

What do Models Learn from Question Answering Datasets?

Priyanka Sen

Amazon Alexa

sepriyan@amazon.com

Amir Saffari

Amazon Alexa

amsafari@amazon.com

Abstract

While models have reached superhuman performance on popular question answering (QA) datasets such as SQuAD, they have yet to outperform humans on the task of question answering itself. In this paper, we investigate if models are learning reading comprehension from QA datasets by evaluating BERT-based models across five datasets. We evaluate models on their generalizability to out-of-domain examples, responses to missing or incorrect data, and ability to handle question variations. We find that no single dataset is robust to all of our experiments and identify shortcomings in both datasets and evaluation methods. Following our analysis, we make recommendations for building future QA datasets that better evaluate the task of question answering through reading comprehension. We also release code to convert QA datasets to a shared format for easier experimentation at <https://github.com/amazon-research/qa-dataset-converter>.

1 Introduction

Question answering (QA) through reading comprehension has seen considerable progress in recent years. This progress is in large part due to the release of large language models like BERT (Devlin et al., 2019) and new datasets that have introduced impossible questions (Rajpurkar et al., 2018), bigger scales (Kwiatkowski et al., 2019), and context (Choi et al., 2018; Reddy et al., 2019) to question answering. At the time of writing this paper, models have outperformed human baselines on the widely-used SQuAD 1.1 and SQuAD 2.0 datasets, and more challenging datasets such as QuAC have models less than 7 F1 points away. Despite these increases in F1 scores, we are still far from saying question answering is a solved problem.

Concerns have been raised over how challenging QA datasets are. Previous work has found that

simple heuristics can give good performance on SQuAD 1.1 (Weissenborn et al., 2017), and successful SQuAD models lack robustness by giving inconsistent answers (Ribeiro et al., 2019) or being vulnerable to adversarial attacks (Jia and Liang, 2017; Wallace et al., 2019). If state-of-the-art models are excelling at test sets but not solving the underlying task of reading comprehension, then our test sets are flawed. We need to better understand what models learn from QA datasets. In this work, we ask three questions: (1) Does performance on individual QA datasets generalize to new datasets? (2) Do models need to learn reading comprehension for QA datasets?, and (3) Can QA models handle variations in questions?

To answer these questions, we evaluate BERT models trained on five QA datasets using simple generalization and robustness probes. We find that (1) Model performance does not generalize well outside of heuristics like question-context overlaps, (2) Removing or corrupting dataset examples does not always harm model performance, showing that models can rely on simpler methods than reading comprehension to answer questions, and (3) No dataset fully prepares models to handle question variations like filler words or negation. These findings suggest that while QA models can learn heuristics around question-context overlaps and named entities, they do not need to learn reading comprehension to perform well on QA datasets. Based on these findings, we make recommendations on how to better create and evaluate QA datasets.

2 Related Work

Our work is inspired by recent trends in NLP to evaluate generalizability and probe what models learn from datasets. In terms of generalizability, prior work has been done by Yogatama et al. (2019) who evaluated a SQuAD 1.1 model across

four datasets and Talmor and Berant (2019), who comprehensively evaluated ten QA datasets. The MRQA 2019 shared task (Fisch et al., 2019) evaluated transferability across multiple datasets, and Khashabi et al. (2020) proposed a method to train one model on 17 different QA datasets. In our work, we focus on question answering through reading comprehension and extend the work on generalizability by including impossible questions in all our datasets and analyzing the effects of question-context overlap on generalizability.

There is also growing interest in probing what models learn from datasets (McCoy et al., 2019; Geirhos et al., 2020; Richardson and Sabharwal, 2020; Si et al., 2020). Previous work in question answering has found that state-of-the-art models can get good performance on incomplete input (Agrawal et al., 2016; Sugawara et al., 2018; Niven and Kao, 2019), under-rely on important words, (Mudrakarta et al., 2018), and over-rely on simple heuristics (Weissenborn et al., 2017; Ko et al., 2020). Experiments on SQuAD in particular have shown that SQuAD models are vulnerable to adversarial attacks (Jia and Liang, 2017; Wallace et al., 2019) and not robust to paraphrases (Ribeiro et al., 2018; Gan and Ng, 2019).

Our work continues exploring what models learn by comprehensively testing multiple QA datasets against a variety of simple but informative probes. We take inspiration from previous studies, and we make novel contributions by using BERT, a state-of-the-art model, and running several experiments against five different QA datasets to investigate the progress made in reading comprehension.

3 Datasets

	Train	Dev
SQuAD	130,319	11,873
TriviaQA	110,647	14,229
NQ	110,857	3,368
QuAC	83,568	7,354
NewsQA	101,707	5,666

Table 1: Train and dev set sizes of the datasets used in our experiments

We compare five datasets in our experiments: SQuAD 2.0, TriviaQA, Natural Questions, QuAC, and NewsQA. All our datasets treat question answering as a reading comprehension task where

	Question	Context	Answer
SQuAD	10	120	3
TriviaQA	15	746	2
NQ	9	96	4
QuAC	7	395	14
NewsQA	8	709	4

Table 2: Comparison of the average number of words in questions, contexts, and answers in each dataset

the question is about a document and the answer is either an extracted span of text or labeled unanswerable. To consistently compare and experiment across models, we convert all datasets into a SQuAD 2.0 JSON format.¹ Since most datasets have a hidden test set, we evaluate models on the dev set and consequently refer to the dev sets as test sets in this paper. The train and dev sets sizes are shown in Table 1

Below we describe each dataset and any modifications we made to run our experiments:

SQuAD 2.0 (Rajpurkar et al., 2018) consists of 150K question-answer pairs on Wikipedia articles. To create SQuAD 1.1, crowd workers wrote questions about a Wikipedia paragraph and highlighted the answer (Rajpurkar et al., 2016). SQuAD 2.0 includes an additional 50K plausible but unanswerable questions written by crowd workers.

TriviaQA (Joshi et al., 2017) includes 95K question-answer pairs from trivia websites. The questions were written by trivia enthusiasts and the evidence documents were retrieved by the authors retrospectively. We use the variant of TriviaQA where the documents are Wikipedia articles.

Natural Questions (NQ) (Kwiatkowski et al., 2019) contains 300K questions from Google search logs. For each question, a crowd worker found a long and short answer on a Wikipedia page. We use the subset of NQ with a long answer and frame the task as finding the short answer in the long answer. We only include examples with answers in the paragraph text (as opposed to a table or list).

QuAC (Choi et al., 2018) contains 100K questions. To create QuAC, one crowd worker asked questions about a Wikipedia article to a second crowd worker, who answered by selecting a text span. To standardize training, we do not model contextual information, but we include QuAC to

¹<https://github.com/amazon-research/qa-dataset-converter>

		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
Fine-tuned on	SQuAD	75.6	46.7	48.7	20.2	41.1
	TriviaQA	49.8	58.7	42.1	20.4	10.5
	NQ	53.5	46.3	73.5	21.6	24.7
	QuAC	39.4	33.1	33.8	33.3	13.8
	NewsQA	52.1	38.4	41.7	20.4	60.1

Table 3: F1 scores of each fine-tuned model evaluated on each test set

see how models trained without context handle context-dependent questions.

NewsQA (Trischler et al., 2017) contains 100K questions on 10K CNN articles. One set of crowd workers wrote questions based on a headline and summary, and a second set of workers found the answer in the article. We reintroduce unanswerable questions that were excluded in the original paper.

There are notable differences among our datasets in terms of genre and how they were built. In Table 2, we see a large variation in the average number of words in questions, contexts, and answers. Despite these differences, all our datasets are reading comprehension tasks. We believe a good reading comprehension model should handle question answering well regardless of dataset differences, and so we compare across all five datasets.

4 Model

Hyperparameter	Value
Batch Size	24
Learning Rate	3e-5
Epochs	2
Max Seq Length	384
Doc Stride	128

Table 4: Hyperparameter values for fine-tuning BERT based on Devlin et al. (2019)

All models are initialized from a pre-trained BERT-Base uncased model² with 110M parameters. For each dataset, we fine-tune on the training set using Devlin et al. (2019)’s default hyperparameters shown in Table 4. We evaluate on the dev set with the SQuAD 2.0 evaluation script (Rajpurkar

²<https://github.com/google-research/bert#pre-trained-models>

et al., 2018). We run our experiments on a single Nvidia Tesla v100 16GB GPU.

In Table 5, we provide a comparison between our models and previously published BERT results. Differences occur when we make modifications to match SQuAD. We simplified NQ by removing the long answer identification task and framed the short answer task in a SQuAD format, so we see higher results than the NQ BERT baseline. For QuAC, we ignored all context-related fields and treated each example as an independent question, so we see lower results than models built on the full dataset. For NewsQA, we introduced impossible questions, resulting in lower performance. We accept these drops in performance since we are interested in comparing changes to a baseline rather than achieving state-of-the-art results.

Dataset	Reference	Ours
SQuAD	76.3 (Liu et al., 2019)	75.6
TriviaQA	56.3 (Yang et al., 2019)	58.7
NQ	52.7 (Alberti et al., 2019)	73.5
QuAC	54.4 (Qu et al., 2019)	33.3
NewsQA	66.8 (Takahashi et al., 2019)	60.1

Table 5: Comparison to previously reported F1 scores. Differences occur when we make modifications to match SQuAD.

5 Experiments

In this section, we discuss the experiments run to evaluate what models learn from QA datasets. All results are reported as F1 scores since they are correlated with Exact Match scores and are more forgiving to sometimes arbitrary cutoffs of answers (for example, we prefer to give some credit to a model for selecting “Charles III” even if the answer was “King Charles III”).

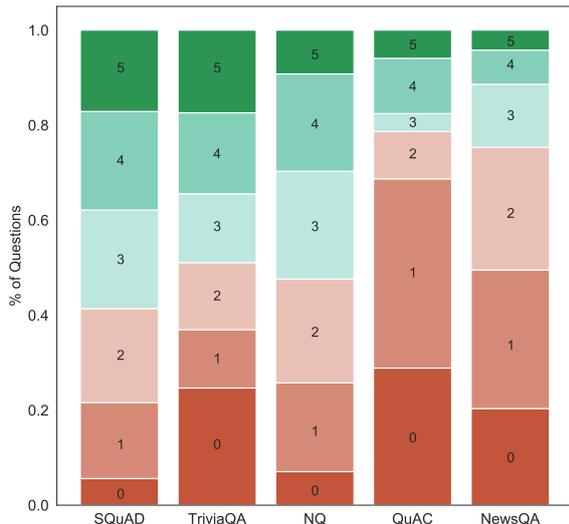


Figure 1: A bar graph of how many questions in each dataset are answered by 0, 1, 2, 3, 4, or 5 models

5.1 Does performance on individual QA datasets generalize to new datasets?

For our first experiment, we evaluate the generalizability of models to out-of-domain examples. While most work in QA has focused on evaluating against a single test set, generalizability is an important feature. If we cannot get good, generalizable performance on research datasets, we will struggle to expand to the variety of questions an open-domain QA system can face. Several papers have focused on generalizability by evaluating transferability across datasets (Talmor and Berant, 2019; Yatskar, 2019), generalizability to out-of-domain data (Fisch et al., 2019), and building cross-dataset evaluation methods (Dua et al., 2019).

We test generalizability by fine-tuning models on each dataset and evaluating against all five test sets. The results are reported as F1 scores in Table 3. The rows show a single model’s performance across all five datasets, and the columns show the performance of all the models on a single test set. The model-on-self baseline is indicated in bold.

All models take a drop in performance when evaluated on an out-of-domain test set. This shows that performance on an individual dataset does not generalize across datasets, confirming results found on different mixes of datasets (Talmor and Berant, 2019; Yogatama et al., 2019). However there is variation in how the models perform. For example, models score up to 53.5 F1 points on SQuAD without seeing SQuAD examples, while models do not score above 21.6 F1 points on QuAC without

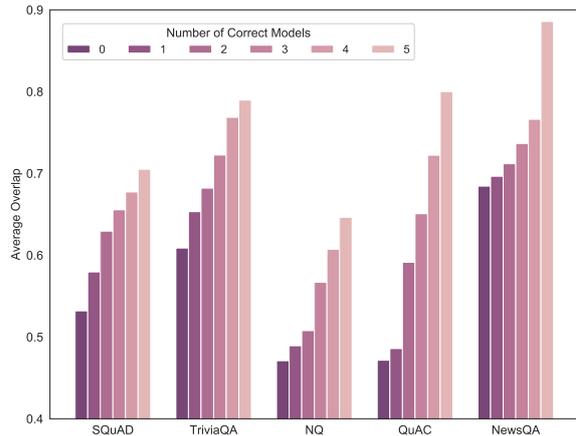


Figure 2: More models correctly answer answerable questions if they have higher question-context overlap.

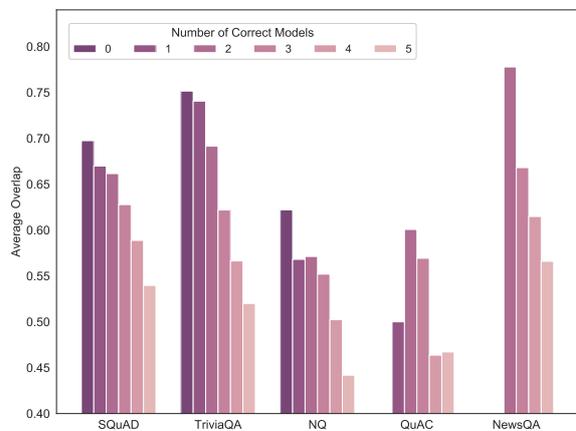


Figure 3: More models correctly answer impossible questions if they have lower question-context overlap. NewsQA has four bars since all impossible NewsQA questions were correctly answered by at least 1 model.

QuAC examples. This suggests that some test sets are easier than others.

To quantify this, we calculate what proportion of each test set can be correctly answered by how many models. This data is represented as a bar graph in Figure 1. Each bar represents one dataset, and the segments show how much of the test set is answered correctly by 0 to 5 of the models.

We consider questions easier if more models correctly answer them. The figure shows that QuAC and NewsQA are more challenging test sets and contain a higher proportion of questions that are answered by 0 or 1 model. In contrast, more than half of SQuAD and NQ and almost half of TriviaQA can be answered correctly by 3 or more models.

While difficult questions pose a challenge for QA models, too many easy questions inflate our understanding of a model’s performance. What

Experiment	Question	Answer Text	Answer Start
1 Original	Who was the Norse leader	Rollo	308
2 Random Label	Who was the Norse leader	succeeding	721
3 Shuffled Context	Who was the Norse leader	Rollo	480
4 Incomplete (first half)	Who was	Rollo	308
5 Incomplete (first word)	Who	Rollo	308
6 Incomplete (no words)		Rollo	308
7 Filler word	Who really was the Norse leader	Rollo	308
8 Negation	Who wasn't the Norse leader	Rollo	308

Table 6: Examples of how question-answer pairs are modified in each experiment

makes a question easy? We identified a trend between question difficulty and the overlap between the question and the context. We measured overlap as the number of words that appeared in both the question and the context divided by the number of words in the question. For answerable questions, Figure 2 shows that more models return correct answers when there is higher overlap, while Figure 3 shows that fewer models correctly identify impossible questions when there is higher overlap. This suggests that models learn to use question-context overlap to identify answers. Models may even over-rely on this strategy and return answers to impossible questions when no answer exists.

Overall, the results show that our models do not generalize well to different datasets. The models do seem to exploit question-context overlap, even on questions that are out-of-domain. Reducing the number of high overlap questions in a dataset could create more challenging datasets in the future and better evaluate generalization and test more complex strategies for question answering.

5.2 Do models need to learn reading comprehension for QA datasets?

State-of-the-art models get good performance on QA datasets, but does good performance mean good reading comprehension? Or are models able to take shortcuts to arrive at the same answers? We explore this by performing three dataset ablation experiments with random labels, shuffled contexts, and incomplete questions. If models can answer test set questions with incorrect or missing information, then the models are likely not learning the task of reading comprehension. The three experiments and their results are discussed in the next sections.

Dataset	Baseline	% of Random Labels		
		10%	50%	90%
SQuAD	78.5	77.1	73.9	32.1
TriviaQA	46.8	36.6	10.9	0.0
NQ	70.6	68.1	60.5	19.4
QuAC	20.3	16.4	1.8	0.3
NewsQA	56.3	50.8	30.2	2.0

Table 7: F1 scores of answered questions decrease as models are fine-tuned on increasingly noisy data.

5.2.1 Random Labels

A robust model should withstand some amount of noise at training time to offset annotation error. However if a model can perform well with a high level of noise, we should be wary of what the model has learned. In our first dataset ablation experiment, we evaluated how various amounts of noise at training time affected model performance.

To introduce noise to the training sets, we randomly selected 10%, 50%, or 90% of the training examples that were answerable and updated the answer to a random string from the same context and of the same length as the original answer. We ensured that the random answer contained no overlaps with the original answer. For simplicity, we did not alter impossible examples. An example of a random label is in the second row of Table 6.

We fine-tuned new models on increasingly noisy training sets and evaluated them on the original test sets. The results are in Table 7 in terms of F1 scores and reported only for answerable questions. On training sets with 10% random labels, all models see an F1 score drop. SQuAD, NQ, and NewsQA achieve over 90% of their baseline score, showing robustness to a reasonable level of noise. TriviaQA

and QuAC take larger F1 hits (achieving only 78% and 81% of their baselines), suggesting that they are less robust to this noise.

As the amount of noise increases, most F1 scores drop to nearly 0. SQuAD and NQ, however, are suspiciously robust even when 90% of their training examples are random. SQuAD achieves 41% of its baseline and NQ achieves 27% of its baseline with training sets that are 90% noise. We find it unlikely that randomly selected strings provide a signal, so this suggests that some examples in each test set are answerable trivially and without learning reading comprehension.

5.2.2 Shuffled Context

Dataset	Baseline	Shuffled Context
SQuAD	75.6	70.5
TriviaQA	58.7	38.7
NQ	73.5	64.5
QuAC	33.3	32.4
NewsQA	60.1	48.2

Table 8: F1 scores decrease, but not dramatically, on test sets with shuffled context sentences.

The task of reading comprehension aims to measure how well a model understands a given passage. If a model is able to answer questions without understanding the logic or structure of a passage, we can get high scores on a test set but be no closer to learning reading comprehension. In our second experiment, we investigate how much of model performance can be accounted for without understanding the full passage.

For each context paragraph in the test set, we split the context by sentence, randomly shuffled the sentences, and rejoined the sentences into a new paragraph. The original answer text remained the same, but the answer start token was updated by locating the correct answer text in the shuffled context. An example is in the third row of Table 6.

We used our models fine-tuned on the original training sets and evaluated on the shuffled context test sets. The results are in Table 8. TriviaQA sees the largest drop in performance, achieving only 66% of its baseline, followed by NewsQA with 80% of its baseline. SQuAD and QuAC, on the other hand, get over 93% of their original baselines even with randomly shuffled contexts. TriviaQA and NewsQA have longer contexts, with an average

of over 700 words, and so shuffling longer contexts seems more detrimental. While these results show that models do not rely on naive approaches, like position, they do show that for many questions, models do not need to understand a paragraph’s structure to correctly predict the answer.

5.2.3 Incomplete Input

Dataset	Baseline	First Half	First Word	No Words
SQuAD	75.6	36.4	22.8	49.5
TriviaQA	58.7	45.8	31.8	30.4
NQ	73.5	61.4	50.3	32.7
QuAC	33.3	25.2	22.4	20.2
NewsQA	60.1	43.6	26.3	13.4

Table 9: F1 scores decrease with incomplete input, but models can return correct answers with no question.

QA dataset creators and their crowd workers spend considerable effort hand-crafting questions that are meant to challenge a model’s ability to understand language. But are models using the questions? In previous work, Agrawal et al. (2016) found that a Visual Question Answering (VQA) model could get good performance with just half the original question, while Sugawara et al. (2018) saw drops in BiDAF model performance on QA datasets with increasingly incomplete questions. We expand on these previous works by using BERT, including impossible questions, and introducing NER baselines for comparison.

We created three variants of each test set containing only the first half, first word, or no words from each question. The answer expectations were not changed. Examples are in the fourth, fifth, and sixth rows of Table 6.

We evaluated models fine-tuned on the original training set on the incomplete test sets. The results are in Table 9. F1 scores mostly decrease on test sets with incomplete input, but models can return correct answers without being given the question. SQuAD achieves 65% of its baseline given no question, an increase from the First Word test set primarily because of higher success on impossible questions. NQ achieves up to 68% of its baseline F1 score given the first word and up to 44% given no question. These results show that not all examples require full or any question understanding to make correct predictions. We also see higher F1

scores compared to Sugawara et al. (2018) when using the first word. In TriviaQA, Sugawara et al. (2018) saw their F1 score drop by 75% (from 49.3 to 12.5) while we see a drop of 46% (from 58.7 to 31.8), which could suggest that our BERT models have overfit more than BiDAF models.

How can models answer without the full question? We investigated our results further by creating two naive named entity recognition (NER) baselines using spaCy³ to see if models can rely on entity types. For our First Word NER baseline, we used the first word of the question to choose an entity as the answer. If a question started with “who”, we returned the first person entity in the context, for “when”, the first date, for “where”, the first location, and for “what”, the first organization, event, or work of art. The results are in the First Word NER column of Table 10. With the exception of NewsQA, we are able to achieve over 40% of baseline performance with this NER system.

Dataset	Baseline	First Word NER	Person NER
SQuAD	75.6	30.0	26.7
TriviaQA	58.7	25.2	8.9
NQ	73.5	35.9	24.1
QuAC	33.3	17.2	6.0
NewsQA	60.1	11.3	8.8

Table 10: NER baselines on QA datasets

Our Person NER baseline returns the first person entity found in each context. The results are shown in Table 10. Both NQ and SQuAD achieve 33-35% of their baseline by only extracting the first person entity. TriviaQA sees a much larger drop when using only person entities, suggesting there is more entity type variety in the TriviaQA test set. These results show that some questions can be answered by extracting entity types and without needing most or all of the question text.

5.3 Can QA models handle variations in questions?

The previous section found that models can perform well on test sets even as seemingly important features are stripped from datasets. This section considers the opposite problem: Can models remain robust as features are added to datasets? To

³<https://spacy.io>

analyze this, we run two experiments where we add filler words and negation to test set questions.

5.3.1 Filler Words

Dataset	Baseline	Filler Words
SQuAD	75.6	69.5
TriviaQA	58.7	56.5
NQ	73.5	67.6
QuAC	33.3	31.2
NewsQA	60.1	54.8

Table 11: F1 scores slightly decrease on test sets where a filler word is added to the question.

If a QA model understands a question, it should handle semantically equivalent questions equally well. While previous works have shown poor model performance on paraphrased questions (Ribeiro et al., 2018; Gan and Ng, 2019), we take an even simpler approach by adding filler words that do not affect the rest of the question. For each question in the test set, we randomly added one of three filler words (*really*, *definitely*, or *actually*) before the main verb, as identified by spaCy. An example is shown in the seventh row of Table 6.

Table 11 shows the results of models fine-tuned on their original training sets and evaluated on the filler word test sets. All models drop between 2 to 5 F1 points. Although these drops may seem small, they do show that even a naive approach can hurt performance. It is no surprise that more sophisticated paraphrases of questions cause models to fail. The SQuAD model in particular had better performance when 50% of the training set was randomly labeled (73.9) than when filler words were added to the test set (69.5), suggesting that our models have become robust to less consequential features.

5.3.2 Negation

Negation is an important grammatical construction for QA systems to understand. Giving the same answer to a question and its negative (Who invented the telescope? vs. Who didn’t invent the telescope?) can frustrate or mislead users. In previous work, Kassner and Schütze (2019) studied negation by manually negating 305 SQuAD 1.1 questions and found that a BERT model largely ignored negation. We expand on this work by using the full SQuAD 2.0 dataset and comparing performance across five datasets.

Dataset	Baseline	Negation
SQuAD	75.6	2.0
TriviaQA	58.7	42.0
NQ	73.5	68.9
QuAC	33.3	16.1
NewsQA	60.1	52.3

Table 12: With the exception of SQuAD, models continue to return the original answer when given a negated question.

For each dataset, we negated every question in the test set by mapping common verbs (i.e. *is*, *did*, *has*) to their contracted negative form (i.e. *isn't*, *didn't*, *hasn't*) or by inserting *never* before the main verb, as identified by spaCy. We keep the original answers in the test set since we want to evaluate how often the original answer is returned for the negative question. In this case, a lower F1 score means better performance. An example is in the last row of Table 6.

We used the models fine-tuned on their original training sets and evaluated them on the negated test sets. The results are in Table 12 and show how often each model continued to return the original answer given a negative question. We see that SQuAD outperforms all the other models by giving the original answer to a negative question less than 3% of the time. Other models return the original answer to the negative question between 48% and 94% of the time, suggesting that the negation is largely ignored.

Dataset	<i>n't</i>	<i>never</i>
SQuAD	0.85	0.89
TriviaQA	0.31	0.48
NQ	0.37	0.34
QuAC	0.17	0.17
NewsQA	0.14	0.06

Table 13: The percentage of questions in the training set containing *n't* or *never* that are impossible

Does the SQuAD model understand negation, or is this a sign of bias? Table 13 shows how often a question containing *n't* or *never* was impossible in the training set. SQuAD has a high bias, with 85% of questions containing *n't* and 89% of questions containing *never* being impossible. This difference could be a result of SQuAD annotators having a

bias to include *n't* or *never* more often in impossible questions than answerable questions, while impossible questions in other datasets were more organically collected. This suggests that SQuAD's performance is due to an annotation artifact. These results find that no dataset adequately prepares a model to understand negation.

6 Conclusions

In this work, we probed five QA datasets with six tasks and found that our models did not learn to generalize well, remained suspiciously robust to incorrect or missing data, and failed to handle variations in questions. These findings show that models learn simple heuristics around question-context overlap or entity types and pick up on underlying patterns in the datasets that allow them to remain robust to corrupt examples but not to valid variations. The shortcomings in datasets and evaluation methods make it difficult to judge if models are learning reading comprehension. Based on our work, we make the following recommendations to researchers who create or evaluate QA datasets:

- **Test for generalizability:** Models are more valuable to real-world applications if they generalize. New QA model should report performance across multiple relevant datasets.
- **Challenge the models:** Evaluating on too many easy questions can inflate our judgement of what a model has learned. Discard questions that can be solved trivially by high overlap or extracting the first named entity.
- **Be wary of cheating:** Good performance does not mean good understanding. Probe datasets by adding noise, shuffling contexts, or providing incomplete input to ensure models aren't taking shortcuts.
- **Include variations:** Models should be prepared to handle a variety of questions. Include variations such as filler words or negation to existing questions to evaluate how well models have understood a question.
- **Standardize dataset formats:** When creating new datasets, consider following a standardized format, such as SQuAD, to make cross-dataset evaluations simpler.

References

- Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. [Analyzing the behavior of visual question answering models](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.
- Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. [QuAC: Question answering in context](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Sameer Singh, and Matt Gardner. 2019. ORB: An open reading benchmark for comprehensive evaluation of machine reading comprehension. In *EMNLP 2019 MRQA Workshop*, page 147.
- Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *EMNLP 2019 MRQA Workshop*, page 1.
- Wee Chung Gan and Hwee Tou Ng. 2019. [Improving the robustness of question answering systems to question paraphrasing](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*.
- Robin Jia and Percy Liang. 2017. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Nora Kassner and Hinrich Schütze. 2019. Negated LAMA: Birds cannot fly. *arXiv preprint arXiv:1911.03343*.
- Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single QA system. *arXiv preprint arXiv:2005.00700*.
- Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. *arXiv preprint arXiv:2004.14602*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. [Did the model understand the question?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.
- Timothy Niven and Hung-Yu Kao. 2019. [Probing neural network comprehension of natural language arguments](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.
- Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. [BERT with history answer embedding for conversational question answering](#). SIGIR’19, pages 1133–1136, New York, NY, USA. Association for Computing Machinery.

- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. **Know what you don't know: Unanswerable questions for SQuAD**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. **SQuAD: 100,000+ questions for machine comprehension of text**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. **CoQA: A conversational question answering challenge**. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. **Are red roses red? Evaluating consistency of question-answering models**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. **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)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.
- Kyle Richardson and Ashish Sabharwal. 2020. **What does my QA model know? Devising controlled probes using expert knowledge**. *Transactions of the Association for Computational Linguistics*, 8:572–588.
- Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2020. **Benchmarking robustness of machine reading comprehension models**. *arXiv preprint arXiv:2004.14004*.
- Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. **What makes reading comprehension questions easier?** In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.
- Takumi Takahashi, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2019. **CLER: Cross-task learning with expert representation to generalize reading and understanding**. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 183–190, Hong Kong, China. Association for Computational Linguistics.
- Alon Talmor and Jonathan Berant. 2019. **MultiQA: An empirical investigation of generalization and transfer in reading comprehension**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy. Association for Computational Linguistics.
- Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordani, Philip Bachman, and Kaheer Suleman. 2017. **NewsQA: A machine comprehension dataset**. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.
- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. **Universal adversarial triggers for attacking and analyzing NLP**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.
- Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. **Making neural QA as simple as possible but not simpler**. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada. Association for Computational Linguistics.
- Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. **Data augmentation for BERT fine-tuning in open-domain question answering**. *arXiv preprint arXiv:1904.06652*.
- Mark Yatskar. 2019. **A qualitative comparison of CoQA, SQuAD 2.0 and QuAC**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2318–2323, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. **Learning and evaluating general linguistic intelligence**. *arXiv preprint arXiv:1901.11373*.