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.

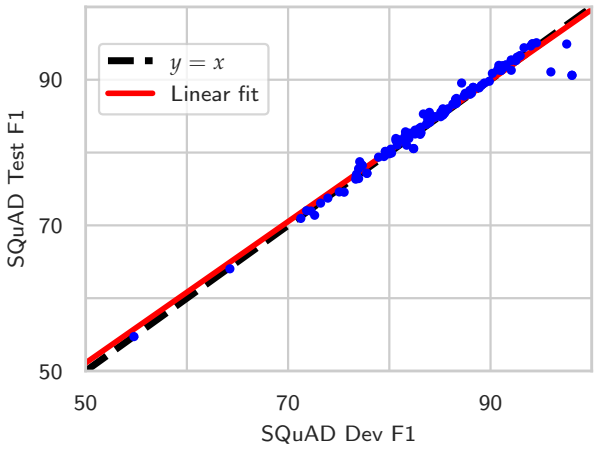

Figure 3. Comparison of F1 scores between the SQuAD v1.1 dev set and the SQuAD v1.1 test set. Despite heavy use of the dev set during model development, the dev set and test set scores closely match, with the exception of three models that were explicitly trained on the dev set, `Common-sense Governed 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

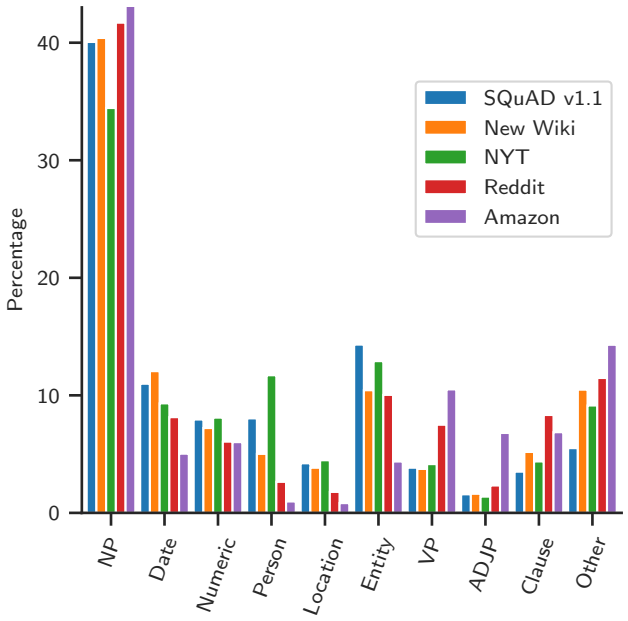

Figure 4. Comparison of answers types in the original and new datasets. We automatically partition our answers into the same categories as Rajpurkar et al. (2016). Although there are differences between the datasets, e.g., New York Times has more person answers, the four datasets are very similar. Moreover, 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.

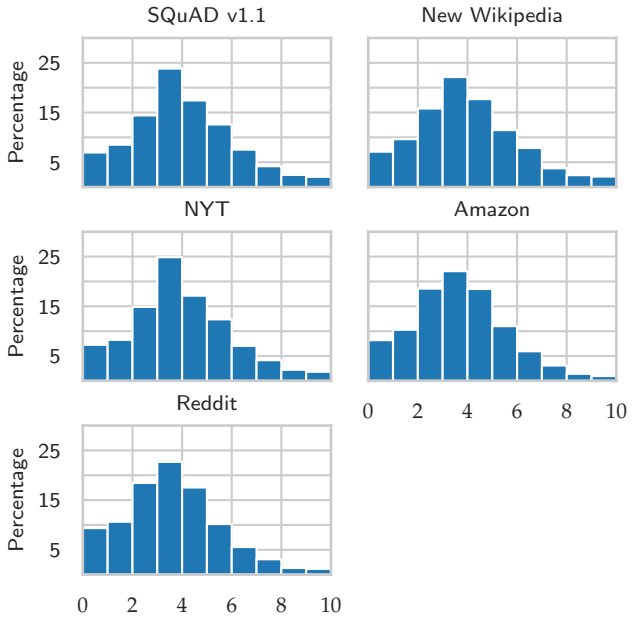

Figure 5. Histograms of syntactic divergence between question and answer sentences for both the original and new datasets. All of the datasets have a similar distribution of syntactic divergence, though the Reddit and Amazon datasets have more question-answers pairs with small (1-2) syntactic divergence.

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

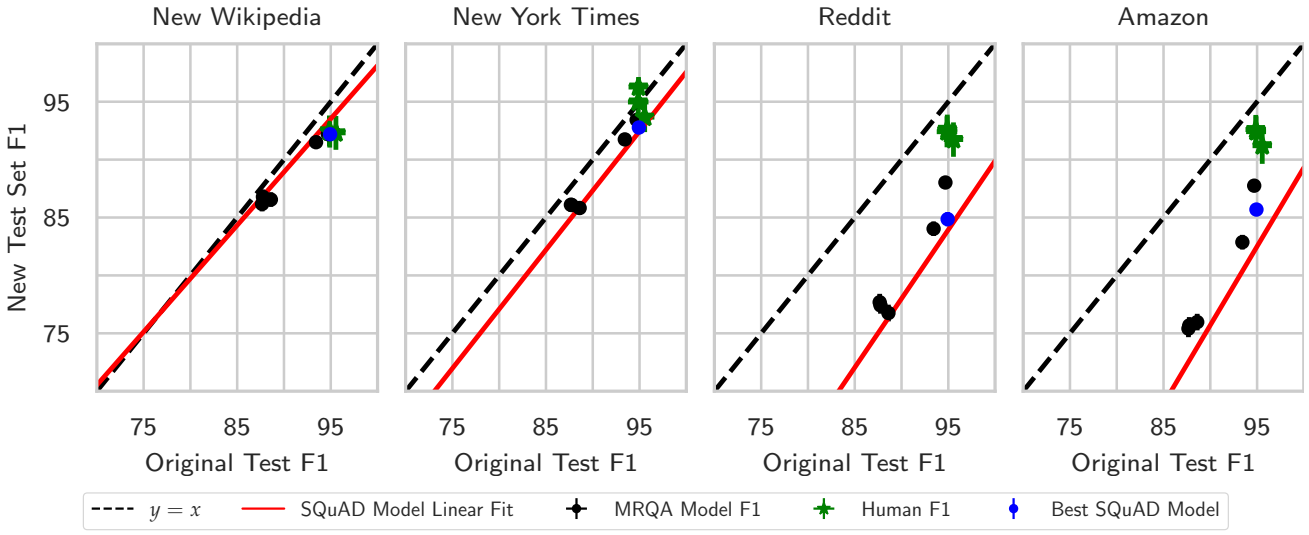

Figure 6. Model from the MRQA Shared Task 2019, trained on 5 datasets beyond SQuAD, and human F1 scores on the original SQuAD test set and each of our new test sets. The error bars are 95% Student's t-confidence intervals. Although the MRQA models still lag human performance and robustness across datasets,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

A. Evaluation Metrics

In this section, we formally define the evaluation metrics used throughout our experiments. Let $(p, q, (a^1, \ldots, a^n)$ denote a passage p, a question q, and a set of n answers $(a^1, \ldots, a^n)$. Let S denote the sampled dataset, let f denote some model, and $f(p, q) = \hat{a}$ be its predicted answer.

F1 Score. F1 measures the average overlap between the prediction and the ground-truth answer. Given answer a and prediction $\hat{a}$, consider a and $\hat{a}$ as bags of words (sets), and let $v(a, \hat{a})$ be their associated F1 score, i.e. the harmonic mean of precision and recall between the two sets. Then,

$$\text{F1}(f) = \frac{1}{|S|} \sum_{(p,q,(a^1,\ldots,a^n))\in S} \max_{i=1,\ldots,n} v(a^i, f(p, q)).$$

Exact match. Exact match measures the percentage of predictions that exactly match any one of the ground truth answers.

$$\text{ExactMatch}(f) = \frac{1}{|S|} \sum_{(p,q,(a^1,\ldots,a^n))\in S} \max_{i=1,\ldots,n} \mathbb{1}\{f(p, q) = a^i\}.$$

All of our results are reported using the evaluation script provided by Rajpurkar et al. (2016), which ignores punctuation and the articles "a", "an", and "the" when computing the above metrics.

B. Comparing Natural and Adversarial Distribution Shift

To contrast natural and adversarial distribution shifts, we evaluated all of the models in our testbed against the adversarial attacks described in Jia & Liang (2017) on the original SQuAD v1.1 dataset.

AddSent. In the `AddSent` attack, for every passage, question, and answer pair (p, q, a), Jia & Liang (2017) procedurally generate up to five new sentences to append to the passage p that do not contradict the correct answer. Each of the sentences are generated to be similar to the correct answer, and ungrammatical or contradictory sentences are removed by crowdworkers. This results in a set of new examples $(\tilde{p}_1, q, a), \ldots, (\tilde{p}_5, q, a)$ for each original example. The adversary evaluates the model f on each of the 5 examples and picks the one that gives the lowest score, $\min_{i=1,\ldots,5} s(f(\tilde{p}_i, q), a)$, where s is the scoring function (exact match or F1). In Figure 7, we compare F1 and EM scores on the original SQuAD v1.1 test set with F1 and EM scores against the adversarial `AddSent` attack.

Similar to the natural distribution shift examples, we observe the relationship between the original test F1 scores and the adversarial F1 test scores broadly follow a linear trend. However, the linear fit is not as good compared to the natural distribution shifts. There is more variability in model performance around the trend line, and this is reflected in lower a R^2 statistic, e.g. 0.72 for AddSent F1, compared to 0.99, 0.97, 0.91, and 0.89 for the New Wikipedia, New York Times, Reddit, and Amazon test sets, respectively. As with the natural distribution shift datasets, the linear fit is better in the probit domain, which we visualize in Figure 8. However, the R^2 statistic is still smaller than the corresponding statistics for our distribution shift datasets in the probit domain: 0.82 compared to $0.99, 0.96, 0.94$, and 0.94, for New Wikipedia, New York Times, Reddit, and Amazon, respectively.

AddOneSent. The `AddOneSent` attack similar to the `AddSent` attack. However, rather than take the worst of the 5 altered passages, it randomly selects one of the five on which to evaluate the model. In Figure 9, we compare F1 and EM scores on the original SQuAD v1.1 test set with F1 and EM scores against the adversarial `AddSent` attack. Since this attack does not require model access or evaluations, it is closer in spirit to the natural distribution shifts we consider. We observe much the same phenomenon as we see with `AddSent`. Model performance broadly follows a linear trend, and there is more variability around the linear trend line than in our natural distribution shift datasets.

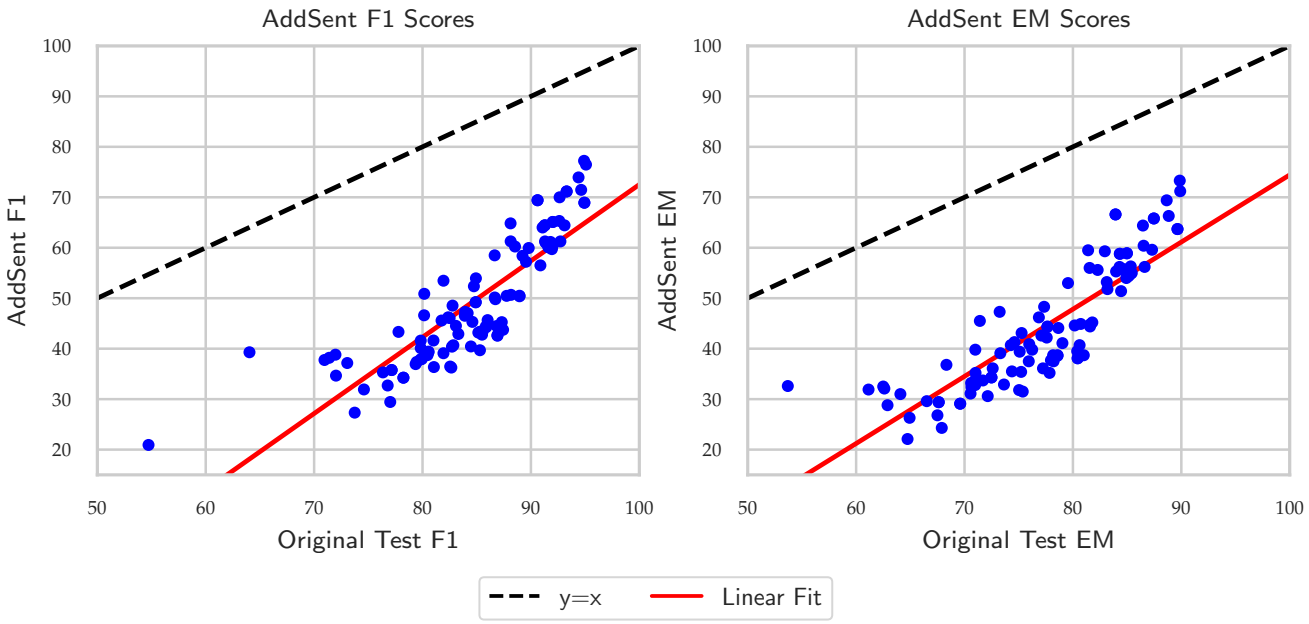

Figure 7. Comparison of F1 and EM scores on the original SQuAD test set versus the *adversarial* `AddSent` attack from Jia & Liang (2017). The models exhibit substantially more variability around the linear trend line compared to natural distribution shifts. For F1 scores, the slope of the linear fit is 1.51, for EM scores, the slope is 1.33. Similarly, the R^2 statistics are 0.73 and 0.74, respectively.

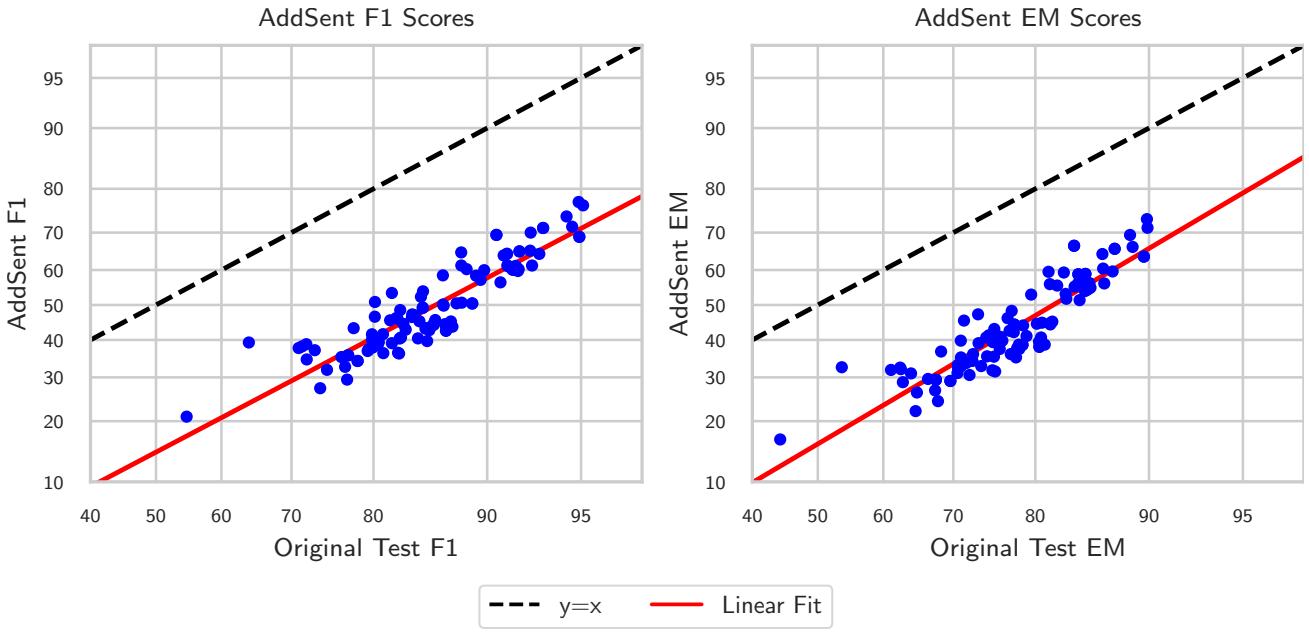

Figure 8. Comparison of F1 and EM scores on the original SQuAD test set versus the *adversarial* `AddSent` attack from Jia & Liang (2017) with *probit* scaling. For F1 scores, the slope of the linear fit is 0.99, and for EM, the slopes is 1.11. In the probit domain, the R^2 statistics are 0.82 and 0.81, respectively.

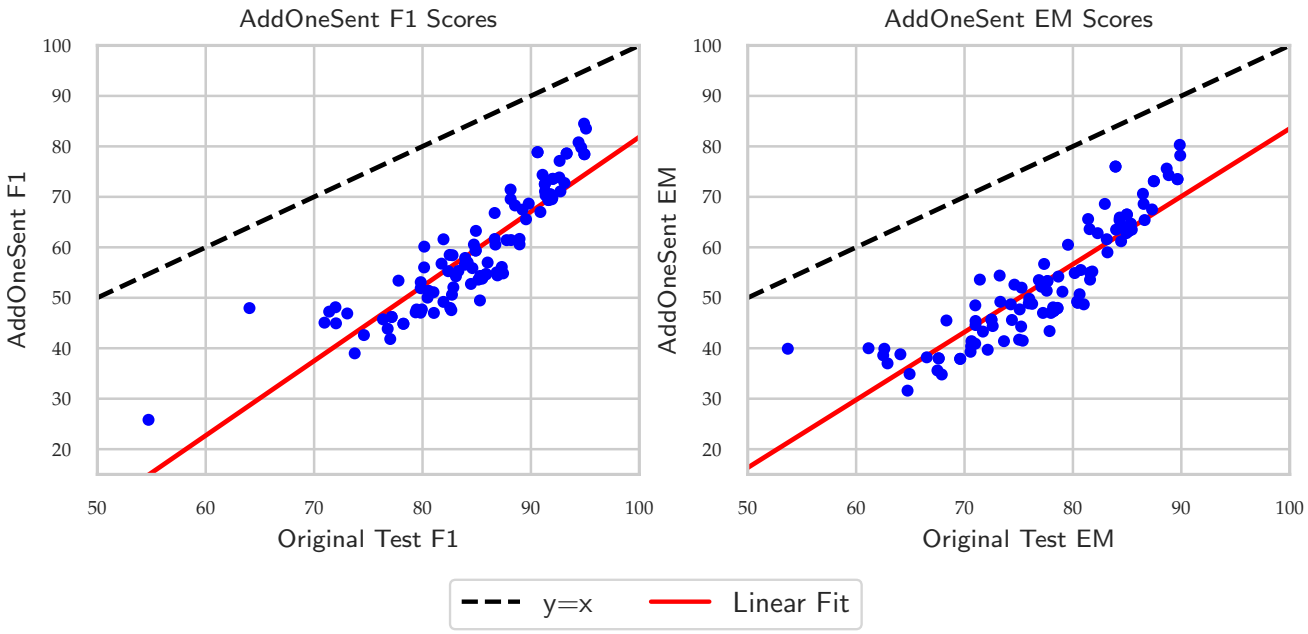

Figure 9. Comparison of F1 and EM scores on the original SQuAD test set versus the *adversarial* `AddOneSent` attack from Jia & Liang (2017). We observe similar phenomenon as with `AddSent`. Model performance broadly follows a linear trend, with more variability around the trend line than with our natural distribution test sets. For F1 scores, the slope of the linear fit is 1.48, and for EM, the slopes is 1.34. The R^2 statistics are 0.79 and 0.80, respectively.

C. Additional Analysis and Results

In this appendix, we present additional results and analysis to better understand our distribution shift experiments.

C.1. Exact Match Scatterplots

Similar to Figure 1 in Section 1, we compare the EM scores of all models in our testbed on the SQuAD v1.1 test set versus the EM scores of all models on each of the new test sets. The results are shown in Figure 10. In each case, we observe a more pronounced drop than the F1 scores with average drops of 4.6, 5.75, 20.0, and 24.8 for each of the new Wikipedia, New York Times, Reddit, and Amazon datasets, respectively. However, the primary trends are the same. In particular, we observe little evidence of overfitting on Wikipedia (the linear model nicely describes the data), and we observe a similar ranking of magnitudes of the drop on each of the other three datasets— New York Times exhibits a small drop, followed by larger drops on Reddit and Amazon.

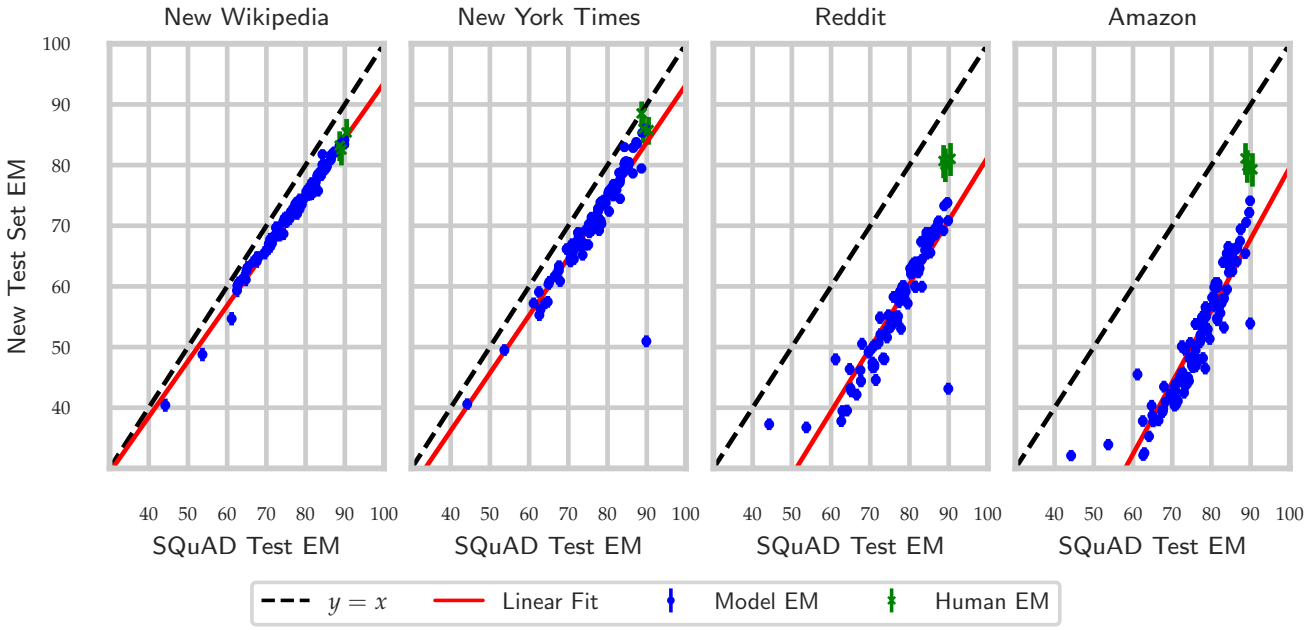

Figure 10. Model and human EM scores on the original SQuAD test set compared to our new test sets (shown with 95% Clopper-Pearson confidence intervals). The slopes of the linear fits are 0.92, 0.95, 1.05, and 1.18, respectively. The R^2 statistics are 0.99, 0.83, 0.82, and 0.85, respectively.

C.2. Linear Fits in the Probit Domain

In many cases, a linear model of F1 or EM scores is not a good fit when the scores span a wide range. In these cases, we find that a probit model describes the data better. In the main text, Figure 11 shows the F1 scores for the Amazon dataset on both the linear scale used throughout the data and a probit scale obtained by transforming all of the F1 scores with the inverse Gaussian CDF. We observe a better linear fit for our data. Figures 12 and Figures 13 show similar probit models for each of our new datasets.

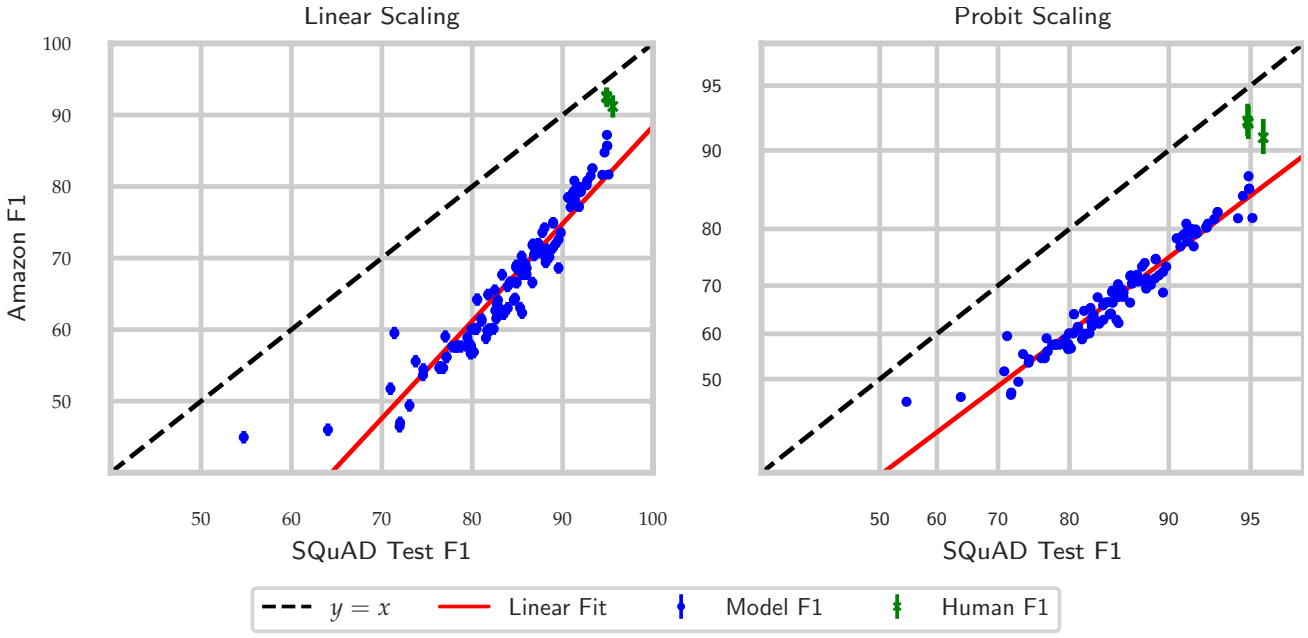

Figure 11. Comparison of model and human F1 scores on the original SQuAD v1.1 test set and our new Amazon test set. Each datapoint corresponds to one model in the testbed and is shown with 95% Student's t-confidence intervals. The left plot shows the model F1 scores under a linear axis scaling, whereas the right plot uses an *probit scale* on both axes. In other words, model F1 score x appears at $\Phi^{-1}(x)$, where Φ^{-1} is the inverse Gaussian CDF. Visual inspection shows the linear fit is better in the probit domain. Quantitatively, the R^2 statistic is 0.89 in the linear domain, compared to 0.94 in the probit domain. See Appendix C.2 for similar comparisons for all datasets.

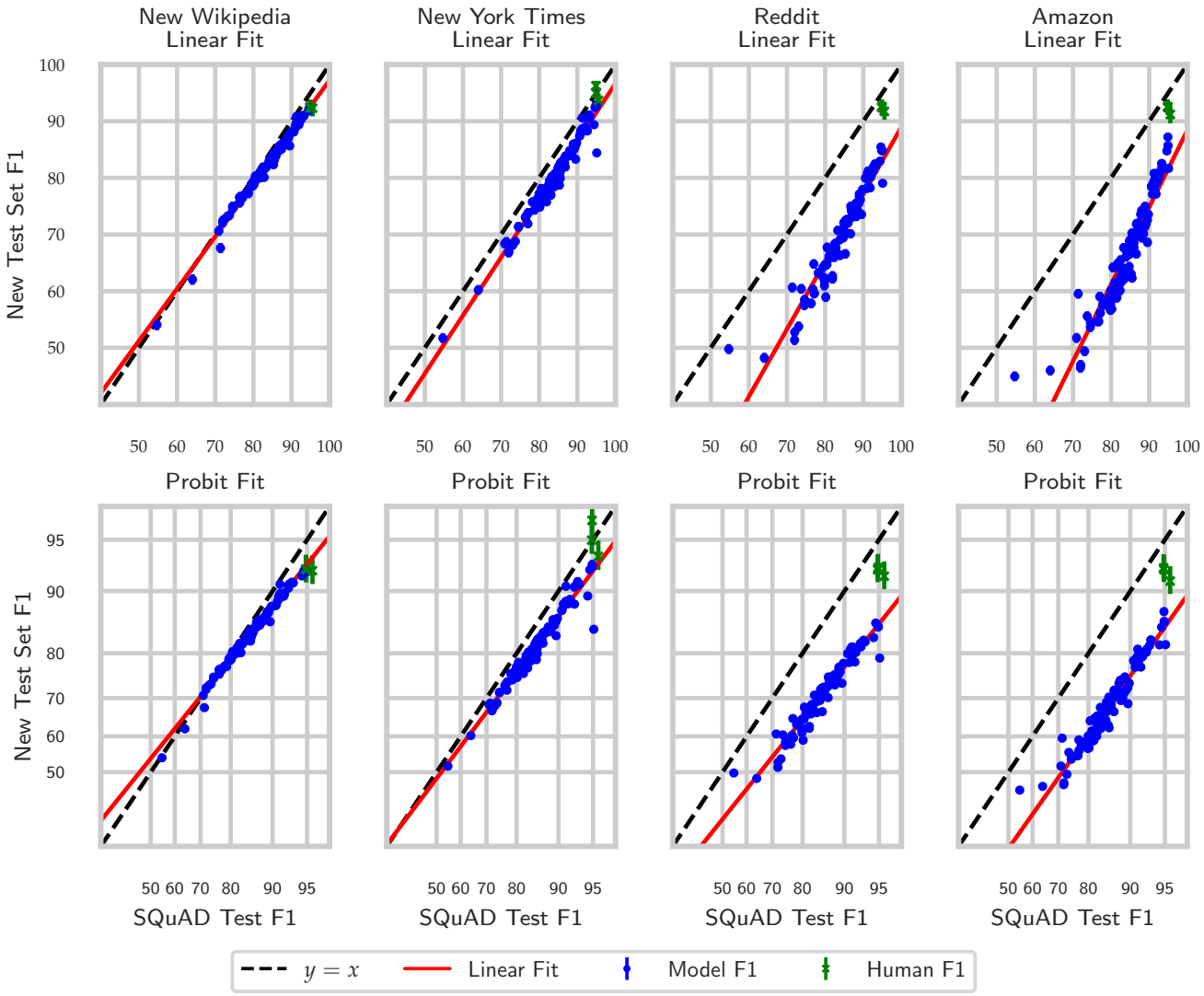

Figure 12. Comparison between linear and probit axis scaling for model and human F1 scores on the original SQuAD test and each of our new test sets. For linear axis scaling, the slopes of the linear fit are 0.92, 1.02, 1.19, and 1.36, respectively, and the R^2 statistics are $0.99, 0.97, 0.91, 0.89$, respectively. Under probit axis scaling, the slopes of the linear fit are 0.83, 0.89, 0.84, and 0.95, respectively, and the R^2 statistics are $0.99, 0.96, 0.94, 0.94$, respectively.

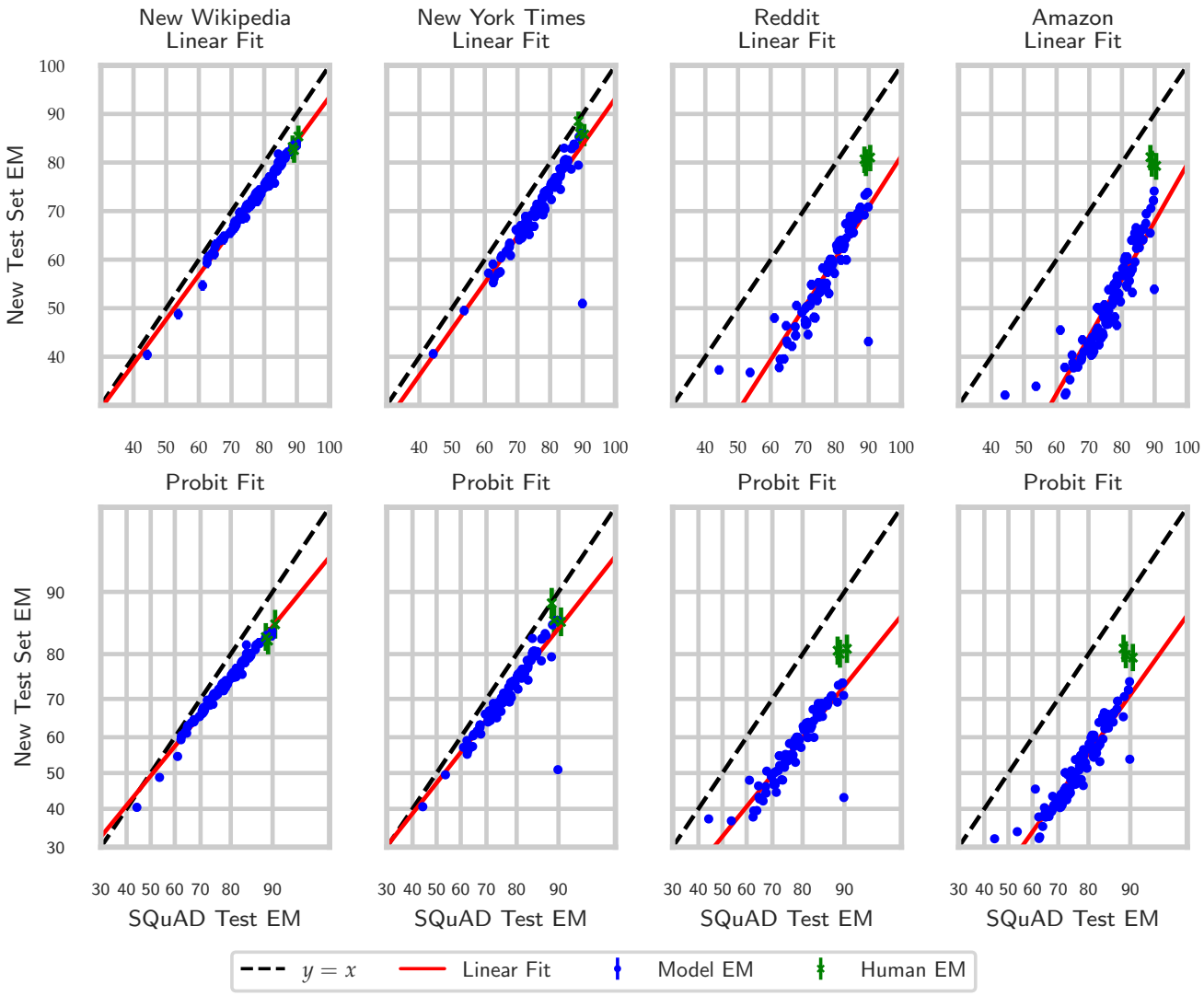

Figure 13. Comparison between linear and probit axis scaling for model and human EM scores on the original SQuAD test and each of our new test sets. Under linear axis scaling, the slopes of the linear fit are 0.92, 0.95, 1.05, and 1.18, respectively. The R^2 statistics are 0.99, 0.83, 0.82, and 0.85, respectively. Under probit scaling, the slopes of the linear fit are 0.82, 0.85, 0.83, and 0.94, respectively. The R^2 statistics are 0.99, 0.82, 0.83, and 0.88, respectively.

C.3. Does Annotator Agreement Correlate with Performance Drops?

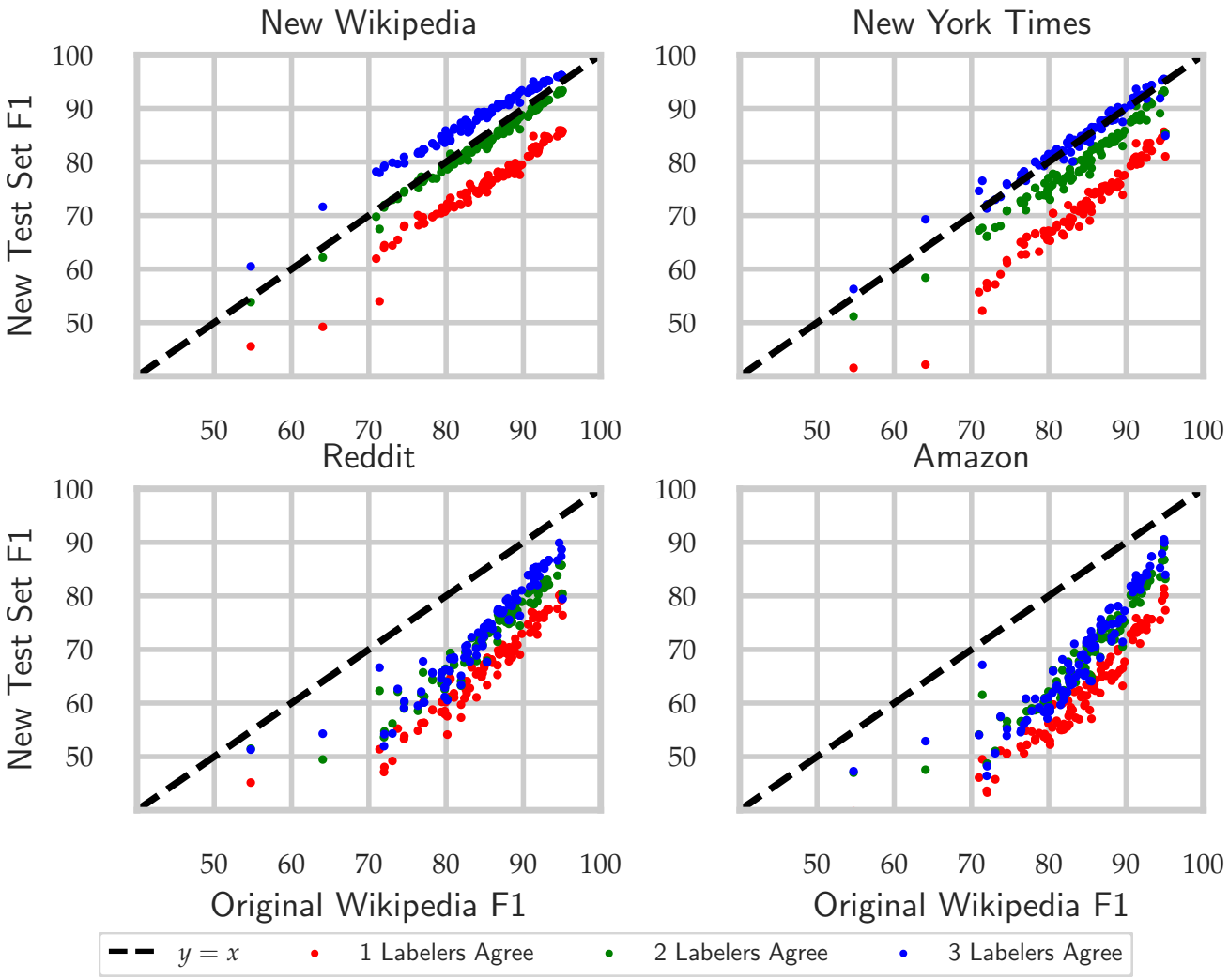

Figure 14. Model and human F1 scores on the original SQuAD v1.1 test set compared to our new test sets, stratified by the agreement between the answers given by the labellers, e.g. if three labellers agree, then three labellers provided identical (up to text normalization) answers to the question. Each point corresponds to a model evaluation. Label agreement roughly corresponds to question difficulty (and ambiguity). For clear and simple questions, all of the labellers typically agree. For more subtle or potentially ambiguous questions, the labeller's answers are more varied and tend to disagree more often. Across each dataset, when the questions are easier or less ambiguous (as measured by higher labeller agreement), the models experience proportionally smaller drops on the new dataset.

C.4. Do Shifts in Answer Category Distributions Predict Performance Drops?

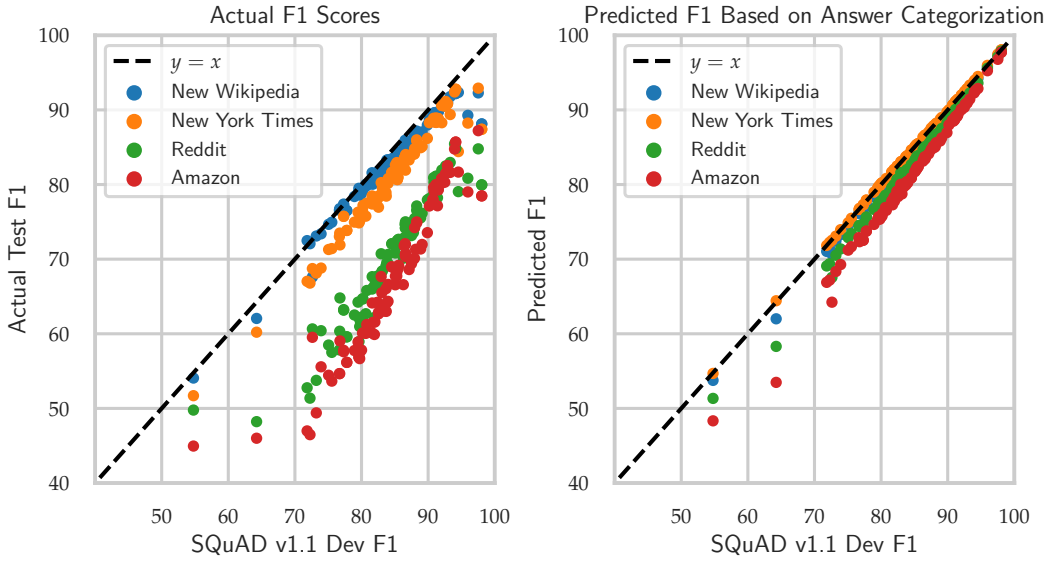

Figure 15. Changes in answer type distributions introduced in Section 6 explain little of the observed performance differences across our new datasets. For each model, we compute the F1 score on each of the answer types on the SQuAD v1.1 dev set, and then we predict the F1 score on the new test set by reweighing these F1 scores based on the frequency of answer types in the new test set. Concretely, if SQuAD v1.1 was 50% `NP` answers and 50% `Places` answers, and a model has average F1 scores of 100 for `NP` and 75 for `Places`, then if a new dataset had 30% `NP` answers and 70% `Places` answers, the predicted F1 score would be 82.5 (versus 87.5 for the original). The $y = x$ line represents the trivial model that predicts the same F1 score on the new test sets as the original. For each of the distribution shift datasets, predictions based on answer category shifts are exceedingly optimistic and explain little of the observed drops. For instance, on the Reddit dataset, answer category shifts suggest models would lose, on average, 2-3 F1 points. However, the average observed shift is 14.0 F1 points.

C.5. Do Shifts in Syntactic Divergence Predict Performance Drops?

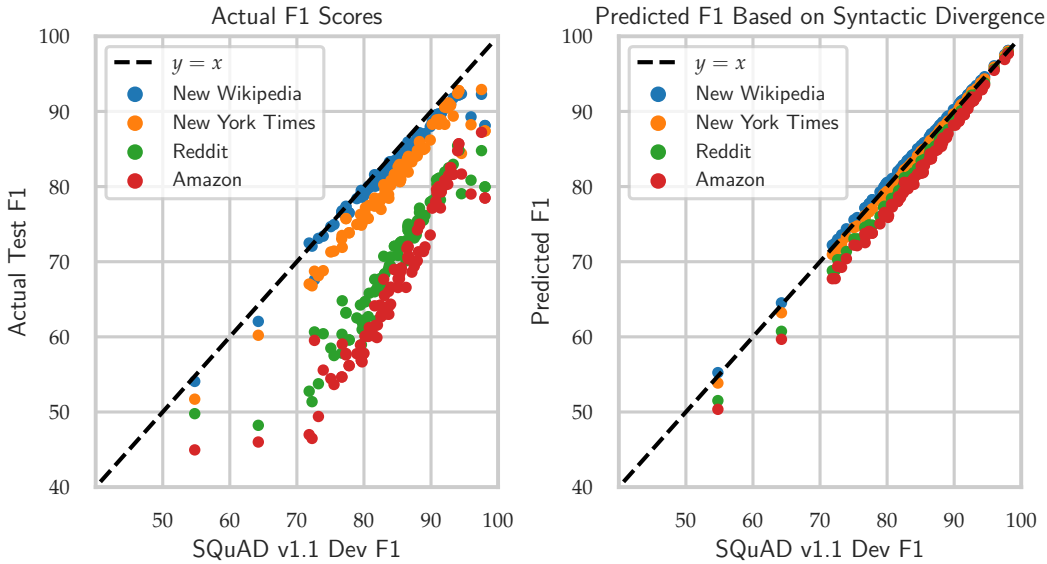

Figure 16. Changes in syntactic distributions introduced in Section 6 explain only a small amount of the observed performance differences across our new datasets. As in the previous plot, for each model, we compute the F1 score for each observed value of syntactic divergence on the SQuAD v1.1 dev set, and then we predict the F1 score on the new test set by reweighing these F1 scores based on the frequency of examples with a given syntactic divergence in the new test set. For each of the distribution shift datasets, predictions based on answer category shifts are optimistic. For instance, on the Reddit dataset, syntactic divergence shifts suggest models would lose, on average, 1.9 F1 points, while the average observed shift is 14.0 F1 points.

C.6. Do Shifts in Reasoning Required Distributions Predict Performance Drops?

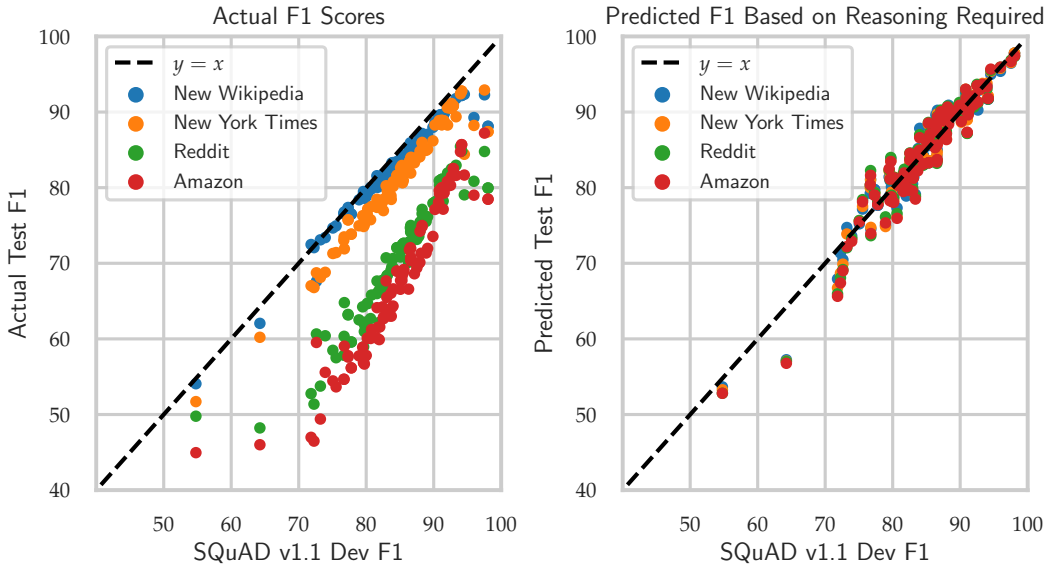

Figure 17. Changes in reasoning required distributions introduced in Section 6 explain little of the observed performance differences across our new datasets. Similar to the previous plot, for each model, we compute the F1 score on each of the reasoning required categories on the SQuAD v1.1 dev set, and then we predict the F1 score on the new test set by reweighing these F1 scores based on the reasoning required distribution of the new test set. For each of the distribution shift datasets, predictions based on reasoning required shifts closely follow the $y = x$ line corresponding to the trivial model that predicts the same F1 score on the new test sets as the original.

C.7. Does Manual Data Curation Reduce Performance Drops?

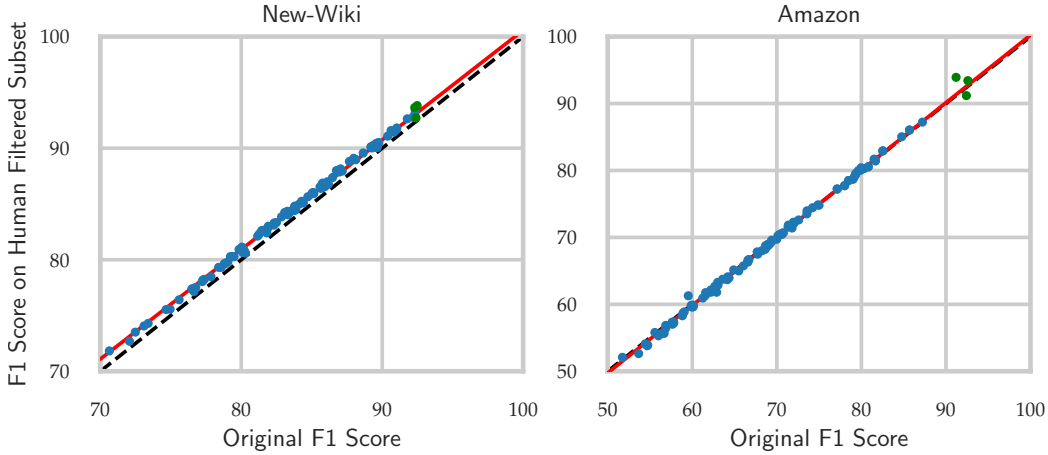

Figure 18. Comparison between model F1 scores on our New-Wikipedia and Amazon datasets and F1 scores on subsets of the datasets with additional human filtering to remove malformed, unanswerable, incorrect, ungrammatical questions and answers. To focus annotator effort on potentially bad questions, if all three MTurk annotators agreed on the answer, the question and answer were automatically marked as valid. For the New Wikipedia dataset, we manually inspected an additional 1,894 questions, removed 85 questions, and removed answers for an 444 questions. For the Amazon dataset, we manually inspected an additional 1,839 questions, removed 46 questions, and removed answers for an 282 questions. This process resulted in human curated subsets of 5574 questions for the New Wikipedia dataset and 6471 questions for the Amazon datasets. On the New Wikipedia dataset, models improve an average of 0.86 F1 points on this filtered dataset. For the filtered Amazon dataset, models slightly decreased their performance by 0.09 F1 points on average. In both cases, the rank order of the models and the linear trend observed on the full datasets without additional human filtering is preserved.

D. Dataset collection details.

In this section, we provide further details regarding our data collection pipeline.

D.1. Passage Length Statistics

We report statistics on various text length statistics. We split each paragraph into individual sentences, words, and characters using spaCy (Honnibal & Montani, 2017) and compute histograms showing the passage sentence, word, and character length distributions across each dataset.

Figures 19, 20, and 21 show the paragraph lengths in characters, words, and sentences across each dataset. In Figures 22 and 23, we show the small differences in the distribution of passage length in terms of words or sentences does not explain the observed performance drops.

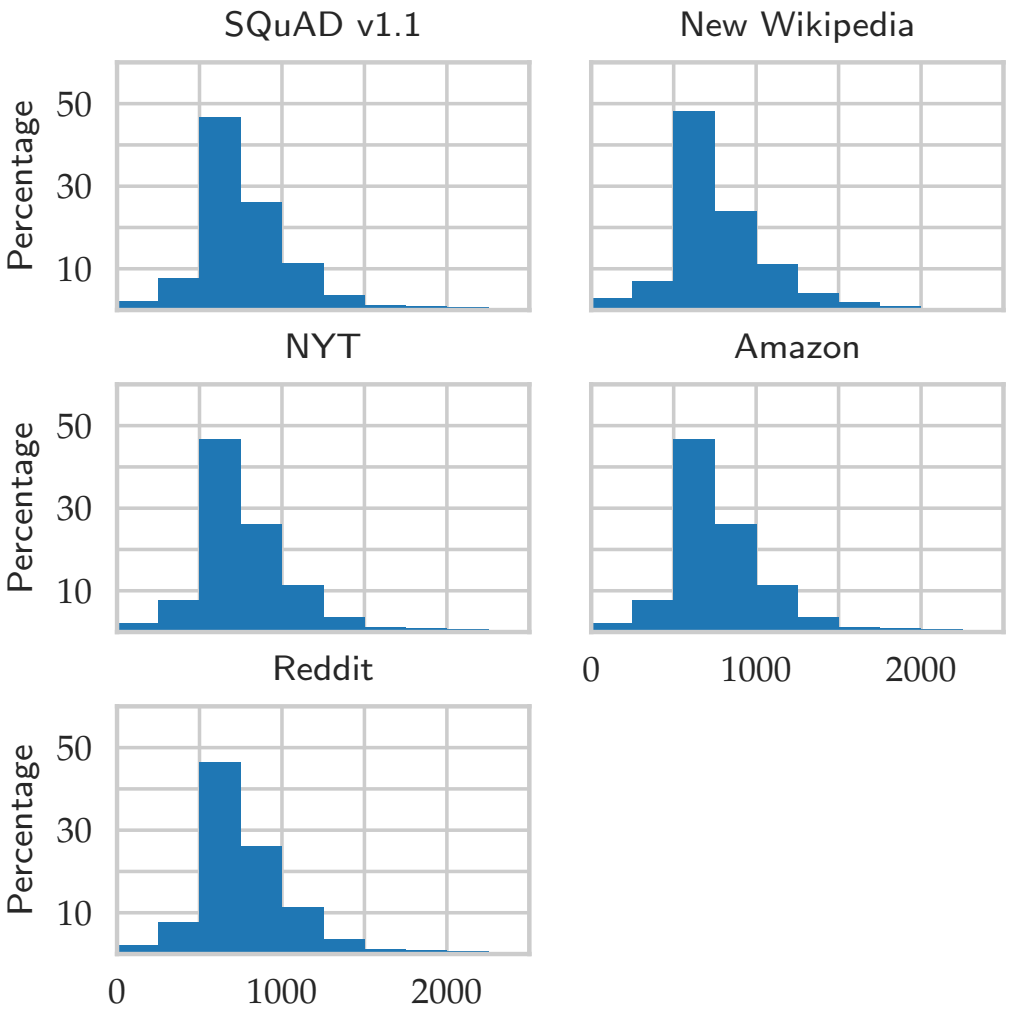

Figure 19. Histograms of the number of characters in each paragraph for the original SQuAD v1.1 development set and our new test sets. The histograms lengths match exactly since we sample in a way that ensures the character length will match for each new dataset.

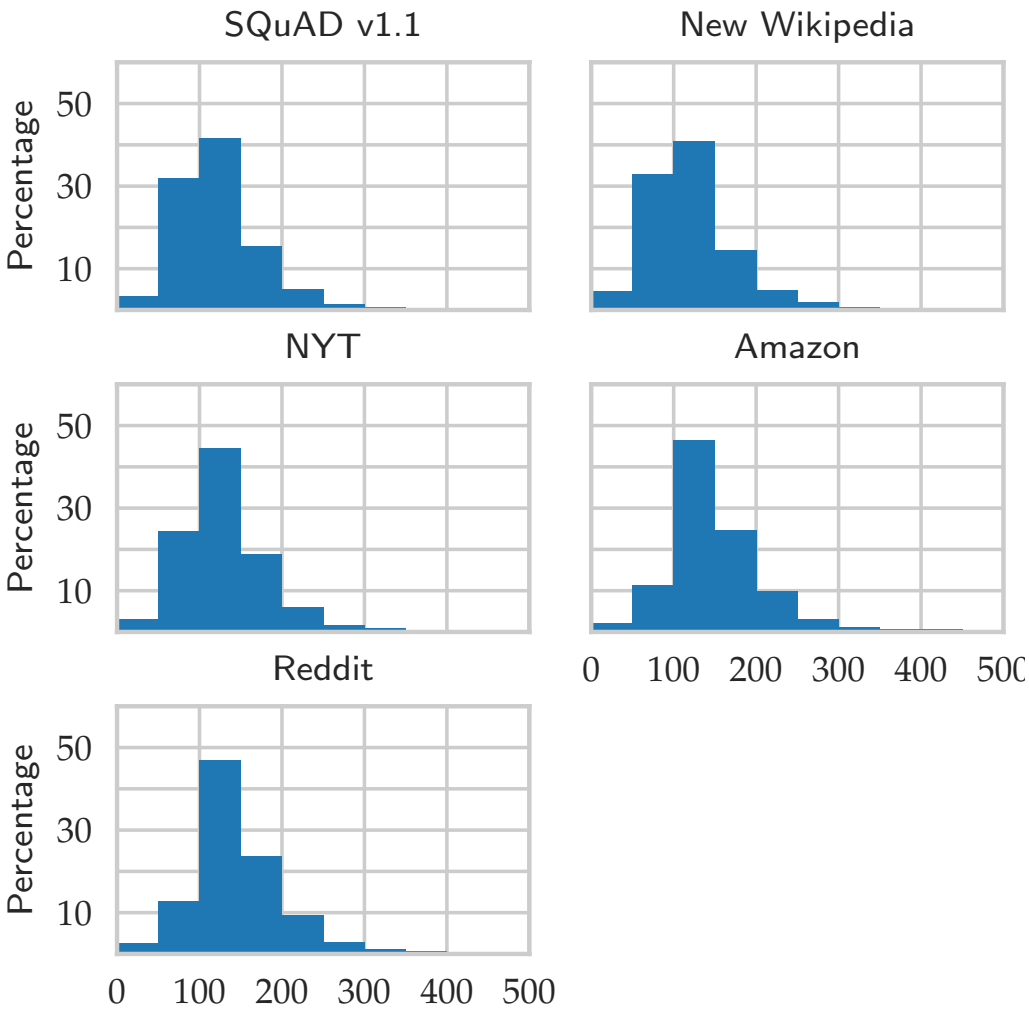

Figure 20. Histograms of the number of words in each paragraph for the original SQuAD v1.1 development set and our new test sets. The Wikipedia histograms match closely, while the Amazon and Reddit datasets' paragraphs have slightly more words. However, these differences do not explain the performance drops we observe, as Figure 22 demonstrates.

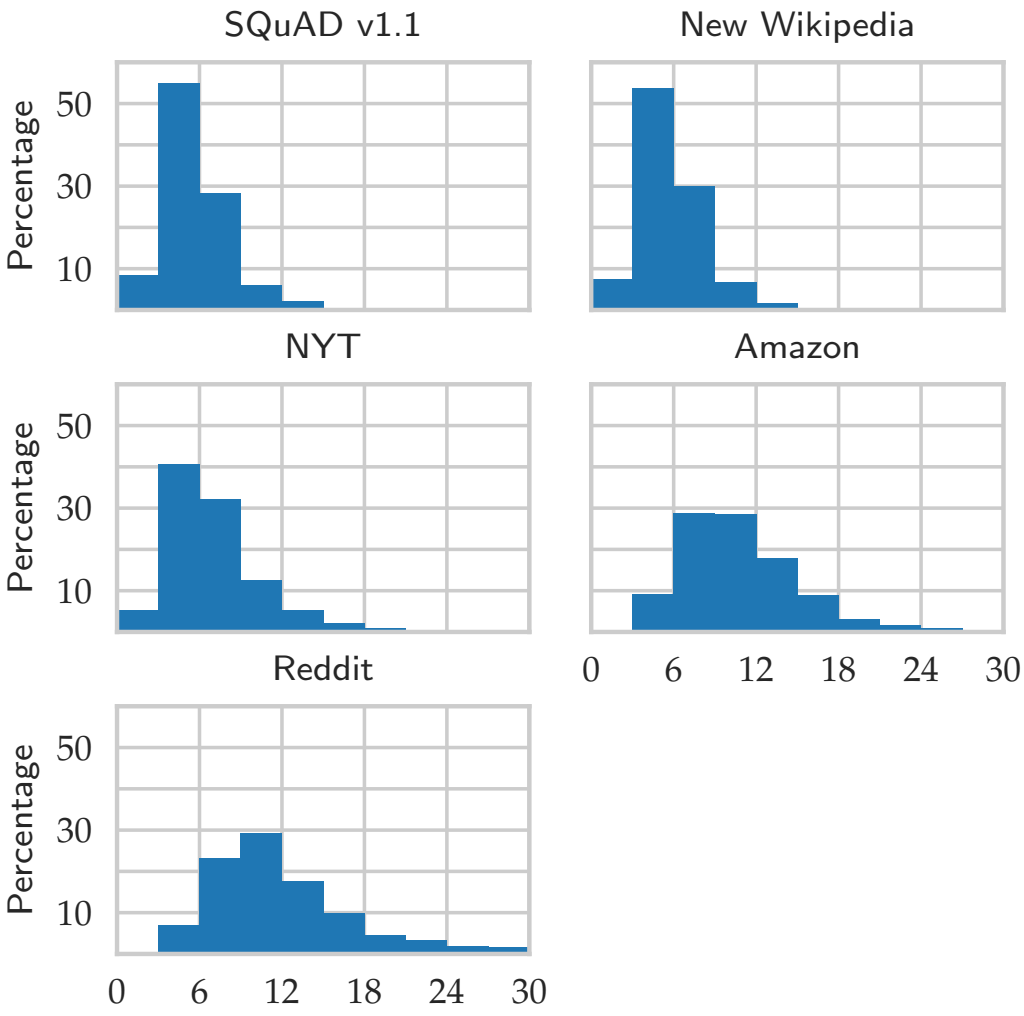

Figure 21. Histograms of the number of sentences in each paragraph for both the original and new datasets. The new Wikipedia dataset matches the SQuAD v1.1 dataset, while the other new test sets have a slightly longer tail. These slight difference in sentences per paragraph do not explain the performance drops we observe, as Figure 23 demonstrates.

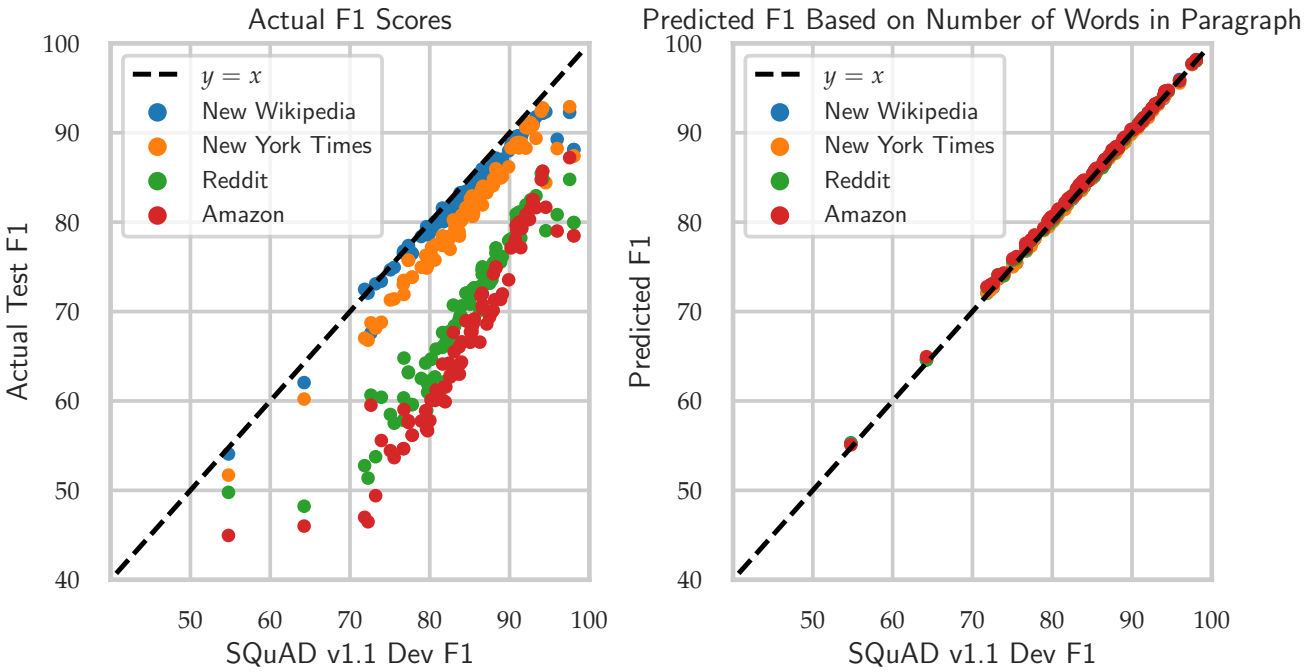

Figure 22. Changes in the distribution of words per paragraph across our new test sets do not explain the differences in F1 scores we observe. Concretely, we stratify the datasets by words per paragraph, and, for each model, we compute the F1 score for each bucket on the SQuAD v1.1 development set. We then predict F1 scores on the new test set by reweighing these F1 scores based on the paragraph length distribution of the new test set.

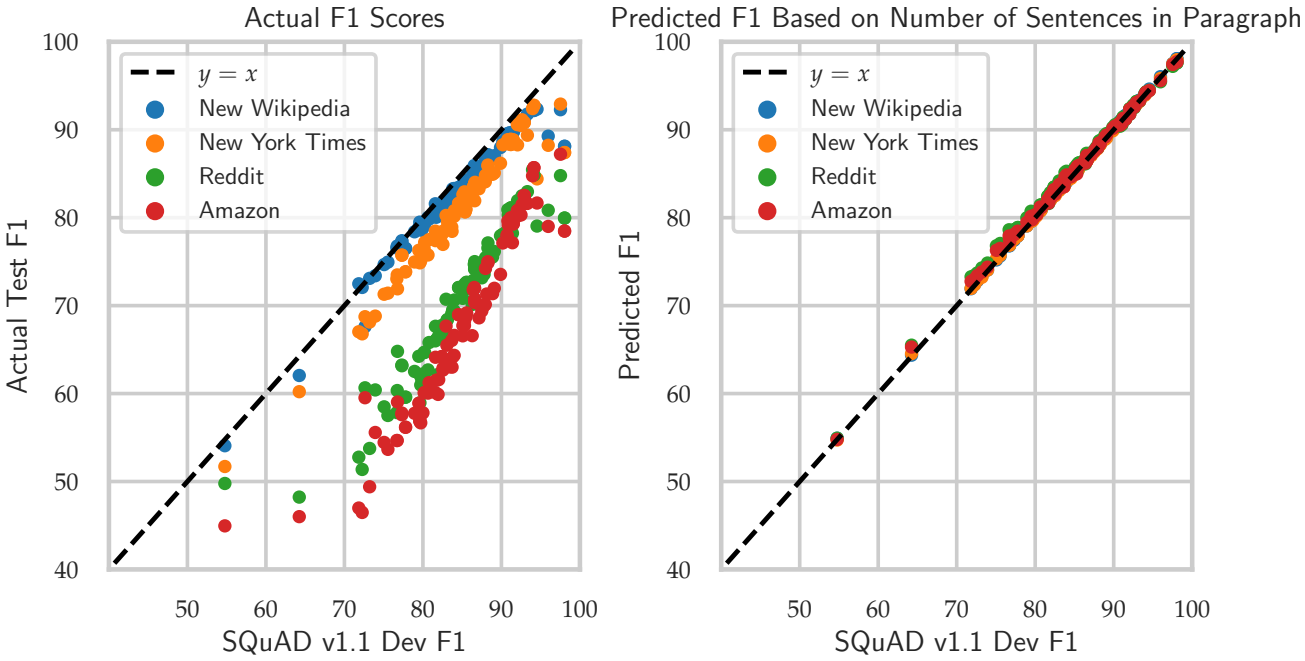

Figure 23. Changes in the distribution of sentences per paragraph across our new test sets do not explain the differences in F1 scores we observe. As in the previous plot, we stratify the datasets by sentence per paragraph, compute the F1 score for each bucket on the SQuAD v1.1 development set, and then predict F1 scores on the new test set by reweighing these F1 scores based on the paragraph length distribution of the new test set.

D.2. MTurk Experiment and UI Examples

Worker Details. Crowdworkers were required to have a 97% HIT acceptance rate, a minimum of 1000 HITs, and be located in the United States or Canada. Workers were asked to spend four minutes per paragraph when asking questions and one minute per question when answering questions. We paid workers $9.60 per hour for the amount of time required to complete each task, using an inflation rate of 6.52% between 2016 and 2019.

UI Examples. The task directions and website UI are identical to the original SQuAD data collection setup with the sole exception that the original tasks had workers ask and answer questions for all of the paragraphs for each article, whereas our tasks limit each worker to at most 5 paragraphs. Figures 24 and 25 show the directions and an example HIT for the Ask task, whereby workers pose questions for the article. Figures 26 and 27 show the directions and an example HIT for the Answer task, whereby workers answer questions posed during the Ask task.

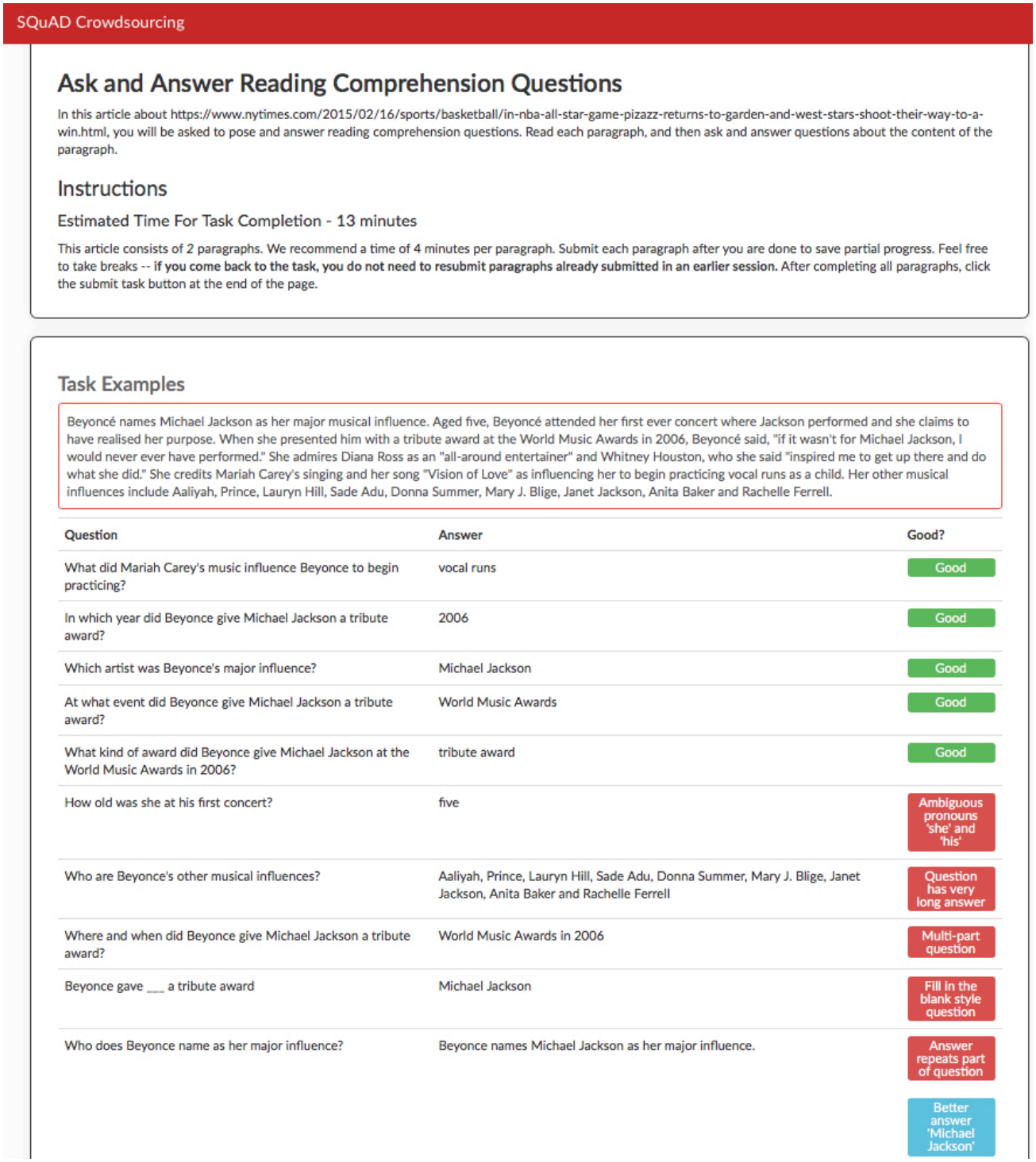

Figure 24. Ask task directions.

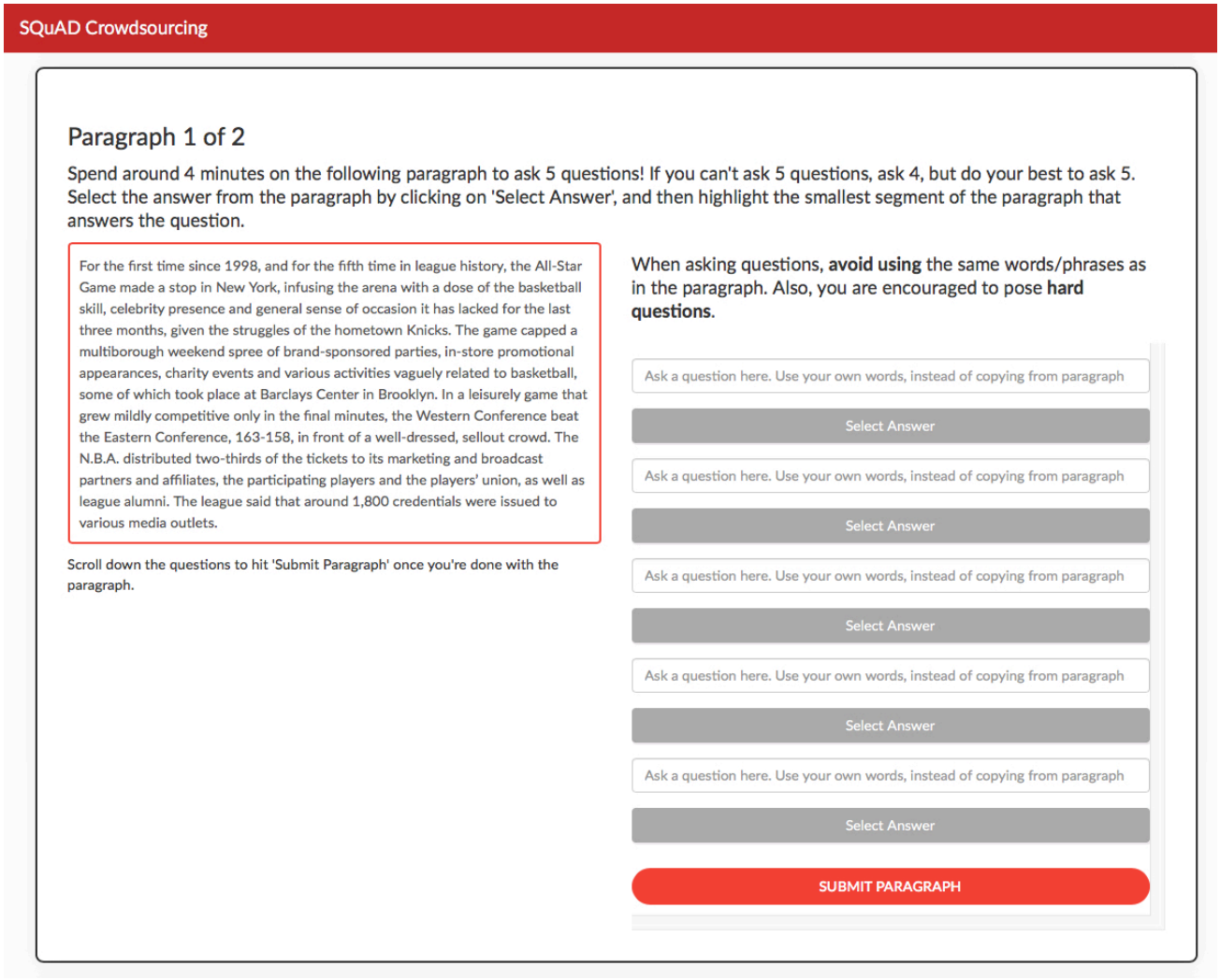

Figure 25. Ask task example.

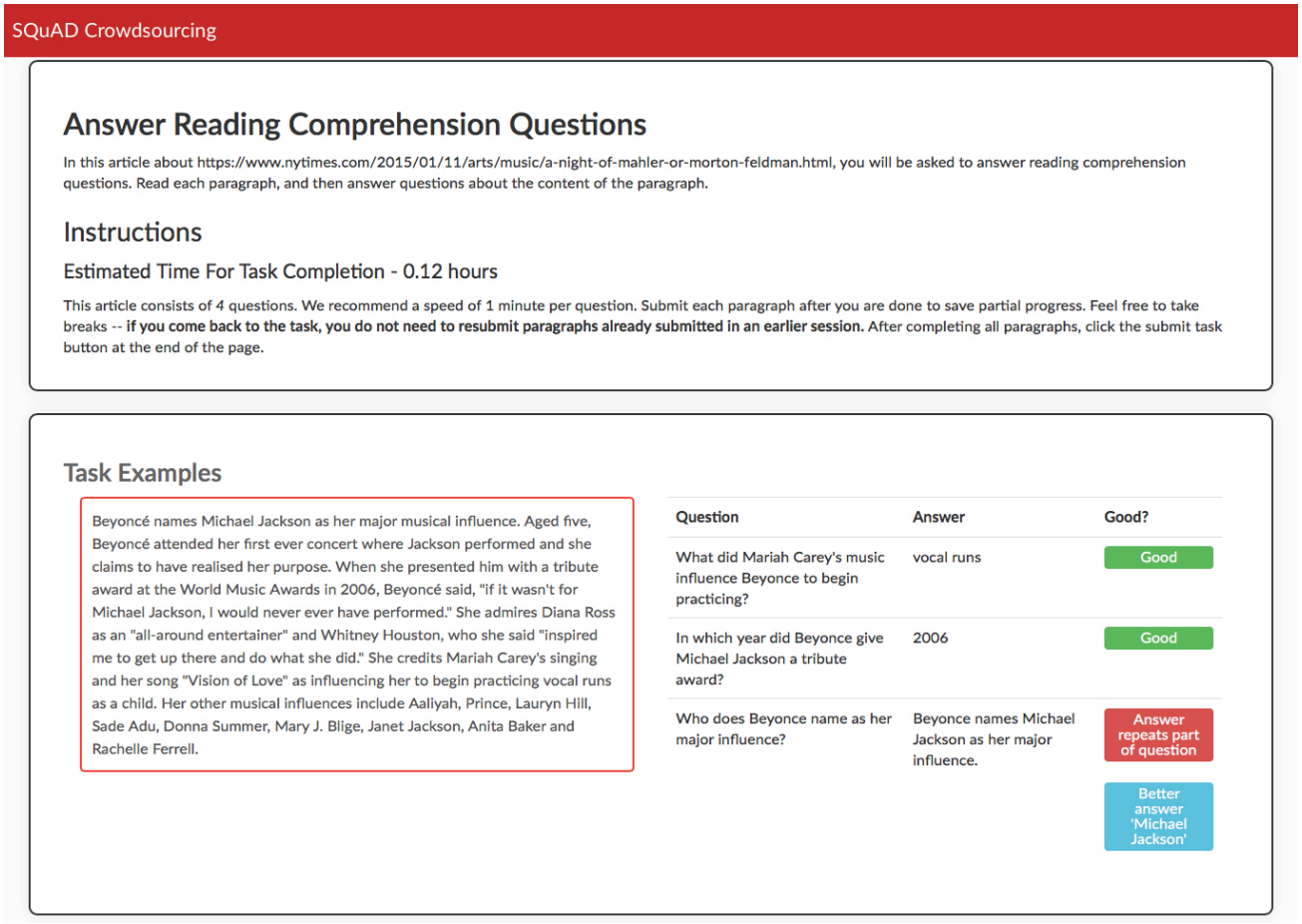

Figure 26. Answer task directions.

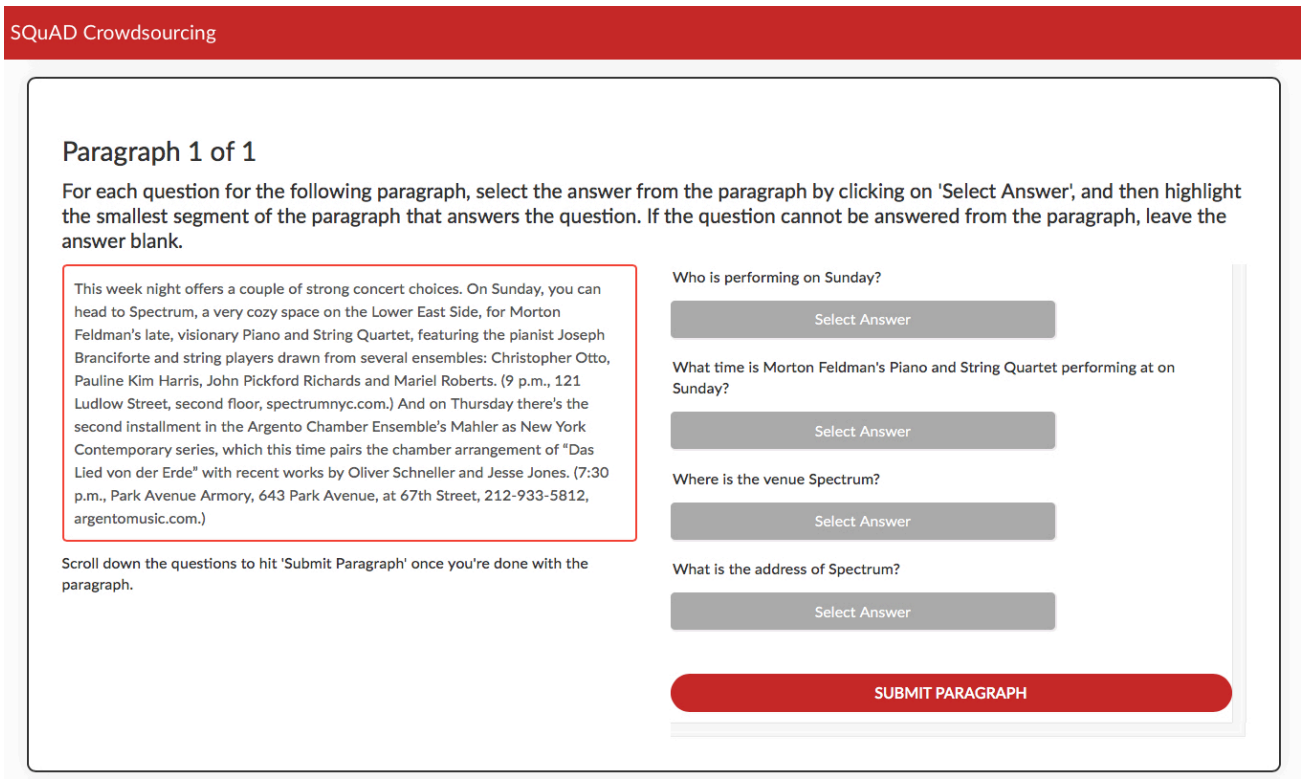

Figure 27. Answer task example.

E. Complete Model Testbed and Results Tables

In this section, we detail the complete model testbed and provide evaluation results for each model on each of our four distribution shift datasets, as well as the adversarial distributions discussed in Section B.

E.1. Models Evaluated

We evaluated a representative subset of over 100 models submitted to the SQuAD leaderboard since 2016. All of the models were submitted to the CodaLab platform, and thus we evaluate every model in the exact same configuration (weights, hyperparameters, command-line arguments, execution environment, etc.) as the original submission. Below, we list all of the models we evaluated with references, where available, and links to the Codalab submission bundle. The models are listed in sorted order based on their SQuAD v1.1 Test F1 score to allow easy reference to the subsequent tables.

1. XLNet (Single) (Yang et al., 2019)
   https://worksheets.codalab.org/bundles/0x74ebcd1a59044db49472900ae9936cf3
2. XLNet-123 (Single)
   https://worksheets.codalab.org/bundles/0x519d3e06a3544b0e85b7477ea512ec01
3. XLNet-123++ (Single)
   https://worksheets.codalab.org/bundles/0x8a03e7cddcea47fa9395ca96870b62fd
4. SpanBERT (Single) (Joshi et al., 2020)
   https://worksheets.codalab.org/bundles/0xe7315e3e35c64097af5351bb2dbdf9a5
5. BERT + WWM + MT (Single)
   https://worksheets.codalab.org/bundles/0x3975475041324f8c8b14626c932d09f4
6. Tuned BERT-1seq Large Cased (Single) (Joshi et al., 2020)
   https://worksheets.codalab.org/bundles/0xa62618d05255460a83adfe1bfd1784f7
7. InfoWord Large (Single) (Kong et al., 2019)
   https://worksheets.codalab.org/bundles/0x4a19b4d7c2fb40ef913bd97f611e66bd
8. BERT-Large Baseline (single model)
   https://worksheets.codalab.org/bundles/0xcd68d4f224b0425ab2b8b34ffb140a75
9. BERT + MT (Single)
   https://worksheets.codalab.org/bundles/0x8e20cbb02fa64883afdb4f8e50357858
10. Tuned BERT Large Cased (Single) (Devlin et al., 2019; Joshi et al., 2020)
    https://worksheets.codalab.org/bundles/0x766e1c3149bd424fb154e31ee530845a
11. DPN (Single)
    https://worksheets.codalab.org/bundles/0xd362627c900146178b5c190161bf61cf
12. ST_bl (Single)
    https://worksheets.codalab.org/bundles/0x79ca106fd7b5402abd3815636368ce2c
13. BERT uncased (Single)
    https://worksheets.codalab.org/bundles/0x1bbff660e00c4445a3dc11277039edc3
14. EL-BERT (Single)
    https://worksheets.codalab.org/bundles/0x2d65a49640394cba8632f765b237a41f
15. BISAN (single model)
    https://worksheets.codalab.org/bundles/0xfd43e046161f4ba89716d5d48b25ca2f
16. BERT + Sparse-Transformer (Single)
    https://worksheets.codalab.org/bundles/0xb1a4af82c1364cc1a41aef78543d0f52
17. InfoWord Base (Single) (Kong et al., 2019)
    https://worksheets.codalab.org/bundles/0xd2067806f74c4da79e81b73eca08bcba
18. InfoWord-Base (single model)
    https://worksheets.codalab.org/bundles/0xa41e1de495f84786a1c84d6f6036af0d

19. InfoWord BERT Large Baseline (Single) (Devlin et al., 2019; Kong et al., 2019)
    https://worksheets.codalab.org/bundles/0x86fb7e7680b6488daa585dcd11e41a36
20. Original BERT Large Cased (Single) (Devlin et al., 2019; Joshi et al., 2020)
    https://worksheets.codalab.org/bundles/0x6603ef1196fd409d81948e3af7b44e58
21. Commonsense Governed BERT-123 (Single; May 8th)
    https://worksheets.codalab.org/bundles/0x8eecf515978a4fd382e077efecbf90e1
22. InfoWord BERT Base Baseline (Single) (Devlin et al., 2019; Kong et al., 2019)
    https://worksheets.codalab.org/bundles/0xb6d9adf28e4241e181602540aeafa5a0
23. Commonsense Governed BERT-123 (Single; April 21st)
    https://worksheets.codalab.org/bundles/0x008044bbd7f74a7a81b51cbcfdf5a654
24. MARS (Ensemble; June 20th)
    https://worksheets.codalab.org/bundles/0xb320588e9f424639b54f1f40de9b0cf9
25. MARS (Single; September 1st)
    https://worksheets.codalab.org/bundles/0xfaf7cb0df0af4bf5a4050a53b81be174
26. MARS (Single; June 21st)
    https://worksheets.codalab.org/bundles/0xfc0c5b744d2a4c6b9f709c98bc2cf4e9
27. MMIPN (Single)
    https://worksheets.codalab.org/bundles/0xc2c7813ec5e241e2a0c43da45c7ecc91
28. MARS (Single; May 9th)
    https://worksheets.codalab.org/bundles/0x6d7c8a0f92374218ab4d419b397d67eb
29. Reinforced Mnemonic Reader (Ensemble) (Hu et al., 2018)
    https://worksheets.codalab.org/bundles/0x0a5ea1308bad49b2bcdd37250bdf844a
30. AttentionReader+ (Ensemble)
    https://worksheets.codalab.org/bundles/0x50985a93bf734c40b76b8cc915fe967b
31. Reinforced Mnemonic Reader + A2D (Single)
    https://worksheets.codalab.org/bundles/0x6ada3ab4807442a4944de1e8ec1f5681
32. Reinforced Mnemonic Reader + A2D + DA (Single)
    https://worksheets.codalab.org/bundles/0xeb52c2067dca498d852ca693eb9fd68a
33. BERT-Compound-DSS (Single)
    https://worksheets.codalab.org/bundles/0xd74488aac2e04d47983cbee5e7a8a106
34. BERT-Compound (Single)
    https://worksheets.codalab.org/bundles/0xc0dc1a25c03e4ba493ec28eef0e643b6
35. BiDAF + Self-Attention + ELMo (Ensemble) (Peters et al., 2018)
    https://worksheets.codalab.org/bundles/0x35b427e3105a46498256e3ccd502e442
36. AVIQA+ (Ensemble)
    https://worksheets.codalab.org/bundles/0x0109d51630ac45599a85523d4690afd1
37. EAZI (Ensemble)
    https://worksheets.codalab.org/bundles/0x55c1434feb8d48dfb990756ee1ce86d8
38. EAZI+ (Ensemble)
    https://worksheets.codalab.org/bundles/0x0b44f79d1e8042dd94943a35a057d7ea
39. MEMEN+ (Ensemble)
    https://worksheets.codalab.org/bundles/0x065d328704784db7b093f3e750ff1b46
40. DNET (Ensemble)
    https://worksheets.codalab.org/bundles/0x5b80aaba5fde4f65823746bb9b8a8fdc
41. BERT-Independent (Single)
    https://worksheets.codalab.org/bundles/0x82178b8ab098491aabec5b3a1ed18994

42. Reinforced Mnemonic Reader (Single) (Hu et al., 2018)
https://worksheets.codalab.org/bundles/0x78c31b2a1b9846a4b9de7dd71124656b

43. FusionNet (Ensemble) (Huang et al., 2018)
https://worksheets.codalab.org/bundles/0xd4ff6ed2458e4df099ea677a20115128

44. MDReader (Single)
https://worksheets.codalab.org/bundles/0xed0bb85059b04ce79db37982b1381801

45. BiDAF + Self Attention + ELMo (single model)
https://worksheets.codalab.org/bundles/0x11f631b3e7cb4a0f8acbd60491f729b6

46. BiDAF + Self-Attention + ELMo (Single) (Peters et al., 2018)
https://worksheets.codalab.org/bundles/0x5ab1fa7d11f04c5991d5011471ebdc4c

47. MDReader0 (Single)
https://worksheets.codalab.org/bundles/0x17bde05ef4b4483a9acf9e1ef8cc9326

48. BiDAF++ + pair2vec (Single) (Joshi et al., 2019)
https://worksheets.codalab.org/bundles/0x1720fa746b0243e19692820fd930b14e

49. Conductor-net (Ensemble) (Liu et al., 2017)
https://worksheets.codalab.org/bundles/0x21d981f8667141b5bf6871714e3d5fd2

50. MEMEN+ (Single)
https://worksheets.codalab.org/bundles/0xf4709036e11843f88f870f4e7dea50a0

51. AVIQA v2 (Ensemble)
https://worksheets.codalab.org/bundles/0x796847815444478c842f63a97cef93a0

52. MEMEN (Single; model submitted after paper) (Pan et al., 2017)
https://worksheets.codalab.org/bundles/0x55fcc3f13d664944969bf05c59f402a4

53. Interactive AoA Reader (Ensemble)
https://worksheets.codalab.org/bundles/0x00599dfa3921413cab3a75a70722234d

54. EAZI (single model)
https://worksheets.codalab.org/bundles/0xad2056e99a0a484f8b8e4bcc2b1b0c14

55. AttentionReader+ (Single)
https://worksheets.codalab.org/bundles/0x334adb7624674e90aff7be232fb52005

56. DNET (Single)
https://worksheets.codalab.org/bundles/0x5eb36fb24feb4911888760f8554f90ac

57. BiDAF++ (Single) (Joshi et al., 2019)
https://worksheets.codalab.org/bundles/0xb9a6b77b0163453c8fb942bafa1e2cfe

58. MARS (Single; January 23rd)
https://worksheets.codalab.org/bundles/0x92ce58765d194debbadc1a165399a454

59. FRC (Single)
https://worksheets.codalab.org/bundles/0x346b188552ed4d1cb6c1bccaa6d243eb

60. Jenga (Ensemble)
https://worksheets.codalab.org/bundles/0xbc23efc53a1f4735bad72aa01546ace1

61. RaSoR + TR + LM (Single) (Salant & Berant, 2018)
https://worksheets.codalab.org/bundles/0xec9321a11b0f44e19ca8d325dcda75eb

62. gqa (single model)
https://worksheets.codalab.org/bundles/0xc8548cd7df0547dd9003a02e5505dd77

63. FusionNet (Single) (Huang et al., 2018)
https://worksheets.codalab.org/bundles/0xbe9fefbe5b544675aafee4e83ccbe1e1

64. Smarnet (Ensemble) (Chen et al., 2017)
https://worksheets.codalab.org/bundles/0x622060479ede4552bf490c942598ac3c

65. AVIQA v2 (Single)
https://worksheets.codalab.org/bundles/0x58ce6e7730b241dea20597b0a0e51b7e

66. DCN+ (Single) (Xiong et al., 2018)
https://worksheets.codalab.org/bundles/0xd38944b81f484cf6a40955778204a0cf

67. Jenga (single model)
https://worksheets.codalab.org/bundles/0x38bce62d659e43d19f56fc2ba34c3c4d

68. MixedModel (Ensemble)
https://worksheets.codalab.org/bundles/0x761449f9e327450e85938688a002bc72

69. Two-Attention + Self-Attention (Ensemble)
https://worksheets.codalab.org/bundles/0xf9087be2e1a34b96809b71e8ccaf9c56

70. MEMEN (Ensemble; original model in paper) (Pan et al., 2017)
https://worksheets.codalab.org/bundles/0x5596d3b1dceb414eab5653c5ec8f1607

71. ReasoNet (Ensemble) (Shen et al., 2017)
https://worksheets.codalab.org/bundles/0xe117260a328f484590e34b91839ce9ad

72. eeAttNet (Single)
https://worksheets.codalab.org/bundles/0x48a65548231d47a1aed7f5554f724064

73. Mnemonic Reader (Ensemble) (Hu et al., 2018)
https://worksheets.codalab.org/bundles/0xa860db3ea8854156b68da2e3a9a2f962

74. Conductor-net (Single) (Liu et al., 2017)
https://worksheets.codalab.org/bundles/0x6fce3642dc574820949b0ae40bbac564

75. Interactive AoA Reader (Single)
https://worksheets.codalab.org/bundles/0x6541c8fd5acb44cf85572d6827c22f44

76. Jenga (Single)
https://worksheets.codalab.org/bundles/0x4b25320ab45d459fb4274c15ed925322

77. SSAE (Ensemble)
https://worksheets.codalab.org/bundles/0x34a9c6dd5f3145ce9130ddba8a951254

78. jNet (Ensemble) (Zhang et al., 2017)
https://worksheets.codalab.org/bundles/0x9ba8c5bbe77c4fd399d670ca11e42695

79. BiDAF + Self-Attention (Single) (Clark & Gardner, 2018)
https://worksheets.codalab.org/bundles/0xe0b60a2436ef407cbf5fa0641c5350ba

80. Two-Attention + Self-Attention (Single)
https://worksheets.codalab.org/bundles/0xfcb73b26ac0049478c0b4ae4f09cb3c9

81. AVIQA (Single)
https://worksheets.codalab.org/bundles/0x513d75fb3d554dd6bc11dafb7ef1f5c3

82. Attention + Self-Attention (Single)
https://worksheets.codalab.org/bundles/0xbd549e52d11b42b39bd3d2fc0bbbe1da

83. Smarnet (Single) (Chen et al., 2017)
https://worksheets.codalab.org/bundles/0x733cef4d589743b8bc95a6108206c8a0

84. Mnemonic Reader (Single) (Hu et al., 2018)
https://worksheets.codalab.org/bundles/0x28ff5339d7164a2ea95db1a4a3a2a750

85. MAMCN (Single) (Yu et al., 2018)
https://worksheets.codalab.org/bundles/0x3d6ebcc7d54d44798d477e94fc840830

86. M-NET (Single)
https://worksheets.codalab.org/bundles/0x978c1865473f4a34bf23c14b152ec4e1

87. jNet (Single) (Zhang et al., 2017)
https://worksheets.codalab.org/bundles/0x8c62efeae93743018965441fe6e7ced0

88. Ruminating Reader (Single) (Gong & Bowman, 2018)
https://worksheets.codalab.org/bundles/0x5abfb433377c45f3b6e3d26c3f6cd050
89. ReasoNet (Single) (Shen et al., 2017)
https://worksheets.codalab.org/bundles/0x2356880cbc5347069d99a8cf38815dbc
90. RaSoR (Single) (Lee et al., 2016)
https://worksheets.codalab.org/bundles/0x9dba642677a4489eb8fc78969601c893
91. SimpleBaseline (Single)
https://worksheets.codalab.org/bundles/0xd78f5da9c45d4fa5bde361f9370b8a40
92. PQMN (Single)
https://worksheets.codalab.org/bundles/0x0f29cad4f3e94dcfb4560e4347d946d5
93. AllenNLP BiDAF (Single) (Seo et al., 2016; Gardner et al., 2018)
https://worksheets.codalab.org/bundles/0x8704f9226d884b5687fba7f73a462195
94. Match-LSTM w/ Ans-Ptr Boundary (Ensemble) (Wang & Jiang, 2016)
https://worksheets.codalab.org/bundles/0x0bbda0093b294c1191a9dda91c0aa9b0
95. Iterative Co-Attention Network (Single)
https://worksheets.codalab.org/bundles/0x801a86cd3dbd44ae930c7134b7ababe5
96. BiDAF-Compound-DSS (Single)
https://worksheets.codalab.org/bundles/0xc46b10050145494aa93708faa40b4013
97. BiDAF-Independent-DSS (Single)
https://worksheets.codalab.org/bundles/0x2254478ccad84effbd92de915ff063be
98. BiDAF-Independent (Single)
https://worksheets.codalab.org/bundles/0x3d6cd49604b8466ca952fda73bfb2527
99. BiDAF-Compound (Single)
https://worksheets.codalab.org/bundles/0x450cd98f9ab548049b8e28c9f225910e
100. Match-LSTM w/ Bi-Ans-Ptr Boundary (Single) (Wang & Jiang, 2016)
https://worksheets.codalab.org/bundles/0x5f678f88703f4eb0b320793ed998dc20
101. OTF Dict + Spelling (Single) (Bahdanau et al., 2017)
https://worksheets.codalab.org/bundles/0xd33f2fbd7eca4819b2c2b45371abcdf4
102. OTF Spelling (Single) (Bahdanau et al., 2017)
https://worksheets.codalab.org/bundles/0x5ce7b655beb0454da5240c17f36bce6c
103. OTF Spelling + Lemma (Single) (Bahdanau et al., 2017)
https://worksheets.codalab.org/bundles/0x308cfd9f735d4965835ec496610ea91d
104. RQA+IDR (single model)
https://worksheets.codalab.org/bundles/0x54e292cee87d4b1488b9cf0df15aeeec
105. Dynamic Chunk Reader (Single) (Yu et al., 2016)
https://worksheets.codalab.org/bundles/0x345be18cbe4541de841de3ac79d5b441
106. UQA (single model)
https://worksheets.codalab.org/bundles/0x64206b3164ea47e7a3d8a2df833c8f9b
107. UnsupervisedQA V1
https://worksheets.codalab.org/bundles/0xe1c53a62c8644e9b9d9fdfd18feb6a85

We also evaluated a subset of five models from the Machine Reading for Question Answering (MRQA) Shared Task (Fisch et al., 2019) on our new test sets. As in our primary experiments, all of the models were submitted to the CodaLab platform, and we evaluated every model in the exact same configuration as the original submission. Below, we list all of the models we evaluated with references and links to the submission bundle.

1. `Delphi` (Longpre et al., 2019)
https://worksheets.codalab.org/bundles/0x9a53e9c50f1244699c4a24aee483bd4c

2. `HierAtt` (Osama et al., 2019)
   https://worksheets.codalab.org/bundles/0x8d851db3255b485c97646c5c0ba812a2
3. `Bert-Large+Adv Train` (Lee et al., 2019)
   https://worksheets.codalab.org/bundles/0xa113983bc3fc42ff89bf3838a6177a0c
4. `BERT-cased-whole-word`
   https://worksheets.codalab.org/bundles/0x456676760aae452cb44ade00bb515b64
5. `BERT-Multi-Finetune`
   https://worksheets.codalab.org/bundles/0x5716df3b477a452a997bcebb9e179c89

The remaining models submitted to the competition were either not publicly accessible or otherwise unable to run on Codalab.

E.2. Full Results Tables

Main Results. In this section, we present the results for each SQuAD model and the 5 MRQA model listed in Appendix E.1, along with results for the three student and postdoc authors of this paper, on each of our new test sets. Tables 5, 6 7, and 8 contain the results for each our models and the three human annotators in terms of F1 score for the New Wikipedia, New York Times, Reddit, and Amazon test sets, respectively. Tables 9, 10, 11, and 12 contain the same data for exact match scores. For a particular dataset, some models are not listed if we were unable to evaluate the model on the dataset in Codalab.

Table 5: Comparison of model F1 scores on the original SQuAD test set and our new Wikipedia test set. Rank refers to the relative ordering using the original SQuAD v1.1 F1 scores, new rank refers to the ordering using the new test set scores, and Δ rank is the relative difference in ranking. The confidence intervals are 95% Student's t-intervals. Unless otherwise noted, all models are `single models`.

New Wikipedia F1 Score Summary

Rank	Name	SQuAD	New Wikipedia	Gap	New Rank	Δ Rank
-	Human-0	94.9 [93.8, 96.0]	92.5 [91.1, 93.8]	2.4	-	-
-	Human-1	94.9 [93.8, 96.0]	92.4 [91.0, 93.8]	2.5	-	-
-	Human-2	95.6 [94.5, 96.6]	92.3 [90.8, 93.8]	3.2	-	-
1	XLNet	95.1	92.3 [91.9, 92.8]	2.7	1	0
2	XLNET-123	94.9	92.2 [91.7, 92.7]	2.7	5	-3
3	XLNET-123++	94.9	92.3 [91.8, 92.7]	2.6	3	0
4	Delphi	94.7	92.2 [91.7, 92.7]	2.5	4	0
5	SpanBERT	94.6	92.3 [91.8, 92.8]	2.3	2	3
6	BERT+WWM+MT	94.4	91.8 [91.3, 92.3]	2.6	6	0
7	BERT-cased-whole-word	93.4	91.5 [91.0, 92.0]	1.9	7	0
8	Tuned BERT-1seq Large Cased	93.3	91.0 [90.5, 91.5]	2.3	9	-1
9	InfoWord (large)	93.1	91.0 [90.5, 91.6]	2.1	8	1
10	BERT-Large Baseline	92.7	90.8 [90.3, 91.3]	1.9	11	-1
11	BERT+MT	92.6	90.4 [89.9, 90.9]	2.3	13	-2
12	Tuned BERT Large Cased	92.6	90.6 [90.1, 91.1]	2.0	12	0
13	DPN	92.0	89.7 [89.2, 90.3]	2.3	14	-1
14	ST_bl	92.0	89.6 [89.0, 90.1]	2.4	16	-2
15	BERT-uncased	91.9	89.4 [88.8, 89.9]	2.6	20	-5
16	EL-BERT	91.8	89.6 [89.0, 90.1]	2.2	17	-1
17	BISAN	91.8	89.4 [88.9, 90.0]	2.3	18	-1
18	BERT+Sparse-Transformer	91.6	89.4 [88.9, 90.0]	2.2	19	-1
19	InfoWord (base)	91.4	89.2 [88.6, 89.8]	2.2	23	-4
20	InfoWord-Base	91.4	89.2 [88.6, 89.8]	2.1	22	-2
21	InfoWord BERT baseline (large)	91.3	90.8 [90.3, 91.3]	0.5	10	11
22	Original BERT Large Cased	91.3	89.6 [89.1, 90.2]	1.6	15	7
23	Common-sense Governed BERT-123 (May 8 2019)	91.1	89.3 [88.7, 89.8]	1.8	21	2
24	InfoWord BERT baseline (base)	90.9	88.7 [88.1, 89.2]	2.2	24	0

Continued on next page

New Wikipedia F1 Score Summary

Rank	Name	SQuAD	New Wikipedia	Gap	New Rank	Δ Rank
25	Common-sense Governed BERT-123	90.6	88.1 [87.5, 88.7]	2.5	25	0
26	MARS (ensemble, June 20 2018)	89.8	88.0 [87.4, 88.6]	1.8	26	0
27	MARS	89.5	85.6 [85.0, 86.3]	3.9	39	-12
28	MARS (June 21 2018)	89.2	87.0 [86.4, 87.7]	2.2	27	1
29	MMIPN	88.9	87.0 [86.3, 87.6]	2.0	28	1
30	MARS (May 9 2018)	88.9	86.8 [86.1, 87.4]	2.1	30	0
31	Bert-Large+Adv. Train	88.6	86.5 [85.9, 87.2]	2.0	31	0
32	Reinforced Mnemonic Reader (ensemble model)	88.5	86.5 [85.8, 87.1]	2.1	32	0
33	AttentionReader+ (ensemble)	88.2	86.1 [85.5, 86.8]	2.1	35	-2
34	Reinforced Mnemonic Reader + A2D	88.1	85.7 [85.1, 86.4]	2.4	38	-4
35	Reinforced Mnemonic Reader + A2D + DA	88.1	86.2 [85.5, 86.8]	1.9	33	2
36	BERT-COMPOUND-DSS	88.0	85.8 [85.1, 86.4]	2.2	37	-1
37	HierAtt	87.8	86.8 [86.2, 87.4]	1.0	29	8
38	BERT-Multi-Finetune	87.7	86.2 [85.5, 86.8]	1.5	34	4
39	BiDAF + Self Attention + ELMo (ensemble)	87.4	85.1 [84.4, 85.8]	2.4	42	-3
40	AVIQA+ (ensemble) (aviqa team)	87.3	85.9 [85.3, 86.6]	1.4	36	4
41	EAZI (ensemble)	86.9	85.1 [84.4, 85.8]	1.8	43	-2
42	MEMEN+ (Ensemble)	86.8	85.0 [84.3, 85.7]	1.8	44	-2
43	DNET (ensemble)	86.7	85.6 [84.9, 86.3]	1.1	40	3
44	BERT-INDEPENDENT	86.7	85.1 [84.5, 85.8]	1.5	41	3
45	Reinforced Mnemonic Reader	86.7	84.7 [84.0, 85.4]	1.9	45	0
46	MDReader	86.0	84.3 [83.6, 85.0]	1.7	47	-1
47	BiDAF + Self Attention + ELMo	85.9	83.8 [83.1, 84.5]	2.1	50	-3
48	BiDAF + Self-Attention + ELMo	85.8	83.8 [83.1, 84.5]	2.0	51	-3
49	MDReader0	85.5	83.7 [83.0, 84.4]	1.8	53	-4
50	Conductor-net (Ensemble)	85.5	83.1 [82.4, 83.8]	2.4	59	-9
51	MEMEN+	85.5	83.7 [82.9, 84.4]	1.8	54	-3
52	aviqa-v2 (ensemble)	85.5	84.4 [83.7, 85.1]	1.1	46	6
53	MEMEN	85.3	83.9 [83.2, 84.6]	1.5	49	4
54	Interactive AoA Reader (Ensemble)	85.3	82.3 [81.6, 83.1]	3.0	63	-9
55	EAZI	85.1	84.1 [83.4, 84.8]	1.0	48	7
56	AttentionReader+	84.9	83.8 [83.1, 84.5]	1.1	52	4
57	DNET	84.9	83.4 [82.7, 84.2]	1.5	55	2
58	BiDAF++	84.9	83.4 [82.6, 84.1]	1.5	56	2
59	MARS (Jan 23)	84.7	83.3 [82.6, 84.0]	1.4	57	2
60	FRC	84.6	83.3 [82.6, 84.0]	1.3	58	2
61	Jenga (ensemble)	84.5	82.9 [82.1, 83.6]	1.6	61	0
62	RaSoR + TR + LM	84.2	83.1 [82.4, 83.8]	1.1	60	2
63	gqa	83.9	82.3 [81.6, 83.1]	1.6	64	-1
64	FusionNet	83.9	82.5 [81.8, 83.2]	1.4	62	2
65	AVIQA-v2	83.3	81.8 [81.1, 82.5]	1.5	68	-3
66	DCN+	83.1	81.8 [81.1, 82.6]	1.3	67	-1
67	Jenga	82.8	80.1 [79.3, 80.9]	2.7	76	-9
68	Mixed model (ensemble)	82.8	81.6 [80.8, 82.3]	1.2	70	-2
69	two-attention-self-attention (ensemble)	82.7	81.9 [81.2, 82.7]	0.8	65	4
70	MEMEN (Ensemble, original model in paper)	82.7	81.4 [80.7, 82.2]	1.2	71	-1
71	ReasoNet (Ensemble)	82.6	81.3 [80.5, 82.0]	1.3	72	-1
72	eeAttNet	82.5	81.9 [81.1, 82.6]	0.6	66	6
73	Mnemonic Reader (Ensemble)	82.4	81.6 [80.8, 82.3]	0.8	69	4
74	Conductor-net	81.9	81.2 [80.4, 81.9]	0.8	73	1
75	Interactive AoA Reader	81.9	80.0 [79.3, 80.8]	1.9	78	-3
76	Jenga	81.8	80.2 [79.4, 80.9]	1.6	75	1
77	BiDAF + Self Attention	81.0	79.8 [79.1, 80.6]	1.2	79	-2
78	two-attention-self-attention	81.0	80.0 [79.3, 80.8]	1.0	77	1
79	AVIQA	80.5	80.3 [79.5, 81.1]	0.3	74	5
80	attention+self-attention	80.5	79.4 [78.6, 80.2]	1.1	81	-1
81	Smarnet	80.2	78.8 [78.0, 79.6]	1.4	86	-5
82	Mnemonic Reader	80.1	79.5 [78.7, 80.3]	0.7	80	2
83	MAMCN	79.9	79.2 [78.4, 80.0]	0.7	82	1
84	M-NET	79.8	78.8 [78.1, 79.6]	1.0	85	-1

Continued on next page

New Wikipedia F1 Score Summary						
Rank	Name	SQuAD	New Wikipedia	Gap	New Rank	Δ Rank
85	JNet	79.8	79.0 [78.3, 79.8]	0.8	83	2
86	Ruminating Reader	79.5	78.9 [78.1, 79.7]	0.6	84	2
87	ReasoNet	79.4	78.4 [77.6, 79.2]	1.0	87	0
88	RaSoR	78.7	77.2 [76.4, 78.1]	1.5	89	-1
89	SimpleBaseline	78.2	77.4 [76.6, 78.2]	0.9	88	1
90	AllenNLP BiDAF	77.2	76.5 [75.7, 77.3]	0.7	93	-3
91	Match-LSTM w/ Ans-Ptr (Ensemble)	77.0	76.6 [75.8, 77.5]	0.4	91	0
92	Iterative Co-Attention Network	76.8	76.8 [76.0, 77.6]	0.0	90	2
93	BIDAF-COMPOUND-DSS	76.4	75.6 [74.8, 76.4]	0.8	94	-1
94	BIDAF-INDEPENDENT-DSS	76.3	76.6 [75.7, 77.4]	-0.2	92	2
95	BIDAF-INDEPENDENT	74.6	74.7 [73.8, 75.5]	-0.1	96	-1
96	BIDAF-COMPOUND	74.6	74.9 [74.1, 75.8]	-0.4	95	1
97	Match-LSTM w/ Bi-Ans-Ptr Boundary	73.7	73.4 [72.5, 74.3]	0.3	97	0
98	OTF dict+spelling	73.1	73.1 [72.2, 74.0]	-0.0	98	0
99	OTF spelling	72.0	72.5 [71.6, 73.4]	-0.5	99	0
100	OTF spelling+lemma	72.0	72.1 [71.2, 72.9]	-0.1	100	0
101	RQA+IDR	71.4	67.6 [66.7, 68.5]	3.8	102	-1
102	Dynamic Chunk Reader	71.0	70.6 [69.7, 71.5]	0.3	101	1
103	UQA	64.0	62.1 [61.1, 63.0]	2.0	103	0
104	UnsupervisedQA V1	54.7	54.1 [53.1, 55.0]	0.7	104	0

Table 6: Comparison of model F1 scores on the original SQuAD test set and our New York Times test set . Rank refers to the relative ordering using the original SQuAD v1.1 F1 scores, new rank refers to the ordering using the new test set scores, and Δ rank is the relative difference in ranking. The confidence intervals are 95% Student's t-intervals. Unless noted, all models are `single models`.

NYT F1 Score Summary						
Rank	Name	SQuAD	NYT	Gap	New Rank	Δ Rank
-	Human-0	94.9 [93.8, 96.0]	95.0 [93.9, 96.1]	-0.1	-	-
-	Human-1	94.9 [93.8, 96.0]	96.3 [95.4, 97.1]	-1.4	-	-
-	Human-2	95.6 [94.5, 96.6]	93.7 [92.4, 95.0]	1.9	-	-
1	XLNet	95.1	84.4 [84.0, 84.9]	10.7	34	-33
2	XLNET-123	94.9	92.8 [92.3, 93.2]	2.2	3	-1
3	XLNET-123++	94.9	92.9 [92.5, 93.3]	2.0	2	1
4	Delphi	94.7	93.4 [93.0, 93.8]	1.3	1	3
5	SpanBERT	94.6	92.4 [92.0, 92.8]	2.2	4	1
6	BERT+WWM+MT	94.4	89.4 [88.9, 89.9]	5.0	11	-5
7	BERT-cased-whole-word	93.4	91.7 [91.3, 92.2]	1.7	5	2
8	Tuned BERT-1seq Large Cased	93.3	90.8 [90.3, 91.3]	2.5	7	1
9	InfoWord (large)	93.1	91.1 [90.7, 91.6]	2.0	6	3
10	BERT-Large Baseline	92.7	90.6 [90.1, 91.1]	2.1	8	2
11	BERT+MT	92.6	88.3 [87.7, 88.8]	4.4	22	-11
12	Tuned BERT Large Cased	92.6	90.5 [90.0, 91.0]	2.1	10	2
13	DPN	92.0	88.8 [88.3, 89.4]	3.2	14	-1
14	ST_bl	92.0	88.9 [88.3, 89.4]	3.1	13	1
15	BERT-uncased	91.9	88.9 [88.4, 89.5]	3.0	12	3
16	EL-BERT	91.8	88.5 [88.0, 89.1]	3.3	18	-2
17	BISAN	91.8	88.4 [87.9, 89.0]	3.3	19	-2
18	BERT+Sparse-Transformer	91.6	88.3 [87.8, 88.9]	3.3	20	-2
19	InfoWord (base)	91.4	88.6 [88.0, 89.1]	2.8	16	3
20	InfoWord-Base	91.4	88.6 [88.0, 89.1]	2.8	17	3
21	InfoWord BERT baseline (large)	91.3	90.6 [90.1, 91.1]	0.7	9	12
22	Original BERT Large Cased	91.3	88.6 [88.1, 89.2]	2.6	15	7
23	Common-sense Governed BERT-123 (May 8 2019)	91.1	88.2 [87.7, 88.8]	2.8	23	0

Continued on next page

NYT F1 Score Summary

Rank	Name	SQuAD	NYT	Gap	New Rank	Δ Rank
24	InfoWord BERT baseline (base)	90.9	88.3 [87.7, 88.8]	2.6	21	3
25	Common-sense Governed BERT-123	90.6	87.4 [86.8, 88.0]	3.2	24	1
26	MARS (ensemble, June 20 2018)	89.8	86.2 [85.6, 86.8]	3.6	25	1
27	MARS	89.5	83.3 [82.6, 83.9]	6.3	44	-17
28	MARS (June 21 2018)	89.2	85.1 [84.5, 85.7]	4.1	30	-2
29	MMIPN	88.9	86.0 [85.4, 86.6]	3.0	28	1
30	MARS (May 9 2018)	88.9	84.9 [84.3, 85.6]	3.9	31	-1
31	Bert-Large+Adv. Train	88.6	85.8 [85.2, 86.4]	2.8	29	2
32	Reinforced Mnemonic Reader (ensemble model)	88.5	84.1 [83.4, 84.7]	4.4	36	-4
33	AttentionReader+ (ensemble)	88.2	84.7 [84.1, 85.3]	3.5	33	0
34	Reinforced Mnemonic Reader + A2D	88.1	84.0 [83.4, 84.6]	4.1	39	-5
35	Reinforced Mnemonic Reader + A2D + DA	88.1	84.0 [83.4, 84.7]	4.1	37	-2
36	BERT-COMPOUND-DSS	88.0	84.8 [84.2, 85.4]	3.2	32	4
37	BERT-COMPOUND	87.8	84.3 [83.7, 84.9]	3.5	35	2
38	HierAtt	87.8	86.1 [85.5, 86.7]	1.7	27	11
39	BERT-Multi-Finetune	87.7	86.1 [85.5, 86.7]	1.6	26	13
40	BiDAF + Self Attention + ELMo (ensemble)	87.4	84.0 [83.4, 84.7]	3.4	38	2
41	AVIQA+ (ensemble) (aviqa team)	87.3	83.7 [83.1, 84.4]	3.6	41	0
42	EAZI (ensemble)	86.9	83.8 [83.2, 84.5]	3.1	40	2
43	MEMEN+ (Ensemble)	86.8	81.9 [81.3, 82.6]	4.9	51	-8
44	DNET (ensemble)	86.7	83.5 [82.8, 84.2]	3.2	42	2
45	BERT-INDEPENDENT	86.7	83.5 [82.8, 84.1]	3.2	43	2
46	Reinforced Mnemonic Reader	86.7	82.6 [81.9, 83.2]	4.1	49	-3
47	MDReader	86.0	82.9 [82.3, 83.6]	3.1	45	2
48	BiDAF + Self Attention + ELMo	85.9	82.7 [82.0, 83.4]	3.2	47	1
49	BiDAF + Self-Attention + ELMo	85.8	82.7 [82.0, 83.4]	3.1	46	3
50	MDReader0	85.5	82.6 [82.0, 83.3]	2.9	48	2
51	Conductor-net (Ensemble)	85.5	79.9 [79.2, 80.6]	5.6	61	-10
52	MEMEN+	85.5	81.0 [80.3, 81.6]	4.5	54	-2
53	MEMEN	85.3	80.7 [80.0, 81.4]	4.7	57	-4
54	Interactive AoA Reader (Ensemble)	85.3	78.7 [78.0, 79.4]	6.6	67	-13
55	EAZI	85.1	82.1 [81.4, 82.8]	3.1	50	5
56	AttentionReader+	84.9	81.0 [80.3, 81.6]	4.0	55	1
57	DNET	84.9	81.6 [80.9, 82.3]	3.3	52	5
58	BiDAF++	84.9	81.4 [80.7, 82.1]	3.5	53	5
59	MARS (Jan 23)	84.7	80.2 [79.5, 80.9]	4.6	60	-1
60	FRC	84.6	80.3 [79.6, 80.9]	4.3	58	2
61	Jenga (ensemble)	84.5	79.8 [79.1, 80.5]	4.7	62	-1
62	RaSoR + TR + LM	84.2	80.8 [80.1, 81.5]	3.4	56	6
63	gqa	83.9	78.4 [77.7, 79.2]	5.5	70	-7
64	FusionNet	83.9	78.9 [78.2, 79.7]	5.0	64	0
65	AVIQA-v2	83.3	80.3 [79.5, 81.0]	3.1	59	6
66	DCN+	83.1	77.0 [76.2, 77.7]	6.1	79	-13
67	Jenga	82.8	77.4 [76.7, 78.1]	5.5	76	-9
68	Mixed model (ensemble)	82.8	78.8 [78.1, 79.5]	4.0	66	2
69	two-attention-self-attention (ensemble)	82.7	79.5 [78.8, 80.2]	3.2	63	6
70	MEMEN (Ensemble, original model in paper)	82.7	78.1 [77.4, 78.9]	4.5	71	-1
71	ReasoNet (Ensemble)	82.6	78.5 [77.7, 79.2]	4.1	69	2
72	eeAttNet	82.5	78.8 [78.1, 79.6]	3.7	65	7
73	Mnemonic Reader (Ensemble)	82.4	78.5 [77.7, 79.2]	3.9	68	5
74	Conductor-net	81.9	77.6 [76.9, 78.4]	4.3	74	0
75	Interactive AoA Reader	81.9	75.8 [75.0, 76.5]	6.2	84	-9
76	Jenga	81.8	77.0 [76.3, 77.7]	4.7	78	-2
77	BiDAF + Self Attention	81.0	77.5 [76.8, 78.2]	3.5	75	2
78	two-attention-self-attention	81.0	77.9 [77.2, 78.7]	3.1	73	5
79	AVIQA	80.5	78.1 [77.4, 78.8]	2.5	72	7
80	attention+self-attention	80.5	76.5 [75.8, 77.3]	3.9	80	0
81	Smarnet	80.2	74.9 [74.1, 75.6]	5.3	88	-7
82	Mnemonic Reader	80.1	76.2 [75.5, 77.0]	3.9	82	0
83	MAMCN	79.9	77.2 [76.5, 77.9]	2.7	77	6

Continued on next page

NYT F1 Score Summary						
Rank	Name	SQuAD	NYT	Gap	New Rank	Δ Rank
84	M-NET	79.8	75.9 [75.2, 76.7]	3.9	83	1
85	JNet	79.8	75.1 [74.3, 75.9]	4.7	86	-1
86	Ruminating Reader	79.5	76.3 [75.5, 77.0]	3.2	81	5
87	ReasoNet	79.4	75.0 [74.2, 75.7]	4.4	87	0
88	RaSoR	78.7	74.3 [73.5, 75.1]	4.4	89	-1
89	SimpleBaseline	78.2	75.7 [75.0, 76.5]	2.5	85	4
90	AllenNLP BiDAF	77.2	73.8 [73.1, 74.6]	3.3	90	0
91	Match-LSTM w/ Ans-Ptr (Ensemble)	77.0	71.9 [71.1, 72.7]	5.1	94	-3
92	Iterative Co-Attention Network	76.8	73.5 [72.7, 74.3]	3.3	91	1
93	BIDAF-COMPOUND-DSS	76.4	73.0 [72.2, 73.8]	3.5	93	0
94	BIDAF-INDEPENDENT-DSS	76.3	73.0 [72.3, 73.8]	3.3	92	2
95	BIDAF-INDEPENDENT	74.6	71.3 [70.5, 72.1]	3.3	96	-1
96	BIDAF-COMPOUND	74.6	71.4 [70.6, 72.2]	3.1	95	1
97	Match-LSTM w/ Bi-Ans-Ptr Boundary	73.7	68.8 [68.0, 69.6]	4.9	97	0
98	OTF dict+spelling	73.1	68.1 [67.3, 69.0]	4.9	100	-2
99	OTF spelling	72.0	67.0 [66.2, 67.9]	5.0	101	-2
100	OTF spelling+lemma	72.0	66.8 [66.0, 67.6]	5.2	102	-2
101	RQA+IDR	71.4	68.7 [67.9, 69.6]	2.6	98	3
102	Dynamic Chunk Reader	71.0	68.4 [67.5, 69.2]	2.6	99	3
103	UQA	64.0	60.2 [59.3, 61.1]	3.8	103	0
104	UnsupervisedQA V1	54.7	51.7 [50.8, 52.6]	3.0	104	0

Table 7: Comparison of model F1 scores on the original SQuAD test set and our Reddit test set. Rank refers to the relative ordering using the original SQuAD v1.1 F1 scores, new rank refers to the ordering using the new test set scores, and Δ rank is the relative difference in ranking. The confidence intervals are 95% Student's t-intervals. Unless noted, all models are `single models`.

Reddit F1 Score Summary						
Rank	Name	SQuAD	Reddit	Gap	New Rank	Δ Rank
-	Human-0	94.9 [93.8, 96.0]	92.4 [91.1, 93.7]	2.5	-	-
-	Human-1	94.9 [93.8, 96.0]	92.6 [91.3, 93.9]	2.3	-	-
-	Human-2	95.6 [94.5, 96.6]	91.7 [90.2, 93.2]	3.8	-	-
1	XLNet	95.1	79.0 [78.5, 79.6]	16.0	21	-20
2	XLNET-123	94.9	84.9 [84.2, 85.5]	10.1	3	-1
3	XLNET-123++	94.9	84.8 [84.2, 85.4]	10.1	4	-1
4	Delphi	94.7	88.0 [87.5, 88.6]	6.7	1	3
5	SpanBERT	94.6	85.4 [84.9, 86.0]	9.2	2	3
6	BERT+WWM+MT	94.4	83.0 [82.3, 83.6]	11.4	6	0
7	BERT-cased-whole-word	93.4	84.0 [83.4, 84.7]	9.4	5	2
8	Tuned BERT-1seq Large Cased	93.3	82.2 [81.5, 82.9]	11.1	8	0
9	InfoWord (large)	93.1	82.5 [81.8, 83.1]	10.6	7	2
10	BERT-Large Baseline	92.7	81.2 [80.6, 81.9]	11.5	11	-1
11	BERT+MT	92.6	81.9 [81.3, 82.6]	10.7	9	2
12	Tuned BERT Large Cased	92.6	81.5 [80.9, 82.2]	11.1	10	2
13	DPN	92.0	80.7 [80.0, 81.4]	11.3	17	-4
14	ST_bl	92.0	80.9 [80.2, 81.6]	11.1	14	0
15	BERT-uncased	91.9	80.1 [79.5, 80.8]	11.8	19	-4
16	EL-BERT	91.8	78.2 [77.5, 78.9]	13.6	24	-8
17	BISAN	91.8	80.3 [79.6, 81.0]	11.5	18	-1
18	BERT+Sparse-Transformer	91.6	81.1 [80.4, 81.8]	10.5	13	5
19	InfoWord (base)	91.4	78.5 [77.8, 79.2]	12.9	22	-3
20	InfoWord-Base	91.4	78.5 [77.8, 79.2]	12.9	23	-3
21	InfoWord BERT baseline (large)	91.3	81.2 [80.6, 81.9]	10.1	12	9
22	Original BERT Large Cased	91.3	80.7 [80.1, 81.4]	10.5	16	6

Continued on next page

Reddit F1 Score Summary

Rank	Name	SQuAD	Reddit	Gap	New Rank	Δ Rank
23	Common-sense Governed BERT-123 (May 8 2019)	91.1	80.8 [80.2, 81.5]	10.2	15	8
24	InfoWord BERT baseline (base)	90.9	78.1 [77.4, 78.8]	12.8	25	-1
25	Common-sense Governed BERT-123	90.6	80.0 [79.3, 80.6]	10.7	20	5
26	MARS (ensemble, June 20 2018)	89.8	77.9 [77.2, 78.7]	11.9	26	0
27	MARS	89.5	73.5 [72.8, 74.3]	16.0	43	-16
28	MARS (June 21 2018)	89.2	76.2 [75.4, 76.9]	13.1	31	-3
29	MMIPN	88.9	76.6 [75.8, 77.3]	12.4	30	-1
30	MARS (May 9 2018)	88.9	75.5 [74.8, 76.3]	13.3	33	-3
31	Bert-Large+Adv. Train	88.6	76.8 [76.0, 77.5]	11.8	29	2
32	Reinforced Mnemonic Reader (ensemble model)	88.5	74.2 [73.4, 75.0]	14.3	40	-8
33	AttentionReader+ (ensemble)	88.2	75.5 [74.8, 76.3]	12.6	32	1
34	Reinforced Mnemonic Reader + A2D	88.1	73.2 [72.4, 73.9]	15.0	44	-10
35	Reinforced Mnemonic Reader + A2D + DA	88.1	73.6 [72.8, 74.3]	14.6	42	-7
36	BERT-COMPOUND-DSS	88.0	75.4 [74.7, 76.2]	12.6	34	2
37	BERT-COMPOUND	87.8	74.8 [74.0, 75.5]	13.0	36	1
38	HierAtt	87.8	77.4 [76.7, 78.1]	10.3	28	10
39	BERT-Multi-Finetune	87.7	77.7 [77.0, 78.4]	10.0	27	12
40	BiDAF + Self Attention + ELMo (ensemble)	87.4	74.5 [73.8, 75.3]	12.9	39	1
41	AVIQA+ (ensemble) (aviqa team)	87.3	74.7 [73.9, 75.5]	12.6	37	4
42	EAZI (ensemble)	86.9	74.2 [73.4, 75.0]	12.7	41	1
43	MEMEN+ (Ensemble)	86.8	74.6 [73.8, 75.4]	12.3	38	5
44	DNET (ensemble)	86.7	75.0 [74.2, 75.8]	11.7	35	9
45	BERT-INDEPENDENT	86.7	72.9 [72.1, 73.6]	13.8	45	0
46	Reinforced Mnemonic Reader	86.7	70.1 [69.3, 70.9]	16.5	57	-11
47	MDReader	86.0	72.0 [71.2, 72.8]	14.0	51	-4
48	BiDAF + Self Attention + ELMo	85.9	72.6 [71.8, 73.4]	13.3	47	1
49	BiDAF + Self-Attention + ELMo	85.8	72.7 [71.9, 73.4]	13.2	46	3
50	MDReader0	85.5	72.0 [71.2, 72.7]	13.6	52	-2
51	MEMEN+	85.5	72.4 [71.6, 73.2]	13.1	49	2
52	MEMEN	85.3	72.5 [71.7, 73.3]	12.9	48	4
53	Interactive AoA Reader (Ensemble)	85.3	66.6 [65.8, 67.4]	18.7	66	-13
54	EAZI	85.1	71.8 [71.0, 72.6]	13.3	53	1
55	AttentionReader+	84.9	70.8 [70.0, 71.6]	14.1	54	1
56	DNET	84.9	72.1 [71.3, 72.8]	12.9	50	6
57	MARS (Jan 23)	84.7	69.4 [68.6, 70.2]	15.3	59	-2
58	FRC	84.6	69.5 [68.7, 70.3]	15.1	58	0
59	RaSoR + TR + LM	84.2	70.6 [69.8, 71.5]	13.5	56	3
60	gqa	83.9	66.3 [65.5, 67.2]	17.6	69	-9
61	FusionNet	83.9	69.1 [68.3, 69.9]	14.8	60	1
62	AVIQA-v2	83.3	70.7 [69.9, 71.5]	12.6	55	7
63	DCN+	83.1	66.5 [65.7, 67.3]	16.6	68	-5
64	Jenga	82.8	67.7 [66.8, 68.5]	15.2	64	0
65	two-attention-self-attention (ensemble)	82.7	68.4 [67.6, 69.2]	14.3	61	4
66	MEMEN (Ensemble, original model in paper)	82.7	66.6 [65.7, 67.4]	16.1	67	-1
67	ReasoNet (Ensemble)	82.6	67.3 [66.4, 68.1]	15.3	65	2
68	eeAttNet	82.5	68.3 [67.5, 69.1]	14.2	62	6
69	Mnemonic Reader (Ensemble)	82.4	66.0 [65.2, 66.8]	16.4	71	-2
70	Conductor-net	81.9	62.2 [61.3, 63.0]	19.8	81	-11
71	Interactive AoA Reader	81.9	62.7 [61.8, 63.5]	19.2	78	-7
72	BiDAF + Self Attention	81.0	65.8 [65.0, 66.6]	15.2	72	0
73	two-attention-self-attention	81.0	66.1 [65.3, 67.0]	14.9	70	3
74	AVIQA	80.5	67.7 [66.8, 68.5]	12.9	63	11
75	attention+self-attention	80.5	64.7 [63.8, 65.5]	15.8	74	1
76	Smarnet	80.2	59.0 [58.1, 59.8]	21.2	88	-12
77	Mnemonic Reader	80.1	62.2 [61.3, 63.0]	18.0	82	-5
78	MAMCN	79.9	64.7 [63.8, 65.5]	15.3	75	3
79	M-NET	79.8	62.2 [61.4, 63.1]	17.6	80	-1
80	JNet	79.8	61.0 [60.1, 61.8]	18.8	83	-3
81	Ruminating Reader	79.5	64.2 [63.4, 65.1]	15.2	76	5
82	ReasoNet	79.4	62.5 [61.7, 63.4]	16.8	79	3

Continued on next page

Reddit F1 Score Summary						
Rank	Name	SQuAD	Reddit	Gap	New Rank	Δ Rank
83	SimpleBaseline	78.2	63.2 [62.3, 64.0]	15.1	77	6
84	AllenNLP BiDAF	77.2	59.6 [58.7, 60.4]	17.6	87	-3
85	Match-LSTM w/ Ans-Ptr (Ensemble)	77.0	64.8 [64.0, 65.6]	12.2	73	12
86	Iterative Co-Attention Network	76.8	60.3 [59.5, 61.2]	16.4	86	0
87	BIDAF-INDEPENDENT-DSS	76.3	57.8 [57.0, 58.7]	18.5	90	-3
88	BIDAF-INDEPENDENT	74.6	58.5 [57.6, 59.3]	16.1	89	-1
89	BIDAF-COMPOUND	74.6	57.5 [56.6, 58.4]	17.0	91	-2
90	Match-LSTM w/ Bi-Ans-Ptr Boundary	73.7	60.4 [59.6, 61.3]	13.3	85	5
91	OTF dict+spelling	73.1	53.8 [52.9, 54.6]	19.3	92	-1
92	OTF spelling	72.0	52.8 [51.9, 53.6]	19.3	93	-1
93	OTF spelling+lemma	72.0	51.4 [50.5, 52.3]	20.6	94	-1
94	RQA+IDR	71.4	60.7 [59.8, 61.5]	10.7	84	10
95	UQA	64.0	48.2 [47.3, 49.1]	15.8	96	-1
96	UnsupervisedQA V1	54.7	49.8 [48.9, 50.7]	4.9	95	1

Table 8: Comparison of model F1 scores on the original SQuAD test set and our Amazon test set. Rank refers to the relative ordering using the original SQuAD v1.1 F1 scores, new rank refers to the ordering using the new test set scores, and Δ rank is the relative difference in ranking. The confidence intervals are 95% Student's t-intervals. Unless noted, all models are `single models`.

Amazon F1 Score Summary						
Rank	Name	SQuAD	Amazon	Gap	New Rank	Δ Rank
-	Human-0	94.9 [93.8, 96.0]	92.6 [91.3, 93.9]	2.3	-	-
-	Human-1	94.9 [93.8, 96.0]	92.4 [91.1, 93.7]	2.5	-	-
-	Human-2	95.6 [94.5, 96.6]	91.2 [89.6, 92.8]	4.4	-	-
1	XLNet	95.1	81.7 [81.1, 82.2]	13.4	7	-6
2	XLNET-123	94.9	85.7 [85.1, 86.3]	9.2	3	-1
3	XLNET-123++	94.9	87.2 [86.7, 87.7]	7.7	2	1
4	Delphi	94.7	87.7 [87.2, 88.3]	6.9	1	3
5	SpanBERT	94.6	84.8 [84.2, 85.3]	9.9	4	1
6	BERT+WWM+MT	94.4	81.6 [81.0, 82.3]	12.8	8	-2
7	BERT-cased-whole-word	93.4	82.9 [82.2, 83.5]	10.6	5	2
8	Tuned BERT-1seq Large Cased	93.3	82.5 [81.9, 83.2]	10.8	6	2
9	InfoWord (large)	93.1	81.5 [80.8, 82.1]	11.6	9	0
10	BERT-Large Baseline	92.7	80.8 [80.2, 81.5]	11.9	10	0
11	BERT+MT	92.6	80.2 [79.5, 80.8]	12.5	13	-2
12	Tuned BERT Large Cased	92.6	80.3 [79.6, 81.0]	12.3	12	0
13	DPN	92.0	79.3 [78.6, 80.0]	12.7	18	-5
14	ST_bl	92.0	79.6 [79.0, 80.3]	12.3	16	-2
15	BERT-uncased	91.9	79.9 [79.3, 80.6]	12.0	15	0
16	EL-BERT	91.8	77.2 [76.4, 77.9]	14.7	24	-8
17	BISAN	91.8	79.2 [78.5, 79.9]	12.6	19	-2
18	BERT+Sparse-Transformer	91.6	80.0 [79.3, 80.6]	11.6	14	4
19	InfoWord (base)	91.4	78.0 [77.3, 78.7]	13.4	23	-4
20	InfoWord-Base	91.4	78.0 [77.3, 78.7]	13.3	22	-2
21	InfoWord BERT baseline (large)	91.3	80.8 [80.2, 81.5]	10.5	11	10
22	Original BERT Large Cased	91.3	79.4 [78.7, 80.0]	11.9	17	5
23	Common-sense Governed BERT-123 (May 8 2019)	91.1	79.0 [78.3, 79.7]	12.1	20	3
24	InfoWord BERT baseline (base)	90.9	77.1 [76.4, 77.8]	13.8	25	-1
25	Common-sense Governed BERT-123	90.6	78.5 [77.8, 79.2]	12.1	21	4
26	MARS (ensemble, June 20 2018)	89.8	73.5 [72.8, 74.3]	16.2	32	-6
27	MARS	89.5	68.6 [67.8, 69.4]	20.9	54	-27
28	MARS (June 21 2018)	89.2	72.0 [71.2, 72.7]	17.3	34	-6
29	MMIPN	88.9	75.0 [74.3, 75.7]	14.0	29	0

Continued on next page

Amazon F1 Score Summary

Rank	Name	SQuAD	Amazon	Gap	New Rank	Δ Rank
30	MARS (May 9 2018)	88.9	71.4 [70.6, 72.1]	17.5	37	-7
31	Bert-Large+Adv. Train	88.6	76.0 [75.3, 76.7]	12.6	26	5
32	Reinforced Mnemonic Reader (ensemble model)	88.5	70.1 [69.3, 70.9]	18.4	44	-12
33	AttentionReader+ (ensemble)	88.2	71.3 [70.5, 72.1]	16.8	38	-5
34	Reinforced Mnemonic Reader + A2D	88.1	69.4 [68.6, 70.2]	18.7	46	-12
35	Reinforced Mnemonic Reader + A2D + DA	88.1	70.0 [69.2, 70.8]	18.1	45	-10
36	BERT-COMPOUND-DSS	88.0	74.2 [73.5, 75.0]	13.8	30	6
37	BERT-COMPOUND	87.8	73.6 [72.8, 74.3]	14.2	31	6
38	HierAtt	87.8	75.7 [75.0, 76.4]	12.1	27	11
39	BERT-Multi-Finetune	87.7	75.4 [74.7, 76.1]	12.3	28	11
40	BiDAF + Self Attention + ELMo (ensemble)	87.4	70.8 [70.0, 71.6]	16.7	39	1
41	AVIQA+ (ensemble) (aviqa team)	87.3	72.1 [71.3, 72.8]	15.2	33	8
42	EAZI (ensemble)	86.9	70.6 [69.8, 71.4]	16.3	40	2
43	EAZI+ (ensemble)	86.9	70.6 [69.8, 71.4]	16.3	41	2
44	MEMEN+ (Ensemble)	86.8	70.3 [69.6, 71.1]	16.5	42	2
45	DNET (ensemble)	86.7	72.0 [71.2, 72.7]	14.8	35	10
46	BERT-INDEPENDENT	86.7	71.8 [71.0, 72.5]	14.9	36	10
47	Reinforced Mnemonic Reader	86.7	66.6 [65.8, 67.4]	20.1	62	-15
48	FusionNet (ensemble)	86.0	68.6 [67.8, 69.4]	17.4	55	-7
49	MDReader	86.0	67.7 [66.9, 68.5]	18.3	57	-8
50	BiDAF + Self Attention + ELMo	85.9	69.2 [68.3, 70.0]	16.7	48	2
51	BiDAF + Self-Attention + ELMo	85.8	69.1 [68.3, 69.9]	16.7	49	2
52	MDReader0	85.5	67.7 [66.9, 68.5]	17.8	58	-6
53	BiDAF++ with pair2vec	85.5	69.3 [68.5, 70.1]	16.2	47	6
54	Conductor-net (Ensemble)	85.5	62.3 [61.5, 63.2]	23.2	76	-22
55	MEMEN+	85.5	68.7 [67.9, 69.5]	16.8	52	3
56	aviqa-v2 (ensemble)	85.5	70.3 [69.5, 71.1]	15.2	43	13
57	MEMEN	85.3	69.0 [68.2, 69.8]	16.4	50	7
58	Interactive AoA Reader (Ensemble)	85.3	63.0 [62.2, 63.9]	22.3	72	-14
59	EAZI	85.1	68.2 [67.4, 69.1]	16.9	56	3
60	AttentionReader+	84.9	66.6 [65.7, 67.4]	18.4	63	-3
61	DNET	84.9	69.0 [68.2, 69.8]	15.9	51	10
62	BiDAF++	84.9	68.7 [67.9, 69.5]	16.2	53	9
63	MARS (Jan 23)	84.7	64.3 [63.5, 65.2]	20.4	67	-4
64	FRC	84.6	64.2 [63.3, 65.0]	20.4	69	-5
65	Jenga (ensemble)	84.5	66.7 [65.9, 67.5]	17.8	60	5
66	RaSoR + TR + LM	84.2	66.6 [65.8, 67.4]	17.5	61	5
67	gqa	83.9	63.0 [62.1, 63.9]	20.9	73	-6
68	FusionNet	83.9	66.0 [65.2, 66.9]	17.9	64	4
69	Smarnet (Ensemble)	83.5	62.2 [61.3, 63.0]	21.3	77	-8
70	AVIQA-v2	83.3	67.7 [66.9, 68.5]	15.6	59	11
71	DCN+	83.1	62.9 [62.0, 63.7]	20.2	74	-3
72	Jenga	82.8	64.1 [63.3, 65.0]	18.7	70	2
73	Mixed model (ensemble)	82.8	62.2 [61.3, 63.0]	20.6	78	-5
74	two-attention-self-attention (ensemble)	82.7	63.6 [62.7, 64.4]	19.1	71	3
75	MEMEN (Ensemble, original model in paper)	82.7	61.6 [60.7, 62.4]	21.1	79	-4
76	ReasoNet (Ensemble)	82.6	62.7 [61.8, 63.5]	19.9	75	1
77	eeAttNet	82.5	65.5 [64.7, 66.3]	17.0	65	12
78	Mnemonic Reader (Ensemble)	82.4	60.1 [59.3, 61.0]	22.3	83	-5
79	Conductor-net	81.9	59.9 [59.1, 60.7]	22.0	86	-7
80	Interactive AoA Reader	81.9	60.1 [59.2, 60.9]	21.9	85	-5
81	Jenga	81.8	64.9 [64.1, 65.7]	16.9	66	15
82	SSAE (Ensemble)	81.7	59.9 [59.0, 60.7]	21.8	87	-5
83	JNet (Ensemble)	81.5	58.8 [58.0, 59.7]	22.7	91	-8
84	BiDAF + Self Attention	81.0	61.2 [60.4, 62.1]	19.8	81	3
85	two-attention-self-attention	81.0	61.5 [60.7, 62.4]	19.5	80	5
86	AVIQA	80.5	64.2 [63.4, 65.0]	16.4	68	18
87	attention+self-attention	80.5	60.1 [59.2, 60.9]	20.4	84	3
88	Smarnet	80.2	56.9 [56.0, 57.7]	23.3	98	-10
89	Mnemonic Reader	80.1	56.9 [56.0, 57.7]	23.3	97	-8

Continued on next page

Amazon F1 Score Summary						
Rank	Name	SQuAD	Amazon	Gap	New Rank	Δ Rank
90	MAMCN	79.9	60.1 [59.3, 61.0]	19.8	82	8
91	M-NET	79.8	57.8 [57.0, 58.7]	22.0	92	-1
92	JNet	79.8	56.7 [55.8, 57.5]	23.1	99	-7
93	Ruminating Reader	79.5	58.9 [58.0, 59.8]	20.6	90	3
94	ReasoNet	79.4	57.8 [56.9, 58.6]	21.6	94	0
95	RaSoR	78.7	57.6 [56.8, 58.5]	21.1	96	-1
96	SimpleBaseline	78.2	57.8 [56.9, 58.6]	20.5	93	3
97	PQMN	77.8	57.6 [56.8, 58.5]	20.1	95	2
98	AllenNLP BiDAF	77.2	56.2 [55.3, 57.0]	21.0	100	-2
99	Match-LSTM w/ Ans-Ptr (Ensemble)	77.0	59.0 [58.2, 59.9]	18.0	89	10
100	Iterative Co-Attention Network	76.8	54.7 [53.8, 55.5]	22.1	103	-3
101	BIDAF-COMPOUND-DSS	76.4	54.7 [53.9, 55.6]	21.7	102	-1
102	BIDAF-INDEPENDENT-DSS	76.3	54.7 [53.8, 55.5]	21.7	104	-2
103	BIDAF-INDEPENDENT	74.6	54.4 [53.6, 55.3]	20.2	105	-2
104	BIDAF-COMPOUND	74.6	53.7 [52.8, 54.5]	20.9	106	-2
105	Match-LSTM w/ Bi-Ans-Ptr Boundary	73.7	55.6 [54.7, 56.4]	18.2	101	4
106	OTF dict+spelling	73.1	49.4 [48.5, 50.3]	23.7	108	-2
107	OTF spelling	72.0	47.0 [46.1, 47.9]	25.0	109	-2
108	OTF spelling+lemma	72.0	46.5 [45.6, 47.3]	25.5	110	-2
109	RQA+IDR	71.4	59.5 [58.7, 60.4]	11.9	88	21
110	Dynamic Chunk Reader	71.0	51.7 [50.9, 52.6]	19.2	107	3
111	UQA	64.0	46.0 [45.1, 46.9]	18.0	111	0
112	UnsupervisedQA V1	54.7	45.0 [44.1, 45.8]	9.8	112	0

Table 9: Comparison of model EM scores on the original SQuAD test set and our new Wikipedia test set. Rank refers to the relative ordering using the original SQuAD v1.1 EM scores, new rank refers to the ordering using the new test set scores, and Δ rank is the relative difference in ranking. The confidence intervals are 95% Clopper-Pearson intervals. Unless noted, all models are `single models`.

New Wiki EM Score Summary						
Rank	Name	SQuAD	New Wiki	Gap	New Rank	Δ Rank
-	Human-0	89.1 [87.1, 91.0]	82.6 [80.0, 85.0]	6.6	-	-
-	Human-1	88.7 [86.6, 90.6]	83.2 [80.6, 85.6]	5.5	-	-
-	Human-2	90.5 [88.5, 92.2]	85.4 [82.9, 87.6]	5.1	-	-
1	XLNet	89.9	84.2 [83.3, 85.0]	5.7	1	0
2	XLNET-123++	89.9	83.4 [82.6, 84.2]	6.4	5	-3
3	XLNET-123	89.6	83.5 [82.7, 84.3]	6.1	3	0
4	Delphi	89.6	84.0 [83.1, 84.8]	5.6	2	2
5	SpanBERT	88.8	83.5 [82.7, 84.3]	5.3	4	1
6	BERT+WWM+MT	88.7	83.2 [82.3, 84.0]	5.5	6	0
7	Tuned BERT-1seq Large Cased	87.5	82.1 [81.3, 83.0]	5.3	9	-2
8	InfoWord (large)	87.3	82.2 [81.4, 83.0]	5.1	8	0
9	BERT-cased-whole-word	87.1	82.3 [81.5, 83.2]	4.8	7	2
10	BERT-Large Baseline	86.6	81.8 [80.9, 82.6]	4.9	11	-1
11	Tuned BERT Large Cased	86.5	81.4 [80.5, 82.2]	5.1	12	-1
12	BERT+MT	86.5	80.9 [80.0, 81.8]	5.6	13	-1
13	ST_bl	85.4	79.9 [79.0, 80.7]	5.6	16	-3
14	EL-BERT	85.3	80.1 [79.2, 81.0]	5.2	14	0
15	BISAN	85.3	79.6 [78.7, 80.5]	5.7	21	-6
16	BERT+Sparse-Transformer	85.1	79.8 [78.9, 80.7]	5.3	17	-1
17	DPN	85.0	79.7 [78.8, 80.6]	5.3	20	-3
18	BERT-uncased	84.9	79.2 [78.3, 80.1]	5.7	23	-5
19	InfoWord-Base	84.7	79.8 [78.9, 80.7]	4.9	19	0
20	InfoWord (base)	84.7	79.8 [78.9, 80.7]	4.9	18	2

Continued on next page

New Wiki EM Score Summary

Rank	Name	SQuAD	New Wiki	Gap	New Rank	Δ Rank
21	InfoWord BERT baseline (base)	84.4	79.3 [78.4, 80.2]	5.2	22	-1
22	Original BERT Large Cased	84.3	80.1 [79.2, 80.9]	4.3	15	7
23	InfoWord BERT baseline (large)	84.3	81.8 [80.9, 82.6]	2.5	10	13
24	MARS (ensemble, June 20 2018)	84.0	79.1 [78.2, 80.0]	4.9	24	0
25	Common-sense Governed BERT-123	83.9	78.3 [77.4, 79.2]	5.6	25	0
26	MARS	83.2	75.8 [74.8, 76.7]	7.4	35	-9
27	MARS (June 21 2018)	83.1	78.1 [77.2, 79.0]	5.0	27	0
28	Common-sense Governed BERT-123 (May 8 2019)	82.9	78.2 [77.3, 79.1]	4.7	26	2
29	MARS (May 9 2018)	82.6	77.2 [76.2, 78.1]	5.4	28	1
30	Reinforced Mnemonic Reader (ensemble model)	82.3	76.9 [76.0, 77.9]	5.3	29	1
31	AttentionReader+ (ensemble)	81.8	76.3 [75.3, 77.2]	5.5	31	0
32	MMIPN	81.6	76.8 [75.9, 77.8]	4.7	30	2
33	Reinforced Mnemonic Reader + A2D	81.5	75.2 [74.2, 76.1]	6.3	39	-6
34	Reinforced Mnemonic Reader + A2D + DA	81.4	76.1 [75.1, 77.0]	5.3	33	1
35	BERT-COMPOUND-DSS	81.0	75.8 [74.8, 76.7]	5.3	34	1
36	BiDAF + Self Attention + ELMo (ensemble)	81.0	75.1 [74.2, 76.1]	5.9	41	-5
37	AVIQA+ (ensemble) (aviqa team)	80.6	76.1 [75.1, 77.0]	4.5	32	5
38	EAZI (ensemble)	80.4	75.2 [74.2, 76.1]	5.3	40	-2
39	MEMEN+ (Ensemble)	80.4	75.1 [74.1, 76.0]	5.3	42	-3
40	DNET (ensemble)	80.2	75.6 [74.7, 76.6]	4.5	37	3
41	Bert-Large+Adv. Train	80.1	74.8 [73.9, 75.8]	5.2	43	-2
42	HierAtt	79.7	75.7 [74.8, 76.7]	4.0	36	6
43	Reinforced Mnemonic Reader	79.5	74.5 [73.5, 75.4]	5.1	44	-1
44	BERT-Multi-Finetune	79.5	75.5 [74.5, 76.4]	4.0	38	6
45	MDReader	79.0	73.5 [72.5, 74.5]	5.5	50	-5
46	BERT-INDEPENDENT	78.7	73.9 [72.9, 74.9]	4.8	46	0
47	BiDAF + Self Attention + ELMo	78.6	73.2 [72.2, 74.2]	5.3	53	-6
48	BiDAF + Self-Attention + ELMo	78.6	73.2 [72.3, 74.2]	5.3	52	-4
49	aviqa-v2 (ensemble)	78.5	74.4 [73.4, 75.3]	4.1	45	4
50	Conductor-net (Ensemble)	78.4	72.7 [71.7, 73.7]	5.7	57	-7
51	MEMEN	78.2	73.7 [72.7, 74.6]	4.6	48	3
52	MEMEN+	78.2	73.1 [72.1, 74.1]	5.1	54	-2
53	MDReader0	78.2	73.0 [72.0, 74.0]	5.2	56	-3
54	EAZI	78.0	73.5 [72.5, 74.4]	4.5	51	3
55	Interactive AoA Reader (Ensemble)	77.8	72.0 [71.0, 73.0]	5.9	63	-8
56	DNET	77.6	72.7 [71.7, 73.7]	4.9	58	-2
57	RaSoR + TR + LM	77.6	73.8 [72.8, 74.7]	3.8	47	10
58	BiDAF++	77.6	72.7 [71.7, 73.7]	4.9	59	-1
59	AttentionReader+	77.3	73.5 [72.5, 74.5]	3.8	49	10
60	Jenga (ensemble)	77.2	72.7 [71.7, 73.7]	4.6	60	0
61	gqa	77.1	73.1 [72.1, 74.0]	4.0	55	6
62	MARS (Jan 23)	76.9	72.2 [71.2, 73.2]	4.6	62	0
63	FRC	76.2	72.3 [71.3, 73.3]	3.9	61	2
64	FusionNet	76.0	71.3 [70.3, 72.3]	4.6	65	-1
65	AVIQA-v2	75.9	71.2 [70.1, 72.1]	4.8	68	-3
66	MEMEN (Ensemble, original model in paper)	75.4	70.9 [69.9, 71.9]	4.5	70	-4
67	Mixed model (ensemble)	75.3	71.0 [70.0, 72.0]	4.3	69	-2
68	two-attention-self-attention (ensemble)	75.2	71.2 [70.2, 72.2]	4.0	67	1
69	DCN+	75.1	71.4 [70.4, 72.4]	3.7	64	5
70	ReasoNet (Ensemble)	75.0	70.6 [69.6, 71.6]	4.4	71	-1
71	eeAttNet	74.6	71.2 [70.2, 72.2]	3.4	66	5
72	Jenga	74.4	68.7 [67.6, 69.7]	5.7	75	-3
73	Mnemonic Reader (Ensemble)	74.3	70.6 [69.6, 71.6]	3.6	72	1
74	Interactive AoA Reader	73.6	68.6 [67.6, 69.7]	5.0	76	-2
75	Jenga	73.3	68.4 [67.4, 69.5]	4.9	78	-3
76	Conductor-net	73.2	69.3 [68.3, 70.3]	4.0	74	2
77	two-attention-self-attention	72.6	68.6 [67.5, 69.6]	4.0	77	0
78	AVIQA	72.5	69.7 [68.7, 70.7]	2.8	73	5
79	BiDAF + Self Attention	72.1	68.4 [67.4, 69.4]	3.7	79	0
80	attention+self-attention	71.7	67.9 [66.8, 68.9]	3.8	80	0

Continued on next page

New Wiki EM Score Summary						
Rank	Name	SQuAD	New Wiki	Gap	New Rank	Δ Rank
81	Smarnet	71.4	67.1 [66.0, 68.1]	4.3	83	-2
82	M-NET	71.0	66.9 [65.9, 67.9]	4.1	85	-3
83	Mnemonic Reader	71.0	67.5 [66.5, 68.6]	3.4	82	1
84	MAMCN	71.0	67.8 [66.7, 68.8]	3.2	81	3
85	RaSoR	70.8	66.9 [65.8, 67.9]	4.0	87	-2
86	Ruminating Reader	70.6	67.0 [66.0, 68.1]	3.6	84	2
87	JNet	70.6	66.9 [65.9, 67.9]	3.7	86	1
88	ReasoNet	70.6	66.3 [65.2, 67.3]	4.3	88	0
89	SimpleBaseline	69.6	65.6 [64.5, 66.6]	4.0	89	0
90	Match-LSTM w/ Ans-Ptr (Ensemble)	67.9	64.8 [63.7, 65.8]	3.1	90	0
91	AllenNLP BiDAF	67.6	64.7 [63.6, 65.7]	3.0	91	0
92	BIDAF-COMPOUND-DSS	67.5	64.1 [63.1, 65.2]	3.4	93	-1
93	Iterative Co-Attention Network	67.5	64.7 [63.6, 65.7]	2.9	92	1
94	BIDAF-INDEPENDENT-DSS	66.5	63.9 [62.8, 64.9]	2.6	94	0
95	BIDAF-COMPOUND	65.2	63.1 [62.0, 64.2]	2.0	95	0
96	BIDAF-INDEPENDENT	64.9	62.6 [61.5, 63.7]	2.3	96	0
97	Match-LSTM w/ Bi-Ans-Ptr Boundary	64.7	61.1 [60.0, 62.2]	3.6	98	-1
98	OTF dict+spelling	64.1	61.3 [60.3, 62.4]	2.7	97	1
99	OTF spelling	62.9	60.6 [59.5, 61.6]	2.3	99	0
100	OTF spelling+lemma	62.6	60.1 [59.0, 61.2]	2.5	100	0
101	Dynamic Chunk Reader	62.5	59.3 [58.2, 60.4]	3.2	101	0
102	RQA+IDR	61.1	54.7 [53.6, 55.8]	6.5	102	0
103	UQA	53.7	48.8 [47.7, 49.9]	4.9	103	0
104	UnsupervisedQA V1	44.2	40.4 [39.3, 41.5]	3.8	104	0

Table 10: Comparison of model EM scores on the original SQuAD test set and our New York Times test set. Rank refers to the relative ordering using the original SQuAD v1.1 EM scores, new rank refers to the ordering using the new test set scores, and Δ rank is the relative difference in ranking. The confidence intervals are 95% Clopper-Pearson intervals. Unless noted, all models are `single models`.

NYT EM Score Summary						
Rank	Name	SQuAD	NYT	Gap	New Rank	Δ Rank
-	Human-0	89.1 [87.1, 91.0]	86.0 [83.6, 88.1]	3.2	-	-
-	Human-1	88.7 [86.6, 90.6]	88.5 [86.3, 90.5]	0.2	-	-
-	Human-2	90.5 [88.5, 92.2]	85.8 [83.4, 87.9]	4.7	-	-
1	XLNet	89.9	50.9 [50.0, 51.9]	38.9	102	-101
2	XLNET-123++	89.9	85.9 [85.2, 86.6]	3.9	3	-1
3	XLNET-123	89.6	86.0 [85.3, 86.7]	3.7	2	1
4	Delphi	89.6	86.9 [86.3, 87.6]	2.7	1	3
5	SpanBERT	88.8	85.3 [84.6, 86.0]	3.5	4	1
6	BERT+WWM+MT	88.7	79.4 [78.6, 80.2]	9.2	21	-15
7	Tuned BERT-1seq Large Cased	87.5	83.5 [82.8, 84.2]	4.0	7	0
8	InfoWord (large)	87.3	83.8 [83.1, 84.5]	3.5	5	3
9	BERT-cased-whole-word	87.1	83.8 [83.1, 84.5]	3.3	6	3
10	BERT-Large Baseline	86.6	83.0 [82.2, 83.7]	3.7	9	1
11	Tuned BERT Large Cased	86.5	82.8 [82.1, 83.6]	3.7	10	1
12	BERT+MT	86.5	78.6 [77.8, 79.4]	7.8	25	-13
13	ST_bl	85.4	80.5 [79.7, 81.3]	4.9	13	0
14	EL-BERT	85.3	80.3 [79.5, 81.0]	5.1	16	-2
15	BISAN	85.3	80.1 [79.3, 80.8]	5.3	19	-4
16	BERT+Sparse-Transformer	85.1	79.7 [78.9, 80.5]	5.4	20	-4
17	DPN	85.0	80.2 [79.4, 80.9]	4.8	18	-1
18	BERT-uncased	84.9	80.4 [79.6, 81.1]	4.5	15	3
19	InfoWord-Base	84.7	80.6 [79.8, 81.4]	4.1	11	8

Continued on next page

NYT EM Score Summary

Rank	Name	SQuAD	NYT	Gap	New Rank	Δ Rank
20	InfoWord (base)	84.7	80.6 [79.8, 81.4]	4.1	12	8
21	InfoWord BERT baseline (base)	84.4	80.5 [79.7, 81.3]	3.9	14	7
22	Original BERT Large Cased	84.3	80.2 [79.4, 81.0]	4.1	17	5
23	InfoWord BERT baseline (large)	84.3	83.0 [82.2, 83.7]	1.3	8	15
24	MARS (ensemble, June 20 2018)	84.0	78.7 [77.9, 79.5]	5.3	24	0
25	Common-sense Governed BERT-123	83.9	79.1 [78.3, 79.9]	4.8	22	3
26	MARS	83.2	74.5 [73.6, 75.3]	8.7	42	-16
27	MARS (June 21 2018)	83.1	77.1 [76.3, 77.9]	6.0	27	0
28	Common-sense Governed BERT-123 (May 8 2019)	82.9	78.7 [77.9, 79.5]	4.2	23	5
29	MARS (May 9 2018)	82.6	76.8 [75.9, 77.6]	5.8	29	0
30	Reinforced Mnemonic Reader (ensemble model)	82.3	75.9 [75.1, 76.7]	6.4	32	-2
31	AttentionReader+ (ensemble)	81.8	76.4 [75.5, 77.2]	5.4	30	1
32	MMIPN	81.6	76.9 [76.1, 77.7]	4.7	28	4
33	Reinforced Mnemonic Reader + A2D	81.5	75.5 [74.6, 76.3]	6.0	37	-4
34	Reinforced Mnemonic Reader + A2D + DA	81.4	74.9 [74.0, 75.7]	6.5	40	-6
35	BERT-COMPOUND-DSS	81.0	76.3 [75.4, 77.1]	4.8	31	4
36	BiDAF + Self Attention + ELMo (ensemble)	81.0	75.9 [75.0, 76.7]	5.1	33	3
37	BERT-COMPOUND	80.7	75.8 [75.0, 76.7]	4.9	34	3
38	AVIQA+ (ensemble) (aviqa team)	80.6	75.3 [74.5, 76.2]	5.3	39	-1
39	EAZI (ensemble)	80.4	75.8 [74.9, 76.6]	4.6	35	4
40	MEMEN+ (Ensemble)	80.4	72.4 [71.5, 73.2]	8.0	53	-13
41	DNET (ensemble)	80.2	75.4 [74.5, 76.2]	4.8	38	3
42	Bert-Large+Adv. Train	80.1	74.5 [73.6, 75.3]	5.6	41	1
43	HierAtt	79.7	75.5 [74.7, 76.4]	4.2	36	7
44	Reinforced Mnemonic Reader	79.5	73.8 [72.9, 74.6]	5.8	47	-3
45	BERT-Multi-Finetune	79.5	77.1 [76.3, 78.0]	2.4	26	19
46	MDReader	79.0	74.3 [73.4, 75.1]	4.8	43	3
47	BERT-INDEPENDENT	78.7	73.5 [72.6, 74.4]	5.2	48	-1
48	BiDAF + Self Attention + ELMo	78.6	74.0 [73.1, 74.9]	4.6	45	3
49	BiDAF + Self-Attention + ELMo	78.6	74.0 [73.2, 74.9]	4.5	44	5
50	Conductor-net (Ensemble)	78.4	70.3 [69.4, 71.2]	8.1	60	-10
51	MEMEN	78.2	70.6 [69.7, 71.5]	7.6	58	-7
52	MEMEN+	78.2	70.9 [70.0, 71.8]	7.3	56	-4
53	MDReader0	78.2	73.9 [73.0, 74.7]	4.3	46	7
54	EAZI	78.0	72.9 [72.1, 73.8]	5.0	49	5
55	Interactive AoA Reader (Ensemble)	77.8	69.2 [68.3, 70.1]	8.6	66	-11
56	DNET	77.6	72.8 [71.9, 73.6]	4.9	51	5
57	RaSoR + TR + LM	77.6	72.8 [72.0, 73.7]	4.7	50	7
58	BiDAF++	77.6	72.4 [71.5, 73.3]	5.2	52	6
59	AttentionReader+	77.3	71.8 [70.9, 72.7]	5.6	54	5
60	Jenga (ensemble)	77.2	70.7 [69.8, 71.6]	6.5	57	3
61	gqa	77.1	70.4 [69.5, 71.3]	6.6	59	2
62	MARS (Jan 23)	76.9	70.1 [69.2, 71.0]	6.7	62	0
63	FRC	76.2	70.1 [69.2, 71.0]	6.1	63	0
64	FusionNet	76.0	69.2 [68.3, 70.1]	6.7	67	-3
65	AVIQA-v2	75.9	71.4 [70.5, 72.3]	4.5	55	10
66	MEMEN (Ensemble, original model in paper)	75.4	68.9 [67.9, 69.8]	6.5	70	-4
67	Mixed model (ensemble)	75.3	69.4 [68.5, 70.3]	5.9	64	3
68	two-attention-self-attention (ensemble)	75.2	70.3 [69.4, 71.2]	5.0	61	7
69	DCN+	75.1	66.9 [65.9, 67.8]	8.2	76	-7
70	ReasoNet (Ensemble)	75.0	69.3 [68.3, 70.2]	5.8	65	5
71	eeAttNet	74.6	69.2 [68.3, 70.1]	5.4	68	3
72	Jenga	74.4	66.8 [65.9, 67.8]	7.5	77	-5
73	Mnemonic Reader (Ensemble)	74.3	69.2 [68.3, 70.1]	5.1	69	4
74	Interactive AoA Reader	73.6	65.2 [64.2, 66.1]	8.5	85	-11
75	Jenga	73.3	66.7 [65.8, 67.6]	6.6	78	-3
76	Conductor-net	73.2	67.6 [66.7, 68.5]	5.6	73	3
77	two-attention-self-attention	72.6	68.1 [67.2, 69.0]	4.5	72	5
78	AVIQA	72.5	68.8 [67.9, 69.7]	3.7	71	7
79	BiDAF + Self Attention	72.1	67.3 [66.4, 68.2]	4.8	74	5

Continued on next page

NYT EM Score Summary						
Rank	Name	SQuAD	NYT	Gap	New Rank	Δ Rank
80	attention+self-attention	71.7	66.3 [65.3, 67.2]	5.4	80	0
81	Smarnet	71.4	64.5 [63.5, 65.4]	6.9	87	-6
82	M-NET	71.0	65.4 [64.4, 66.3]	5.6	83	-1
83	Mnemonic Reader	71.0	65.6 [64.7, 66.5]	5.4	82	1
84	MAMCN	71.0	67.0 [66.1, 67.9]	4.0	75	9
85	RaSoR	70.8	65.3 [64.4, 66.2]	5.6	84	1
86	Ruminating Reader	70.6	66.3 [65.4, 67.2]	4.3	79	7
87	JNet	70.6	64.8 [63.8, 65.7]	5.8	86	1
88	ReasoNet	70.6	64.1 [63.2, 65.0]	6.5	88	0
89	SimpleBaseline	69.6	66.1 [65.2, 67.1]	3.5	81	8
90	Match-LSTM w/ Ans-Ptr (Ensemble)	67.9	60.9 [59.9, 61.8]	7.0	93	-3
91	AllenNLP BiDAF	67.6	63.3 [62.4, 64.3]	4.3	89	2
92	BIDAF-COMPOUND-DSS	67.5	62.6 [61.6, 63.5]	5.0	91	1
93	Iterative Co-Attention Network	67.5	62.8 [61.8, 63.7]	4.7	90	3
94	BIDAF-INDEPENDENT-DSS	66.5	61.7 [60.7, 62.6]	4.8	92	2
95	BIDAF-COMPOUND	65.2	60.8 [59.8, 61.7]	4.4	94	1
96	BIDAF-INDEPENDENT	64.9	60.4 [59.4, 61.3]	4.6	95	1
97	Match-LSTM w/ Bi-Ans-Ptr Boundary	64.7	57.5 [56.5, 58.4]	7.3	97	0
98	OTF dict+spelling	64.1	57.2 [56.3, 58.2]	6.9	98	0
99	OTF spelling	62.9	56.2 [55.2, 57.2]	6.7	100	-1
100	OTF spelling+lemma	62.6	55.3 [54.3, 56.3]	7.3	101	-1
101	Dynamic Chunk Reader	62.5	59.1 [58.1, 60.0]	3.4	96	5
102	RQA+IDR	61.1	57.2 [56.2, 58.2]	3.9	99	3
103	UQA	53.7	49.5 [48.5, 50.5]	4.2	103	0
104	UnsupervisedQA V1	44.2	40.6 [39.6, 41.6]	3.6	104	0

Table 11: Comparison of model EM scores on the original SQuAD test set and our Reddit test set. Rank refers to the relative ordering using the original SQuAD v1.1 EM scores, new rank refers to the ordering using the new test set scores, and Δ rank is the relative difference in ranking. The confidence intervals are 95% Clopper-Pearson intervals. Unless noted, all models are `single models`.

Reddit EM Score Summary						
Rank	Name	SQuAD	Reddit	Gap	New Rank	Δ Rank
-	Human-0	89.1 [87.1, 91.0]	80.1 [77.2, 82.7]	9.1	-	-
-	Human-1	88.7 [86.6, 90.6]	80.7 [77.8, 83.3]	8.0	-	-
-	Human-2	90.5 [88.5, 92.2]	81.0 [78.2, 83.6]	9.4	-	-
1	XLNet	89.9	43.1 [42.1, 44.1]	46.8	89	-88
2	XLNET-123++	89.9	70.8 [69.9, 71.7]	19.0	5	-3
3	XLNET-123	89.6	73.8 [72.9, 74.7]	15.8	2	1
4	Delphi	89.6	77.9 [77.1, 78.8]	11.7	1	3
5	SpanBERT	88.8	73.3 [72.4, 74.2]	15.6	3	2
6	BERT+WWM+MT	88.7	69.2 [68.3, 70.1]	19.4	9	-3
7	Tuned BERT-1seq Large Cased	87.5	70.8 [69.9, 71.7]	16.7	6	1
8	InfoWord (large)	87.3	70.5 [69.6, 71.4]	16.8	7	1
9	BERT-cased-whole-word	87.1	72.0 [71.1, 72.9]	15.1	4	5
10	BERT-Large Baseline	86.6	68.9 [68.0, 69.8]	17.7	10	0
11	Tuned BERT Large Cased	86.5	69.5 [68.6, 70.4]	17.0	8	3
12	BERT+MT	86.5	68.4 [67.4, 69.3]	18.1	14	-2
13	ST_bl	85.4	68.7 [67.7, 69.6]	16.8	13	0
14	EL-BERT	85.3	65.6 [64.6, 66.5]	19.8	24	-10
15	BISAN	85.3	67.7 [66.8, 68.7]	17.6	16	-1
16	BERT+Sparse-Transformer	85.1	68.9 [67.9, 69.8]	16.2	12	4
17	DPN	85.0	67.2 [66.3, 68.2]	17.8	20	-3
18	BERT-uncased	84.9	67.2 [66.3, 68.2]	17.7	19	-1

Continued on next page

Reddit EM Score Summary

Rank	Name	SQuAD	Reddit	Gap	New Rank	Δ Rank
19	InfoWord-Base	84.7	65.8 [64.9, 66.8]	18.9	23	-4
20	InfoWord (base)	84.7	65.8 [64.9, 66.8]	18.9	22	-2
21	InfoWord BERT baseline (base)	84.4	65.4 [64.4, 66.3]	19.1	25	-4
22	Original BERT Large Cased	84.3	68.3 [67.3, 69.2]	16.1	15	7
23	InfoWord BERT baseline (large)	84.3	68.9 [68.0, 69.8]	15.4	11	12
24	MARS (ensemble, June 20 2018)	84.0	66.0 [65.0, 66.9]	18.0	21	3
25	Common-sense Governed BERT-123	83.9	67.7 [66.7, 68.6]	16.2	17	8
26	MARS	83.2	60.0 [59.0, 60.9]	23.2	45	-19
27	MARS (June 21 2018)	83.1	64.4 [63.5, 65.4]	18.7	27	0
28	Common-sense Governed BERT-123 (May 8 2019)	82.9	67.4 [66.4, 68.3]	15.6	18	10
29	MARS (May 9 2018)	82.6	63.0 [62.0, 63.9]	19.6	31	-2
30	Reinforced Mnemonic Reader (ensemble model)	82.3	62.3 [61.3, 63.3]	20.0	38	-8
31	AttentionReader+ (ensemble)	81.8	63.0 [62.0, 63.9]	18.8	32	-1
32	MMIPN	81.6	63.5 [62.6, 64.5]	18.0	29	3
33	Reinforced Mnemonic Reader + A2D	81.5	59.9 [58.9, 60.9]	21.6	46	-13
34	Reinforced Mnemonic Reader + A2D + DA	81.4	60.0 [59.0, 61.0]	21.4	42	-8
35	BERT-COMPOUND-DSS	81.0	63.8 [62.8, 64.8]	17.2	28	7
36	BiDAF + Self Attention + ELMo (ensemble)	81.0	62.3 [61.4, 63.3]	18.7	37	-1
37	BERT-COMPOUND	80.7	62.8 [61.8, 63.8]	17.9	34	3
38	AVIQA+ (ensemble) (aviqa team)	80.6	62.7 [61.8, 63.7]	17.9	35	3
39	EAZI (ensemble)	80.4	62.1 [61.1, 63.1]	18.3	40	-1
40	MEMEN+ (Ensemble)	80.4	62.5 [61.6, 63.5]	17.9	36	4
41	DNET (ensemble)	80.2	63.0 [62.0, 63.9]	17.2	33	8
42	Bert-Large+Adv. Train	80.1	62.2 [61.2, 63.1]	17.9	39	3
43	HierAtt	79.7	63.5 [62.6, 64.5]	16.2	30	13
44	Reinforced Mnemonic Reader	79.5	57.2 [56.2, 58.2]	22.4	56	-12
45	BERT-Multi-Finetune	79.5	65.2 [64.3, 66.2]	14.3	26	19
46	MDReader	79.0	59.2 [58.2, 60.2]	19.8	49	-3
47	BERT-INDEPENDENT	78.7	58.5 [57.5, 59.5]	20.1	53	-6
48	BiDAF + Self Attention + ELMo	78.6	60.0 [59.0, 61.0]	18.6	44	4
49	BiDAF + Self-Attention + ELMo	78.6	60.0 [59.0, 61.0]	18.6	43	6
50	MEMEN	78.2	60.0 [59.0, 61.0]	18.2	41	9
51	MEMEN+	78.2	59.9 [58.9, 60.8]	18.4	47	4
52	MDReader0	78.2	58.7 [57.7, 59.7]	19.4	50	2
53	EAZI	78.0	58.5 [57.6, 59.5]	19.4	52	1
54	Interactive AoA Reader (Ensemble)	77.8	53.0 [52.0, 54.0]	24.8	67	-13
55	DNET	77.6	59.2 [58.2, 60.2]	18.4	48	7
56	RaSoR + TR + LM	77.6	58.6 [57.6, 59.6]	19.0	51	5
57	AttentionReader+	77.3	57.4 [56.4, 58.4]	19.9	55	2
58	gqa	77.1	55.1 [54.1, 56.1]	22.0	59	-1
59	MARS (Jan 23)	76.9	54.3 [53.3, 55.3]	22.6	63	-4
60	FRC	76.2	55.1 [54.2, 56.1]	21.1	58	2
61	FusionNet	76.0	54.5 [53.5, 55.5]	21.5	62	-1
62	AVIQA-v2	75.9	58.3 [57.3, 59.2]	17.7	54	8
63	MEMEN (Ensemble, original model in paper)	75.4	53.5 [52.5, 54.5]	21.9	65	-2
64	two-attention-self-attention (ensemble)	75.2	55.1 [54.1, 56.1]	20.2	60	4
65	DCN+	75.1	53.8 [52.8, 54.8]	21.3	64	1
66	ReasoNet (Ensemble)	75.0	53.3 [52.3, 54.3]	21.8	66	0
67	eeAttNet	74.6	55.2 [54.2, 56.2]	19.4	57	10
68	Jenga	74.4	52.9 [51.9, 53.9]	21.5	68	0
69	Mnemonic Reader (Ensemble)	74.3	51.6 [50.6, 52.6]	22.7	70	-1
70	Interactive AoA Reader	73.6	48.0 [47.0, 49.0]	25.6	78	-8
71	Conductor-net	73.2	48.1 [47.1, 49.1]	25.1	77	-6
72	two-attention-self-attention	72.6	52.1 [51.1, 53.1]	20.5	69	3
73	AVIQA	72.5	54.8 [53.8, 55.8]	17.7	61	12
74	BiDAF + Self Attention	72.1	50.7 [49.7, 51.7]	21.4	71	3
75	attention+self-attention	71.7	50.6 [49.6, 51.6]	21.1	72	3
76	Smarnet	71.4	44.6 [43.6, 45.6]	26.8	86	-10
77	M-NET	71.0	47.1 [46.1, 48.1]	23.9	81	-4
78	Mnemonic Reader	71.0	46.9 [46.0, 47.9]	24.1	82	-4

Continued on next page

Reddit EM Score Summary						
Rank	Name	SQuAD	Reddit	Gap	New Rank	Δ Rank
79	MAMCN	71.0	50.3 [49.3, 51.3]	20.7	74	5
80	Ruminating Reader	70.6	50.0 [49.0, 51.0]	20.7	75	5
81	JNet	70.6	46.7 [45.7, 47.7]	23.9	83	-2
82	ReasoNet	70.6	47.4 [46.4, 48.4]	23.1	80	2
83	SimpleBaseline	69.6	49.1 [48.1, 50.1]	20.5	76	7
84	Match-LSTM w/ Ans-Ptr (Ensemble)	67.9	50.5 [49.5, 51.5]	17.4	73	11
85	AllenNLP BiDAF	67.6	44.4 [43.4, 45.4]	23.2	87	-2
86	Iterative Co-Attention Network	67.5	46.2 [45.2, 47.2]	21.3	85	1
87	BIDAF-INDEPENDENT-DSS	66.5	42.2 [41.2, 43.1]	24.4	91	-4
88	BIDAF-COMPOUND	65.2	42.7 [41.7, 43.7]	22.5	90	-2
89	BIDAF-INDEPENDENT	64.9	43.2 [42.2, 44.2]	21.7	88	1
90	Match-LSTM w/ Bi-Ans-Ptr Boundary	64.7	46.4 [45.4, 47.4]	18.4	84	6
91	OTF dict+spelling	64.1	39.5 [38.6, 40.5]	24.6	92	-1
92	OTF spelling	62.9	39.4 [38.5, 40.4]	23.4	93	-1
93	OTF spelling+lemma	62.6	37.8 [36.8, 38.8]	24.8	94	-1
94	RQA+IDR	61.1	48.0 [47.0, 49.0]	13.2	79	15
95	UQA	53.7	36.8 [35.8, 37.7]	16.9	96	-1
96	UnsupervisedQA V1	44.2	37.3 [36.3, 38.2]	6.9	95	1

Table 12: Comparison of model EM scores on the original SQuAD test set and our Amazon test set. Rank refers to the relative ordering using the original SQuAD v1.1 EM scores, new rank refers to the ordering using the new test set scores, and Δ rank is the relative difference in ranking. The confidence intervals are 95% Clopper-Pearson intervals. Unless noted, all models are `single models`.

Amazon EM Score Summary						
Rank	Name	SQuAD	Amazon	Gap	New Rank	Δ Rank
-	Human-0	89.1 [87.1, 91.0]	79.9 [77.1, 82.4]	9.3	-	-
-	Human-1	88.7 [86.6, 90.6]	81.1 [78.4, 83.6]	7.6	-	-
-	Human-2	90.5 [88.5, 92.2]	79.3 [76.5, 82.0]	11.2	-	-
1	XLNet	89.9	53.9 [52.9, 54.9]	36.0	53	-52
2	XLNET-123++	89.9	74.1 [73.2, 75.0]	15.7	2	0
3	XLNET-123	89.6	72.2 [71.3, 73.1]	17.5	3	0
4	Delphi	89.6	75.7 [74.8, 76.5]	13.9	1	3
5	SpanBERT	88.8	70.6 [69.7, 71.5]	18.3	4	1
6	BERT+WWM+MT	88.7	65.5 [64.5, 66.4]	23.2	13	-7
7	Tuned BERT-1seq Large Cased	87.5	69.4 [68.5, 70.3]	18.0	5	2
8	InfoWord (large)	87.3	67.5 [66.5, 68.4]	19.8	7	1
9	BERT-cased-whole-word	87.1	69.3 [68.4, 70.2]	17.8	6	3
10	BERT-Large Baseline	86.6	66.6 [65.6, 67.5]	20.1	8	2
11	Tuned BERT Large Cased	86.5	66.3 [65.3, 67.2]	20.3	10	1
12	BERT+MT	86.5	64.1 [63.1, 65.0]	22.4	18	-6
13	ST_bl	85.4	65.5 [64.6, 66.5]	19.9	12	1
14	EL-BERT	85.3	62.5 [61.5, 63.4]	22.8	23	-9
15	BISAN	85.3	65.0 [64.1, 66.0]	20.3	15	0
16	BERT+Sparse-Transformer	85.1	66.1 [65.2, 67.1]	19.0	11	5
17	DPN	85.0	63.9 [62.9, 64.8]	21.1	20	-3
18	BERT-uncased	84.9	64.6 [63.7, 65.6]	20.3	17	1
19	InfoWord-Base	84.7	63.1 [62.2, 64.1]	21.6	22	-3
20	InfoWord (base)	84.7	63.1 [62.2, 64.1]	21.6	21	-1
21	InfoWord BERT baseline (base)	84.4	62.3 [61.3, 63.3]	22.1	24	-3
22	Original BERT Large Cased	84.3	64.8 [63.9, 65.8]	19.5	16	6
23	InfoWord BERT baseline (large)	84.3	66.6 [65.6, 67.5]	17.7	9	14
24	MARS (ensemble, June 20 2018)	84.0	59.5 [58.5, 60.5]	24.5	31	-7
25	Common-sense Governed BERT-123	83.9	65.4 [64.5, 66.4]	18.5	14	11

Continued on next page

Amazon EM Score Summary

Rank	Name	SQuAD	Amazon	Gap	New Rank	Δ Rank
26	MARS	83.2	53.2 [52.2, 54.2]	30.0	57	-31
27	MARS (June 21 2018)	83.1	57.9 [56.9, 58.9]	25.2	34	-7
28	Common-sense Governed BERT-123 (May 8 2019)	82.9	64.0 [63.0, 64.9]	19.0	19	9
29	MARS (May 9 2018)	82.6	57.2 [56.2, 58.2]	25.4	35	-6
30	Reinforced Mnemonic Reader (ensemble model)	82.3	55.6 [54.6, 56.6]	26.7	43	-13
31	AttentionReader+ (ensemble)	81.8	56.9 [55.9, 57.8]	24.9	38	-7
32	MMIPN	81.6	60.6 [59.6, 61.5]	21.0	27	5
33	Reinforced Mnemonic Reader + A2D	81.5	54.4 [53.4, 55.4]	27.1	52	-19
34	Reinforced Mnemonic Reader + A2D + DA	81.4	54.8 [53.8, 55.7]	26.6	49	-15
35	BERT-COMPOUND-DSS	81.0	60.6 [59.6, 61.5]	20.5	26	9
36	BiDAF + Self Attention + ELMo (ensemble)	81.0	57.1 [56.1, 58.1]	23.9	36	0
37	BERT-COMPOUND	80.7	59.9 [58.9, 60.8]	20.8	28	9
38	AVIQA+ (ensemble) (aviqa team)	80.6	58.3 [57.3, 59.3]	22.3	32	6
39	EAZI (ensemble)	80.4	56.9 [55.9, 57.9]	23.5	37	2
40	EAZI+ (ensemble)	80.4	56.7 [55.7, 57.7]	23.7	40	0
41	MEMEN+ (Ensemble)	80.4	56.8 [55.8, 57.7]	23.6	39	2
42	DNET (ensemble)	80.2	58.2 [57.2, 59.1]	22.0	33	9
43	Bert-Large+Adv. Train	80.1	59.8 [58.8, 60.8]	20.3	29	14
44	HierAtt	79.7	59.7 [58.8, 60.7]	20.0	30	14
45	Reinforced Mnemonic Reader	79.5	51.4 [50.4, 52.3]	28.2	62	-17
46	BERT-Multi-Finetune	79.5	61.6 [60.6, 62.5]	18.0	25	21
47	MDReader	79.0	52.9 [52.0, 53.9]	26.1	58	-11
48	FusionNet (ensemble)	79.0	52.9 [51.9, 53.9]	26.1	60	-12
49	BERT-INDEPENDENT	78.7	56.1 [55.1, 57.1]	22.5	42	7
50	BiDAF + Self Attention + ELMo	78.6	55.0 [54.1, 56.0]	23.5	46	4
51	BiDAF + Self-Attention + ELMo	78.6	55.0 [54.1, 56.0]	23.5	45	6
52	aviqa-v2 (ensemble)	78.5	56.6 [55.6, 57.6]	21.9	41	11
53	Conductor-net (Ensemble)	78.4	46.5 [45.5, 47.5]	32.0	79	-26
54	MEMEN	78.2	54.9 [53.9, 55.9]	23.3	48	6
55	BiDAF++ with pair2vec	78.2	55.1 [54.1, 56.1]	23.1	44	11
56	MEMEN+	78.2	54.7 [53.7, 55.7]	23.5	50	6
57	MDReader0	78.2	52.9 [51.9, 53.9]	25.3	59	-2
58	EAZI	78.0	53.8 [52.8, 54.8]	24.2	55	3
59	Interactive AoA Reader (Ensemble)	77.8	48.2 [47.2, 49.2]	29.7	70	-11
60	DNET	77.6	55.0 [54.0, 55.9]	22.7	47	13
61	RaSoR + TR + LM	77.6	53.3 [52.4, 54.3]	24.2	56	5
62	BiDAF++	77.6	54.6 [53.6, 55.6]	23.0	51	11
63	AttentionReader+	77.3	51.3 [50.3, 52.3]	26.1	63	0
64	Jenga (ensemble)	77.2	52.0 [51.1, 53.0]	25.2	61	3
65	gqa	77.1	50.8 [49.8, 51.8]	26.3	64	1
66	MARS (Jan 23)	76.9	47.9 [46.9, 48.9]	29.0	72	-6
67	FRC	76.2	47.7 [46.7, 48.6]	28.6	73	-6
68	Smarnet (Ensemble)	76.0	46.8 [45.9, 47.8]	29.2	76	-8
69	FusionNet	76.0	49.5 [48.5, 50.5]	26.5	68	1
70	AVIQA-v2	75.9	53.8 [52.8, 54.8]	22.1	54	16
71	MEMEN (Ensemble, original model in paper)	75.4	46.8 [45.8, 47.8]	28.6	77	-6
72	Mixed model (ensemble)	75.3	46.7 [45.7, 47.7]	28.6	78	-6
73	two-attention-self-attention (ensemble)	75.2	48.4 [47.5, 49.4]	26.8	69	4
74	DCN+	75.1	47.6 [46.6, 48.6]	27.5	74	0
75	ReasoNet (Ensemble)	75.0	46.9 [45.9, 47.9]	28.1	75	0
76	eeAttNet	74.6	50.7 [49.7, 51.7]	23.9	65	11
77	Jenga	74.4	48.1 [47.1, 49.1]	26.2	71	6
78	Mnemonic Reader (Ensemble)	74.3	44.4 [43.4, 45.4]	29.9	84	-6
79	SSAE (Ensemble)	74.1	44.9 [43.9, 45.9]	29.2	83	-4
80	Interactive AoA Reader	73.6	43.8 [42.8, 44.8]	29.8	87	-7
81	Jenga	73.3	49.5 [48.6, 50.5]	23.8	67	14
82	Conductor-net	73.2	43.7 [42.7, 44.7]	29.5	88	-6
83	JNet (Ensemble)	73.0	42.5 [41.5, 43.5]	30.5	92	-9
84	two-attention-self-attention	72.6	45.9 [44.9, 46.9]	26.7	80	4
85	AVIQA	72.5	50.1 [49.1, 51.1]	22.4	66	19

Continued on next page

Amazon EM Score Summary						
Rank	Name	SQuAD	Amazon	Gap	New Rank	Δ Rank
86	BiDAF + Self Attention	72.1	45.2 [44.2, 46.2]	26.9	82	4
87	attention+self-attention	71.7	44.4 [43.4, 45.4]	27.3	85	2
88	Smarnet	71.4	41.1 [40.1, 42.1]	30.3	95	-7
89	M-NET	71.0	41.3 [40.4, 42.3]	29.7	94	-5
90	Mnemonic Reader	71.0	40.5 [39.6, 41.5]	30.4	98	-8
91	MAMCN	71.0	44.0 [43.0, 45.0]	27.0	86	5
92	RaSoR	70.8	42.8 [41.8, 43.7]	28.1	91	1
93	Ruminating Reader	70.6	43.5 [42.5, 44.4]	27.2	89	4
94	JNet	70.6	40.3 [39.4, 41.3]	30.3	99	-5
95	ReasoNet	70.6	40.9 [39.9, 41.8]	29.7	97	-2
96	SimpleBaseline	69.6	42.3 [41.3, 43.3]	27.3	93	3
97	PQMN	68.3	41.0 [40.0, 42.0]	27.3	96	1
98	Match-LSTM w/ Ans-Ptr (Ensemble)	67.9	43.5 [42.5, 44.4]	24.4	90	8
99	AllenNLP BiDAF	67.6	39.8 [38.8, 40.7]	27.9	101	-2
100	BIDAF-COMPOUND-DSS	67.5	39.3 [38.3, 40.3]	28.3	103	-3
101	Iterative Co-Attention Network	67.5	39.4 [38.4, 40.4]	28.1	102	-1
102	BIDAF-INDEPENDENT-DSS	66.5	37.9 [36.9, 38.9]	28.6	105	-3
103	BIDAF-COMPOUND	65.2	37.7 [36.8, 38.7]	27.4	107	-4
104	BIDAF-INDEPENDENT	64.9	38.7 [37.8, 39.7]	26.2	104	0
105	Match-LSTM w/ Bi-Ans-Ptr Boundary	64.7	40.3 [39.4, 41.3]	24.4	100	5
106	OTF dict+spelling	64.1	35.3 [34.3, 36.2]	28.8	108	-2
107	OTF spelling	62.9	32.6 [31.6, 33.5]	30.3	110	-3
108	OTF spelling+lemma	62.6	32.2 [31.2, 33.1]	30.4	111	-3
109	Dynamic Chunk Reader	62.5	37.8 [36.8, 38.8]	24.7	106	3
110	RQA+IDR	61.1	45.5 [44.5, 46.5]	15.7	81	29
111	UQA	53.7	33.9 [33.0, 34.9]	19.8	109	2
112	UnsupervisedQA V1	44.2	32.1 [31.2, 33.0]	12.1	112	0

Adversarial Results. The model evaluation data used to construct the adversarial plots from Appendix B is summarized in Tables 13 and 14 for F1 scores and Tables 15 and 16 for EM scores.

Table 13: Model F1 scores for against the *adversarial* distribution shifts `AddSent`. Unless otherwise noted, all models are `single models`.

AddSent F1 Score Summary						
Rank	Name	SQuAD	AddSent	Gap	New Rank	Δ Rank
1	XLNet	95.1	76.5	18.6	2	-1
2	XLNET-123	94.9	68.9	26.0	8	-6
3	XLNET-123++	94.9	77.2	17.7	1	2
4	SpanBERT	94.6	71.5	23.2	4	0
5	BERT+WWM+MT	94.4	73.9	20.5	3	2
6	Tuned BERT-1seq Large Cased	93.3	71.2	22.1	5	1
7	InfoWord (large)	93.1	64.4	28.7	12	-5
8	BERT-Large Baseline	92.7	61.3	31.5	16	-8
9	BERT+MT	92.6	70.0	22.6	6	3
10	Tuned BERT Large Cased	92.6	65.3	27.3	9	1
11	DPN	92.0	65.1	26.9	10	1
12	ST_bl	92.0	60.2	31.8	22	-10
13	BERT-uncased	91.9	59.7	32.2	26	-13
14	EL-BERT	91.8	61.1	30.7	18	-4
15	BISAN	91.8	60.1	31.6	23	-8
16	BERT+Sparse-Transformer	91.6	60.1	31.5	24	-8
17	InfoWord (base)	91.4	61.0	30.4	19	-2

Continued on next page

AddSent F1 Score Summary

Rank	Name	SQuAD	AddSent	Gap	New Rank	Δ Rank
18	InfoWord-Base	91.4	61.0	30.3	20	-2
19	InfoWord BERT baseline (large)	91.3	61.3	30.0	17	2
20	Original BERT Large Cased	91.3	64.4	26.8	13	7
21	Common-sense Governed BERT-123 (May 8 2019)	91.1	64.0	27.1	14	7
22	InfoWord BERT baseline (base)	90.9	56.5	34.4	30	-8
23	Common-sense Governed BERT-123	90.6	69.4	21.2	7	16
24	MARS (ensemble, June 20 2018)	89.8	59.9	29.8	25	-1
25	MARS	89.5	57.2	32.3	29	-4
26	MARS (June 21 2018)	89.2	58.4	30.8	28	-2
27	MMIPN	88.9	50.4	38.6	37	-10
28	Reinforced Mnemonic Reader (ensemble model)	88.5	60.2	28.3	21	7
29	AttentionReader+ (ensemble)	88.2	50.7	37.5	35	-6
30	Reinforced Mnemonic Reader + A2D	88.1	61.3	26.9	15	15
31	Reinforced Mnemonic Reader + A2D + DA	88.1	64.8	23.3	11	20
32	BERT-COMPOUND	87.8	50.5	37.3	36	-4
33	BiDAF + Self Attention + ELMo (ensemble)	87.4	43.7	43.7	56	-23
34	AVIQA+ (ensemble) (aviqa team)	87.3	45.3	42.1	51	-17
35	EAZI (ensemble)	86.9	42.7	44.2	63	-28
36	MEMEN+ (Ensemble)	86.8	44.5	42.4	53	-17
37	DNET (ensemble)	86.7	49.8	36.9	39	-2
38	BERT-INDEPENDENT	86.7	50.1	36.5	38	0
39	Reinforced Mnemonic Reader	86.7	58.5	28.2	27	12
40	MDReader	86.0	45.7	40.3	48	-8
41	BiDAF + Self Attention + ELMo	85.9	44.3	41.6	55	-14
42	BiDAF + Self-Attention + ELMo	85.8	44.4	41.4	54	-12
43	MDReader0	85.5	43.4	42.2	57	-14
44	MEMEN+	85.5	42.8	42.7	62	-18
45	MEMEN	85.3	43.4	42.0	58	-13
46	Interactive AoA Reader (Ensemble)	85.3	39.7	45.6	69	-23
47	EAZI	85.1	43.2	42.0	60	-13
48	AttentionReader+	84.9	53.9	31.0	31	17
49	DNET	84.9	49.2	35.7	40	9
50	MARS (Jan 23)	84.7	52.4	32.4	33	17
51	FRC	84.6	45.3	39.3	50	1
52	Jenga (ensemble)	84.5	40.4	44.0	67	-15
53	RaSoR + TR + LM	84.2	47.0	37.1	43	10
54	gqa	83.9	47.3	36.6	42	12
55	FusionNet	83.9	46.5	37.4	45	10
56	AVIQA-v2	83.3	42.9	40.4	61	-5
57	DCN+	83.1	44.5	38.6	52	5
58	Jenga	82.8	40.7	42.2	66	-8
59	Mixed model (ensemble)	82.8	48.5	34.2	41	18
60	two-attention-self-attention (ensemble)	82.7	40.4	42.3	68	-8
61	MEMEN (Ensemble, original model in paper)	82.7	36.3	46.4	84	-23
62	ReasoNet (Ensemble)	82.6	36.5	46.1	82	-20
63	eeAttNet	82.5	46.1	36.4	47	16
64	Mnemonic Reader (Ensemble)	82.4	46.2	36.2	46	18
65	Conductor-net	81.9	53.5	28.5	32	33
66	Interactive AoA Reader	81.9	39.1	42.8	72	-6
67	Jenga	81.8	45.6	36.2	49	18
68	BiDAF + Self Attention	81.0	36.4	44.7	83	-15
69	two-attention-self-attention	81.0	41.6	39.4	64	5
70	AVIQA	80.5	39.3	41.3	70	0
71	attention+self-attention	80.5	38.7	41.7	74	-3
72	Smarnet	80.2	50.9	29.3	34	38
73	Mnemonic Reader	80.1	46.6	33.5	44	29
74	MAMCN	79.9	37.9	42.0	76	-2
75	M-NET	79.8	41.6	38.2	65	10
76	JNet	79.8	37.9	41.9	77	-1
77	Ruminating Reader	79.5	37.4	42.1	79	-2

Continued on next page

AddSent F1 Score Summary						
Rank	Name	SQuAD	AddSent	Gap	New Rank	Δ Rank
78	ReasoNet	79.4	36.9	42.4	81	-3
79	SimpleBaseline	78.2	34.3	44.0	88	-9
80	PQMN	77.8	43.3	34.5	59	21
81	AllenNLP BiDAF	77.2	35.8	41.4	85	-4
82	Match-LSTM w/ Ans-Ptr (Ensemble)	77.0	29.4	47.6	91	-9
83	Iterative Co-Attention Network	76.8	32.7	44.1	89	-6
84	BIDAF-INDEPENDENT-DSS	76.3	35.3	41.0	86	-2
85	BIDAF-INDEPENDENT	74.6	31.9	42.7	90	-5
86	Match-LSTM w/ Bi-Ans-Ptr Boundary	73.7	27.3	46.4	92	-6
87	OTF dict+spelling	73.1	37.2	35.9	80	7
88	OTF spelling	72.0	34.7	37.4	87	1
89	OTF spelling+lemma	72.0	38.8	33.2	73	16
90	RQA+IDR	71.4	38.2	33.2	75	15
91	Dynamic Chunk Reader	71.0	37.8	33.2	78	13
92	UQA	64.0	39.3	24.7	71	21
93	UnsupervisedQA V1	54.7	20.9	33.8	93	0

Table 14: Model F1 scores for against the *adversarial* distribution shifts `AddOneSent`. Unless otherwise noted, all models are `single models`.

AddOneSent F1 Score Summary						
Rank	Name	SQuAD	AddOneSent	Gap	New Rank	Δ Rank
1	XLNet	95.1	83.5	11.6	2	-1
2	XLNET-123	94.9	78.4	16.5	7	-5
3	XLNET-123++	94.9	84.5	10.4	1	2
4	SpanBERT	94.6	79.8	14.9	4	0
5	BERT+WWM+MT	94.4	80.8	13.6	3	2
6	Tuned BERT-1seq Large Cased	93.3	78.6	14.7	6	0
7	InfoWord (large)	93.1	72.7	20.4	12	-5
8	BERT-Large Baseline	92.7	71.1	21.6	15	-7
9	BERT+MT	92.6	77.1	15.5	8	1
10	Tuned BERT Large Cased	92.6	73.8	18.8	10	0
11	DPN	92.0	73.6	18.5	11	0
12	ST_bl	92.0	69.8	22.2	21	-9
13	BERT-uncased	91.9	69.5	22.4	23	-10
14	EL-BERT	91.8	70.5	21.3	17	-3
15	BISAN	91.8	70.5	21.3	18	-3
16	BERT+Sparse-Transformer	91.6	69.4	22.3	24	-8
17	InfoWord (base)	91.4	70.3	21.1	19	-2
18	InfoWord-Base	91.4	70.3	21.0	20	-2
19	InfoWord BERT baseline (large)	91.3	71.1	20.2	16	3
20	Original BERT Large Cased	91.3	72.5	18.7	13	7
21	Common-sense Governed BERT-123 (May 8 2019)	91.1	74.4	16.7	9	12
22	InfoWord BERT baseline (base)	90.9	67.0	23.9	28	-6
23	Common-sense Governed BERT-123	90.6	78.8	11.8	5	18
24	MARS (ensemble, June 20 2018)	89.8	68.7	21.1	25	-1
25	MARS	89.5	65.6	24.0	30	-5
26	MARS (June 21 2018)	89.2	67.5	21.8	27	-1
27	MMIPN	88.9	60.6	28.4	36	-9
28	Reinforced Mnemonic Reader (ensemble model)	88.5	68.3	20.2	26	2
29	AttentionReader+ (ensemble)	88.2	61.4	26.7	34	-5
30	Reinforced Mnemonic Reader + A2D	88.1	69.5	18.6	22	8
31	Reinforced Mnemonic Reader + A2D + DA	88.1	71.4	16.7	14	17
32	BERT-COMPOUND	87.8	61.4	26.3	35	-3

Continued on next page

AddOneSent F1 Score Summary

Rank	Name	SQuAD	AddOneSent	Gap	New Rank	Δ Rank
33	BiDAF + Self Attention + ELMo (ensemble)	87.4	54.8	32.6	54	-21
34	AVIQA+ (ensemble) (aviqa team)	87.3	56.1	31.2	48	-14
35	EAZI (ensemble)	86.9	54.4	32.5	57	-22
36	MEMEN+ (Ensemble)	86.8	55.0	31.8	53	-17
37	DNET (ensemble)	86.7	60.5	26.2	38	-1
38	BERT-INDEPENDENT	86.7	61.6	25.0	32	6
39	Reinforced Mnemonic Reader	86.7	66.8	19.9	29	10
40	MDReader	86.0	57.0	29.0	45	-5
41	BiDAF + Self Attention + ELMo	85.9	54.7	31.2	56	-15
42	BiDAF + Self-Attention + ELMo	85.8	54.7	31.2	55	-13
43	MDReader0	85.5	54.0	31.6	60	-17
44	MEMEN+	85.5	53.8	31.7	61	-17
45	MEMEN	85.3	54.3	31.0	58	-13
46	Interactive AoA Reader (Ensemble)	85.3	49.5	35.8	71	-25
47	EAZI	85.1	53.6	31.5	62	-15
48	AttentionReader+	84.9	63.3	21.7	31	17
49	DNET	84.9	59.4	25.5	40	9
50	MARS (Jan 23)	84.7	60.5	24.2	37	13
51	FRC	84.6	55.9	28.7	50	1
52	Jenga (ensemble)	84.5	52.7	31.7	65	-13
53	RaSoR + TR + LM	84.2	57.0	27.1	44	9
54	gqa	83.9	57.9	26.0	43	11
55	FusionNet	83.9	56.6	27.3	47	8
56	AVIQA-v2	83.3	55.3	28.0	51	5
57	DCN+	83.1	54.3	28.8	59	-2
58	Jenga	82.8	52.1	30.8	66	-8
59	Mixed model (ensemble)	82.8	58.4	24.3	42	17
60	two-attention-self-attention (ensemble)	82.7	50.6	32.1	69	-9
61	MEMEN (Ensemble, original model in paper)	82.7	47.6	35.1	78	-17
62	ReasoNet (Ensemble)	82.6	48.0	34.6	74	-12
63	eeAttNet	82.5	58.5	24.0	41	22
64	Mnemonic Reader (Ensemble)	82.4	55.3	27.1	52	12
65	Conductor-net	81.9	61.6	20.3	33	32
66	Interactive AoA Reader	81.9	49.2	32.7	72	-6
67	Jenga	81.8	56.8	25.0	46	21
68	BiDAF + Self Attention	81.0	47.0	34.1	82	-14
69	two-attention-self-attention	81.0	51.1	29.9	68	1
70	AVIQA	80.5	51.3	29.2	67	3
71	attention+self-attention	80.5	50.0	30.4	70	1
72	Smarnet	80.2	60.1	20.0	39	33
73	Mnemonic Reader	80.1	56.0	24.1	49	24
74	MAMCN	79.9	47.7	32.3	76	-2
75	M-NET	79.8	53.1	26.7	64	11
76	JNet	79.8	47.0	32.8	81	-5
77	Ruminating Reader	79.5	47.7	31.8	77	0
78	ReasoNet	79.4	47.1	32.3	80	-2
79	SimpleBaseline	78.2	44.9	33.4	88	-9
80	PQMN	77.8	53.4	24.4	63	17
81	AllenNLP BiDAF	77.2	46.2	31.0	84	-3
82	Match-LSTM w/ Ans-Ptr (Ensemble)	77.0	41.8	35.2	91	-9
83	Iterative Co-Attention Network	76.8	43.9	32.9	89	-6
84	BIDAF-INDEPENDENT-DSS	76.3	45.7	30.6	85	-1
85	BIDAF-INDEPENDENT	74.6	42.6	32.0	90	-5
86	Match-LSTM w/ Bi-Ans-Ptr Boundary	73.7	39.0	34.8	92	-6
87	OTF dict+spelling	73.1	46.9	26.2	83	4
88	OTF spelling	72.0	44.9	27.1	87	1
89	OTF spelling+lemma	72.0	48.1	23.8	73	16
90	RQA+IDR	71.4	47.3	24.1	79	11
91	Dynamic Chunk Reader	71.0	45.1	25.9	86	5
92	UQA	64.0	48.0	16.1	75	17

Continued on next page

AddOneSent F1 Score Summary						
Rank	Name	SQuAD	AddOneSent	Gap	New Rank	Δ Rank
93	UnsupervisedQA V1	54.7	25.8	28.9	93	0

Table 15: Model EM scores for against the *adversarial* distribution shifts `AddSent`. Unless otherwise noted, all models are `single models`.

AddSent EM Score Summary						
Rank	Name	SQuAD	AddSent	Gap	New Rank	Δ Rank
1	XLNet	89.9	71.2	18.7	2	-1
2	XLNET-123++	89.9	73.3	16.6	1	1
3	XLNET-123	89.6	63.7	25.9	8	-5
4	SpanBERT	88.8	66.3	22.5	5	-1
5	BERT+WWM+MT	88.7	69.4	19.2	3	2
6	Tuned BERT-1seq Large Cased	87.5	65.8	21.7	6	0
7	InfoWord (large)	87.3	59.6	27.7	10	-3
8	BERT-Large Baseline	86.6	56.2	30.4	16	-8
9	Tuned BERT Large Cased	86.5	60.4	26.1	9	0
10	BERT+MT	86.5	64.4	22.1	7	3
11	ST_bl	85.4	55.0	30.4	23	-12
12	EL-BERT	85.3	56.3	29.0	15	-3
13	BISAN	85.3	54.8	30.5	24	-11
14	BERT+Sparse-Transformer	85.1	54.4	30.7	25	-11
15	DPN	85.0	58.9	26.1	13	2
16	BERT-uncased	84.9	54.0	30.9	26	-10
17	InfoWord-Base	84.7	55.7	29.0	19	-2
18	InfoWord (base)	84.7	55.7	29.0	20	-2
19	InfoWord BERT baseline (base)	84.4	51.4	33.0	30	-11
20	Original BERT Large Cased	84.3	58.8	25.5	14	6
21	InfoWord BERT baseline (large)	84.3	56.2	28.1	17	4
22	MARS (ensemble, June 20 2018)	84.0	55.3	28.7	22	0
23	Common-sense Governed BERT-123	83.9	66.6	17.3	4	19
24	MARS	83.2	51.8	31.4	29	-5
25	MARS (June 21 2018)	83.1	53.2	29.9	27	-2
26	Common-sense Governed BERT-123 (May 8 2019)	82.9	59.3	23.6	12	14
27	Reinforced Mnemonic Reader (ensemble model)	82.3	55.6	26.7	21	6
28	AttentionReader+ (ensemble)	81.8	45.2	36.6	35	-7
29	MMIPN	81.6	44.4	37.2	38	-9
30	Reinforced Mnemonic Reader + A2D	81.5	56.0	25.5	18	12
31	Reinforced Mnemonic Reader + A2D + DA	81.4	59.5	21.9	11	20
32	BiDAF + Self Attention + ELMo (ensemble)	81.0	38.7	42.3	54	-22
33	BERT-COMPOUND	80.7	44.9	35.8	36	-3
34	AVIQA+ (ensemble) (aviqa team)	80.6	40.7	39.9	47	-13
35	EAZI (ensemble)	80.4	38.1	42.3	58	-23
36	MEMEN+ (Ensemble)	80.4	39.5	40.9	51	-15
37	DNET (ensemble)	80.2	44.6	35.6	37	0
38	Reinforced Mnemonic Reader	79.5	53.0	26.5	28	10
39	MDReader	79.0	41.1	37.9	45	-6
40	BERT-INDEPENDENT	78.7	44.1	34.6	40	0
41	BiDAF + Self Attention + ELMo	78.6	38.6	40.0	57	-16
42	BiDAF + Self-Attention + ELMo	78.6	38.7	39.9	55	-13
43	MEMEN	78.2	38.0	40.2	59	-16
44	MEMEN+	78.2	37.5	40.7	61	-17
45	MDReader0	78.2	38.7	39.5	56	-11
46	EAZI	78.0	37.7	40.3	60	-14
47	Interactive AoA Reader (Ensemble)	77.8	35.2	42.6	68	-21

Continued on next page

AddSent EM Score Summary						
Rank	Name	SQuAD	AddSent	Gap	New Rank	Δ Rank
48	DNET	77.6	44.3	33.3	39	9
49	RaSoR + TR + LM	77.6	42.2	35.4	43	6
50	AttentionReader+	77.3	48.3	29.0	31	19
51	Jenga (ensemble)	77.2	36.1	41.1	64	-13
52	gqa	77.1	42.6	34.5	42	10
53	MARS (Jan 23)	76.9	46.2	30.7	33	20
54	FRC	76.2	39.8	36.4	49	5
55	FusionNet	76.0	40.9	35.1	46	9
56	AVIQA-v2	75.9	37.5	38.4	62	-6
57	MEMEN (Ensemble, original model in paper)	75.4	31.5	43.9	81	-24
58	Mixed model (ensemble)	75.3	43.1	32.2	41	17
59	two-attention-self-attention (ensemble)	75.2	35.4	39.8	67	-8
60	DCN+	75.1	39.4	35.7	52	8
61	ReasoNet (Ensemble)	75.0	31.8	43.2	80	-19
62	eeAttNet	74.6	41.3	33.3	44	18
63	Jenga	74.4	35.5	38.9	66	-3
64	Mnemonic Reader (Ensemble)	74.3	40.7	33.6	48	16
65	Interactive AoA Reader	73.6	32.9	40.7	73	-8
66	Jenga	73.3	39.1	34.2	53	13
67	Conductor-net	73.2	47.3	25.9	32	35
68	two-attention-self-attention	72.6	36.1	36.5	65	3
69	AVIQA	72.5	34.3	38.2	70	-1
70	BiDAF + Self Attention	72.1	30.6	41.5	84	-14
71	attention+self-attention	71.7	33.7	38.0	71	0
72	Smarnet	71.4	45.5	25.9	34	38
73	M-NET	71.0	35.2	35.8	69	4
74	Mnemonic Reader	71.0	39.8	31.2	50	24
75	MAMCN	71.0	32.8	38.2	74	1
76	Ruminating Reader	70.6	32.4	38.2	77	-1
77	JNet	70.6	33.2	37.4	72	5
78	ReasoNet	70.6	31.1	39.5	82	-4
79	SimpleBaseline	69.6	29.1	40.5	87	-8
80	PQMN	68.3	36.8	31.5	63	17
81	Match-LSTM w/ Ans-Ptr (Ensemble)	67.9	24.3	43.6	91	-10
82	AllenNLP BiDAF	67.6	29.4	38.2	86	-4
83	Iterative Co-Attention Network	67.5	26.8	40.7	89	-6
84	BIDAF-INDEPENDENT-DSS	66.5	29.6	36.9	85	-1
85	BIDAF-INDEPENDENT	64.9	26.3	38.6	90	-5
86	Match-LSTM w/ Bi-Ans-Ptr Boundary	64.7	22.1	42.6	92	-6
87	OTF dict+spelling	64.1	31.0	33.1	83	4
88	OTF spelling	62.9	28.8	34.1	88	0
89	OTF spelling+lemma	62.6	32.1	30.5	78	11
90	Dynamic Chunk Reader	62.5	32.5	30.0	76	14
91	RQA+IDR	61.1	31.9	29.2	79	12
92	UQA	53.7	32.6	21.1	75	17
93	UnsupervisedQA V1	44.2	16.5	27.7	93	0

Table 16: Model EM scores for against the *adversarial* distribution shifts `AddOneSent`. Unless otherwise noted, all models are `single models`.

AddOneSent EM Score Summary						
Rank	Name	SQuAD	AddOneSent	Gap	New Rank	Δ Rank
1	XLNet	89.9	78.2	11.7	2	-1
2	XLNET-123++	89.9	80.3	9.6	1	1

Continued on next page

AddOneSent EM Score Summary

Rank	Name	SQuAD	AddOneSent	Gap	New Rank	Δ Rank
3	XLNET-123	89.6	73.5	16.1	6	-3
4	SpanBERT	88.8	74.3	14.5	5	-1
5	BERT+WWM+MT	88.7	75.6	13.1	4	1
6	Tuned BERT-1seq Large Cased	87.5	73.1	14.4	7	-1
7	InfoWord (large)	87.3	67.5	19.8	11	-4
8	BERT-Large Baseline	86.6	65.4	21.2	15	-7
9	Tuned BERT Large Cased	86.5	68.6	17.9	9	0
10	BERT+MT	86.5	70.6	15.9	8	2
11	ST_bl	85.4	63.5	21.9	22	-11
12	EL-BERT	85.3	64.7	20.6	17	-5
13	BISAN	85.3	64.6	20.7	18	-5
14	BERT+Sparse-Transformer	85.1	63.2	21.9	24	-10
15	DPN	85.0	66.5	18.5	12	3
16	BERT-uncased	84.9	62.8	22.1	25	-9
17	InfoWord-Base	84.7	63.8	20.9	19	-2
18	InfoWord (base)	84.7	63.8	20.9	20	-2
19	InfoWord BERT baseline (base)	84.4	61.2	23.2	28	-9
20	Original BERT Large Cased	84.3	65.9	18.4	13	7
21	InfoWord BERT baseline (large)	84.3	65.4	18.9	16	5
22	MARS (ensemble, June 20 2018)	84.0	63.5	20.5	23	-1
23	Common-sense Governed BERT-123	83.9	76.0	7.9	3	20
24	MARS	83.2	59.0	24.2	30	-6
25	MARS (June 21 2018)	83.1	61.6	21.5	27	-2
26	Common-sense Governed BERT-123 (May 8 2019)	82.9	68.6	14.3	10	16
27	Reinforced Mnemonic Reader (ensemble model)	82.3	62.8	19.5	26	1
28	AttentionReader+ (ensemble)	81.8	55.2	26.6	33	-5
29	MMIPN	81.6	53.6	28.0	37	-8
30	Reinforced Mnemonic Reader + A2D	81.5	63.6	17.9	21	9
31	Reinforced Mnemonic Reader + A2D + DA	81.4	65.6	15.8	14	17
32	BiDAF + Self Attention + ELMo (ensemble)	81.0	48.7	32.3	53	-21
33	BERT-COMPOUND	80.7	55.5	25.2	32	1
34	AVIQA+ (ensemble) (aviqa team)	80.6	50.7	29.9	46	-12
35	EAZI (ensemble)	80.4	49.0	31.4	50	-15
36	MEMEN+ (Ensemble)	80.4	49.3	31.1	48	-12
37	DNET (ensemble)	80.2	54.9	25.3	34	3
38	Reinforced Mnemonic Reader	79.5	60.5	19.0	29	9
39	MDReader	79.0	51.2	27.8	45	-6
40	BERT-INDEPENDENT	78.7	54.2	24.5	36	4
41	BiDAF + Self Attention + ELMo	78.6	47.9	30.7	58	-17
42	BiDAF + Self-Attention + ELMo	78.6	48.0	30.6	57	-15
43	MEMEN	78.2	47.4	30.8	60	-17
44	MEMEN+	78.2	47.4	30.8	61	-17
45	MDReader0	78.2	48.1	30.1	56	-11
46	EAZI	78.0	47.0	31.0	62	-16
47	Interactive AoA Reader (Ensemble)	77.8	43.4	34.4	70	-23
48	DNET	77.6	53.3	24.3	40	8
49	RaSoR + TR + LM	77.6	51.4	26.2	44	5
50	AttentionReader+	77.3	56.7	20.6	31	19
51	Jenga (ensemble)	77.2	47.0	30.2	63	-12
52	gqa	77.1	52.2	24.9	42	10
53	MARS (Jan 23)	76.9	53.5	23.4	39	14
54	FRC	76.2	48.8	27.4	52	2
55	FusionNet	76.0	49.8	26.2	47	8
56	AVIQA-v2	75.9	48.9	27.0	51	5
57	MEMEN (Ensemble, original model in paper)	75.4	41.5	33.9	73	-16
58	Mixed model (ensemble)	75.3	52.0	23.3	43	15
59	two-attention-self-attention (ensemble)	75.2	44.3	30.9	69	-10
60	DCN+	75.1	47.7	27.4	59	1
61	ReasoNet (Ensemble)	75.0	41.7	33.3	72	-11
62	eeAttNet	74.6	52.6	22.0	41	21

Continued on next page

AddOneSent EM Score Summary

Rank	Name	SQuAD	AddOneSent	Gap	New Rank	Δ Rank
63	Jenga	74.4	45.6	28.8	65	-2
64	Mnemonic Reader (Ensemble)	74.3	48.7	25.6	54	10
65	Interactive AoA Reader	73.6	41.4	32.2	74	-9
66	Jenga	73.3	49.2	24.1	49	17
67	Conductor-net	73.2	54.4	18.8	35	32
68	two-attention-self-attention	72.6	44.4	28.2	68	0
69	AVIQA	72.5	45.7	26.8	64	5
70	BiDAF + Self Attention	72.1	39.7	32.4	81	-11
71	attention+self-attention	71.7	43.3	28.4	71	0
72	Smarnet	71.4	53.6	17.8	38	34
73	M-NET	71.0	45.4	25.6	67	6
74	Mnemonic Reader	71.0	48.5	22.5	55	19
75	MAMCN	71.0	40.9	30.1	76	-1
76	Ruminating Reader	70.6	41.4	29.2	75	1
77	JNet	70.6	40.4	30.2	77	0
78	ReasoNet	70.6	39.3	31.3	82	-4
79	SimpleBaseline	69.6	37.9	31.7	87	-8
80	PQMN	68.3	45.5	22.8	66	14
81	Match-LSTM w/ Ans-Ptr (Ensemble)	67.9	34.8	33.1	91	-10
82	AllenNLP BiDAF	67.6	38.0	29.6	86	-4
83	Iterative Co-Attention Network	67.5	35.6	31.9	89	-6
84	BIDAF-INDEPENDENT-DSS	66.5	38.2	28.3	85	-1
85	BIDAF-INDEPENDENT	64.9	34.9	30.0	90	-5
86	Match-LSTM w/ Bi-Ans-Ptr Boundary	64.7	31.6	33.1	92	-6
87	OTF dict+spelling	64.1	38.8	25.3	83	4
88	OTF spelling	62.9	37.0	25.9	88	0
89	OTF spelling+lemma	62.6	39.9	22.7	79	10
90	Dynamic Chunk Reader	62.5	38.6	23.9	84	6
91	RQA+IDR	61.1	40.0	21.1	78	13
92	UQA	53.7	39.9	13.8	80	12
93	UnsupervisedQA V1	44.2	20.3	23.9	93	0