, L., Ling, W., Dai, Z., and Yogatama, D. A mutual information maximization perspective of language representation learning. In *International Conference on Learning Representations*, 2019.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Lee, K., Salant, S., Kwiatkowski, T., Parikh, A., Das, D., and Berant, J. Learning recurrent span representations for extractive question answering. *arXiv preprint arXiv:1611.01436*, 2016.

Lee, S., Kim, D., and Park, J. Domain-agnostic question-answering with adversarial training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 196–202, 2019.

Liu, R., Wei, W., Mao, W., and Chikina, M. Phase conductor on multi-layered attentions for machine comprehension. *arXiv preprint arXiv:1710.10504*, 2017.

Longpre, S., Lu, Y., Tu, Z., and DuBois, C. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 220–227, Hong Kong, China, November 2019. Association for Computational Linguistics.

Mania, H., Miller, J., Schmidt, L., Hardt, M., and Recht, B. Model similarity mitigates test set overuse. In *Advances in Neural Information Processing Systems*, pp. 9993–10002, 2019.

McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

Osama, R., El-Makky, N., and Torki, M. Question answering using hierarchical attention on top of BERT features. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, November 2019.

Pan, B., Li, H., Zhao, Z., Cao, B., Cai, D., and He, X. Memen: Multi-layer embedding with memory networks for machine comprehension. *arXiv preprint arXiv:1707.09098*, 2017.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pp. 2227–2237, 2018.

Qiu, R. Personal Communication, 2020.

Rajpurkar, P. Personal Communication, 2019.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Rajpurkar, P., Jia, R., and Liang, P. Know what you dont know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 856–865, 2018.

Richardson, M., Burges, C. J., and Renshaw, E. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 193–203, 2013.

Roelofs, R., Fridovich-Keil, S., Miller, J., Shankar, V., Hardt, M., Recht, B., and Schmidt, L. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems*, pp. 9175–9185. 2019.

Salant, S. and Berant, J. Contextualized word representations for reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 554–559, 2018.

Sen, P. and Saffari, A. What do models learn from question answering datasets? *arXiv preprint arXiv:2004.03490*, 2020.

Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

Shen, Y., Huang, P.-S., Gao, J., and Chen, W. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1047–1055, 2017.

Talmor, A. and Berant, J. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4911–4921, 2019.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, 2017.

Wang, S. and Jiang, J. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*, 2016.

Xiong, C., Zhong, V., and Socher, R. Dcn+: Mixed objective and deep residual coattention for question answering. In *International Conference on Learning Representations*, 2018.

Yadav, C. and Bottou, L. Cold case: The lost mnist digits. In *Advances in Neural Information Processing Systems*, pp. 13443–13452, 2019.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pp. 5753–5763, 2019.

Yogatama, D., d'Autume, C. d. M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., et al. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, 2019.

Yu, S., Indurthi, S. R., Back, S., and Lee, H. A multi-stage memory augmented neural network for machine reading comprehension. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pp. 21–30, 2018.

Yu, Y., Zhang, W., Hasan, K., Yu, M., Xiang, B., and Zhou, B. End-to-end answer chunk extraction and ranking for reading comprehension. *arXiv preprint arXiv:1610.09996*, 2016.

Zhang, J., Zhu, X., Chen, Q., Dai, L., Wei, S., and Jiang, H. Exploring question understanding and adaptation in neural-network-based question answering. *arXiv preprint arXiv:1703.04617*, 2017.

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., and Van Durme, B. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018.

Zrnic, T. and Hardt, M. Natural analysts in adaptive data analysis. In *International Conference on Machine Learning (ICML)*, 2019.