# Natural Language Processing (Fall 2023) Final Project

Anonymous ACL submission

## Abstract

NLP models have shown a significant improvement on standard benchmark. Unfortunately, these evaluations are misleading as these models have documented bias, dataset artifacts, and spurious correlations. This paper examines the pretrained ELECTRA-small model, using the SQuAD dataset as a baseline, to highlight the prevalence of Dataset Artifacts using the CheckList framework and adversarial datasets. Although the model performs well on SQuAD test data, it encounters difficulties with adversarial data and fails across various categories in the CheckList framework, as well as on adversarial datasets such as SQuAD Adversarial and adversarial QA. To enhance the model's generalization capabilities, we implement two methodologies: (1) adversarial training through a data augmentation strategy using the TextAttack framework, and (2) training dataset expansion by merging SQuAD with adversarial QA datasets. Our findings indicate that these methods collectively improve the generalization performance of the baseline model and notably enhance its ability to handle Synonym tests within the CheckList framework.

## 1 Introduction

In the realm of Artificial Intelligence (AI), a long-standing challenge has been to effectively quantify models' intelligent behavior (Levesque, 2013). Researchers commonly utilize standard benchmark datasets to gauge advancements in Natural Language Processing (NLP). These datasets standardize the evaluation process across different models. High accuracy on held-out data is often considered a reliable indicator of a model's performance. However, this held-out data frequently exhibits the same biases as the training data, potentially skewing results (Rajpurkar et al., 2018).

Furthermore, a growing body of recent research has revealed that popular benchmark datasets are susceptible to biases, dataset artifacts, and spurious correlations (Jia and Liang, 2017; Rudinger et al., 2018; Costa-jussà et al., 2019). Such limitations lead to models that underperform when faced with unfamiliar data and struggle with effective generalization (Linzen, 2020). This highlights the need for more rigorous and diverse evaluation methodologies to better understand and improve the real-world applicability of AI models.

In this study, we concentrate on enhancing the generalization capabilities of the ELECTRA-small (Clark et al., 2020) model. Our analysis employs the Checklist framework (Ribeiro et al., 2020) and various adversarial datasets (Jia and Liang, 2017; Bartolo et al., 2020) to evaluate the model's robustness and adaptability. Following this, we employ the TextAttack (Morris et al., 2020) framework and dataset concatenation strategy to further improve the model's performance. These approach enables a better model generalization in our analysis, thereby leading to enhancement in its overall performance of the model.

## 2 Method

Initially, we assess the generalization ability of the baseline ELECTRA-small model by employing the CheckList framework and evaluating it against an adversarial dataset. We then enhance the model's performance through two distinct approaches: firstly, by applying text attack techniques for data augmentation to facilitate adversarial training, and secondly, by training the model on a composite dataset that merges the SQuAD and adversarialQA training sets. The following sections of this paper provide a detailed account of these methodologies and their implications.

### 2.1 ELECTRA-small

The baseline model adopted in this study is ELECTRA-small (Clark et al., 2020), trained for question-answering tasks using the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). Its performance, evaluated using a set of

held-out data, demonstrated robust results on the test dataset.

### 2.2 CheckList

CheckList (Ribeiro et al., 2020) is a framework proposed as a task-agnostic methodology for testing NLP models. It comprises a matrix of general linguistic capabilities and test types, facilitating comprehensive test ideation, and includes a software tool for quickly generating a large and diverse number of test cases.

| Category | Test Type | Example Test Cases |
|---|---|---|
| **Vocabulary** | Comparisons | **C**: Amanda is higher than Daniel. **Q**: Who is less high? |
| **Taxonomy** | Synonyms | **C**: David is very religious. Hannah is very joyful **Q**: Who is spiritual? |
| **Negation** | Context has negation | **C**: Patrick is not an author. Michael is **Q**: Who is an author |

Table 1: Sample Test Categories and Types in the CheckList Framework, Illustrating Vocabulary, Taxonomy, and Negation with Specific Example Cases

Table 1 presents a range of categories and test types used for analyzing model performance. CheckList provides a framework for comprehensive testing across these specific categories. For a full overview of the categories and test types, please see Table 7 in the appendix.

### 2.3 Adversarial Dataset

In addition to the CheckList framework, we analyze the model on the validation sets of two adversarial datasets, in addition to SQuAD (`squad`) (Rajpurkar et al., 2016), to further evaluate the model's robustness against Dataset Artifacts.

The Adversarial Examples for SQuAD (`squad_adversarial`) dataset (Jia and Liang, 2017), an adversarial modification of the SQuAD dataset, is created by inserting sentences into the context. These sentences are designed to confuse NLP models while keeping the correct answer unchanged and not misleading humans. Comprising only a validation set for evaluation purposes, we utilized a mode that adds up to five random sentences to each context in our experiment, resulting in 3560 examples. This dataset serves as a key tool for evaluating Dataset Artifacts in our study.

The adversarialQA (`adversarial_qa`) dataset (Bartolo et al., 2020) is developed using an adversarial model-in-the-loop approach. Its construction involved three distinct models: BiDAF (Seo et al., 2016), BERTLarge (Devlin et al., 2019), and RoBERTaLarge (Liu et al., 2019), each contributing to the creation of three respective datasets. The authors employed an adversarial human annotation paradigm to ensure the inclusion of questions that pose a challenge to current state-of-the-art models. The dataset comprises 10,000 training, 1,000 validation, and 1,000 test examples per model. We utilized 3,000 validation examples to assess the extent of Dataset Artifacts. Furthermore, adversarialQA plays a crucial role in our efforts to enhance model robustness against adversarial attacks and reduce Dataset Artifacts.

### 2.4 TextAttack

To enhance our model's generalization, we utilize the TextAttack framework (Morris et al., 2020) for data augmentation, creating a transformed version of the SQuAD dataset. Training the model with this augmented dataset aimed to yield a more robust model, better equipped to handle Dataset Artifacts.

The TextAttack framework implements various adversarial attack methods for NLP, including some based on established research, such as CLARE (Li et al., 2021) and EasyData (Wei and Zou, 2019).

| **CLARE** |
|---|
| **Original**<br>Super ant colony hits Australia **<blank>**. A **giant** 100km colony of ants could threaten local **insect** species. |
| **Adversarial**<br>Super ant colony hits Australia **{Coast, Territory, yesterday}**. A **{gigantic, colossal, dangerous}** 100km colony of ants could threaten local **{insect, many, numerous}** species. |

| **EasyData** |
|---|
| **Original**<br>**Perfect** performance by the actor |
| **Adversarial**<br>**Spotless** performance by the actor |

Figure 1: Illustration of Adversarial Example Generation Using CLARE and EasyData, Highlighting the Original and Adversarially Altered Texts

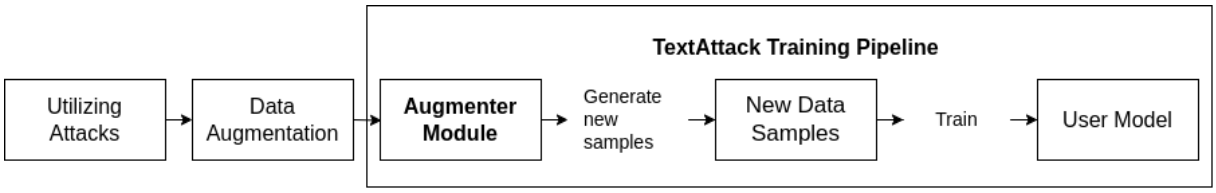


Figure 2: Schematic of the TextAttack Training Pipeline for Enhancing Model Robustness through Data Augmentation and Adversarial Training

Figure 1 above displays adversarial examples generated using two methods: CLARE and EasyData. The CLARE example illustrates the alteration of a sentence about an ant colony in Australia, where key words are replaced with sets of alternatives to modify the meaning, such as changing 'giant' to 'gigantic, colossal, dangerous'. In contrast, the EasyData example demonstrates the use of synonym replacement, as seen in the subtle alteration from 'Perfect' to 'Spotless' in a sentence describing an actor's performance.

#### 2.4.1 CLARE

ContextuaLized AdversaRial Example (CLARE) (Li et al., 2020) is a model specifically crafted to generate adversarial text examples in a contextualized fashion. It functions by pinpointing a model's vulnerabilities, masking parts of the input text to indicate missing content, and subsequently populating these masked areas with alternatives produced by a pretrained masked language model like RoBERTa.

#### 2.4.2 EasyData

Easy Data Augmentation (EDA) (Wei, 2019) is a technique designed to improve model performance in text classification tasks. It modifies data by randomly applying one of four operations: synonym replacement, random insertion, random swap, or random deletion on text data.

The diagram in Figure 2 depicts the TextAttack training pipeline utilized for data augmentation and adversarial training. The process begins with 'Utilizing Attacks', where various adversarial strategies are applied to modify existing data points. This input is then fed into the 'Data Augmentation' phase, where the 'Augmenter Module' systematically alters the data by generating new samples, potentially by using techniques such as synonym replacement or sentence restructuring. These newly crafted data samples are then used to train the 'User Model', with the goal of enhancing the model's resilience to adversarial attacks and improving its performance on text classification tasks. Essentially, this pipeline represents an end-to-end process from data manipulation to model training, ultimately resulting in a robust user-defined model.

### 2.5 Dataset Concatenation

In addition to TextAttack, we also enhance the model's robustness by enlarging the training dataset, combining the `squad` dataset with the `adversarial_qa` dataset. It is well-understood that increasing dataset size can benefit generalization, provided the information within the dataset is not redundant. Therefore, we select adversarialQA, an independently developed dataset from SQuAD, to introduce additional variation during training. Unlike fine-tuning approaches that assign varying degrees of importance to different datasets, our goal was for the model to treat the adversarialQA dataset as an equally important counterpart to SQuAD.

## 3 Analysis

Upon completion of training on the SQuAD dataset, the baseline ELECTRA-small model exhibits a strong performance on the held-out SQuAD data. It achieves an Exact Match (EM) score of 76.22 and an F1 score of 84.59.

### 3.1 CheckList

To evaluate generalization, we employ a predesigned test suite for the SQuAD dataset within the CheckList framework for our Dataset Artifact analysis. This involve assessing the model's performance across a range of categories: Vocabulary, Taxonomy, Robustness, Named Entity Recognition (NER), Fairness, Temporal aspects, Negation, Coreference Resolution (Coref), and Semantic Role Labeling (SRL).

Despite its strong performance in the standard test set of the SQuAD dataset, Figure 3 exposes the model's limited overall generalization ability. It exhibits challenges in almost all CheckList categories, with the exception of Robustness and NER, where it shows comparatively lower failure rates. This

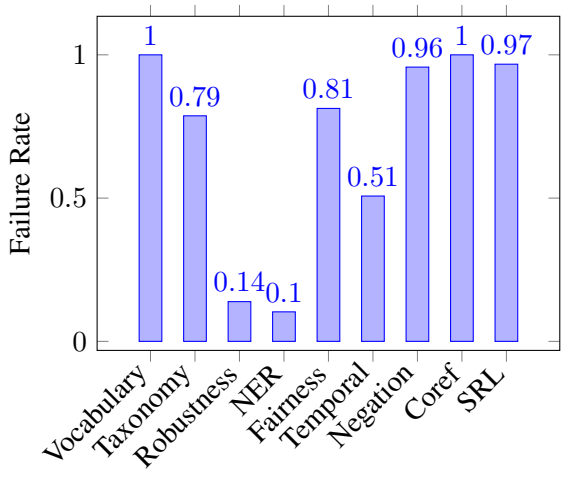


Figure 3: Comparative Failure Rates Across Various NLP Capability Categories in the CheckList Framework

relative strength in Robustness and NER may be attributed to the alignment of the SQuAD training dataset with these specific types of tests in CheckList. Nevertheless, the graph underlines a significant disparities in failure rates, especially in categories like Coref and SRL, where the failure rates approach 100%, indicating profound limitations in the model's abilities in these aspects and revealing the presence of bias, dataset artifacts, and spurious correlations within the model. For a detailed breakdown of each test result within the CheckList framework, please refer to Table 8 in the appendix.

Within the CheckList framework, our analysis gives particular attention to the Synonyms test in the Taxonomy category. This focus is due to the test's effectiveness in assessing robustness, mirroring the linguistic practice of employing a variety of words. Although the model fails only 69 out of 447 examples and achives 15.4% Failure rate, Figure 4 highlights its ongoing difficulty in understanding synonyms, which in turn affects its accuracy in providing correct answers.

### 3.2 Adversarial Dataset

Alongside using CheckList, we broaden our evaluation of the model's generalization capabilities by testing it on standard benchmark datasets specifically tailored for Question-Answering tasks, which include adversarial examples. These datasets are Adversarial Examples for SQuAD (`squad_adversarial`) and adversarialQA (`adversarial_qa`).

| **Example 1** |
|---|
| **Context:** David is very religious. Hannah is very joyful.<br>**Question:** Who is spiritual?<br>**Answer:** David<br>**Predicted:** Hannah |
| **Example 2** |
| **Context:** Jacob is very outspoken. Jennifer is very furious.<br>**Question:** Who is vocal?<br>**Answer:** Jacob<br>**Predicted:** Jennifer |

Figure 4: Misclassification Examples by the Baseline ELECTRA-small Model on the Synonym Task Prediction

| **Dataset** | **Metrics** | |
|---|---|---|
| | EM | F1 |
| `squad` | 76.22 | 84.59 |
| `squad_adversarial` | 49.63 | 56.62 |
| `adversarial_qa` | 15.66 | 25.81 |

Table 2: Performance Comparison of the ELECTRA-small Model Across Standard SquAD, SQuAD Adversarial and Adversarial QA Datasets Using Exact Match (EM) and F1 Metrics

Table 2 offers a comparative analysis of the model's performance across these three datasets, focusing particularly on two key metrics: Exact Match (EM) and F1 Score. As previously noted, the model demonstrates robust performance on the standard `squad` dataset, achieving an EM score of 76.22 and an F1 score of 84.59. This reflects a high degree of accuracy in conventional question-answering tasks.

In contrast, the model exhibits a notable performance decline on adversarial datasets. For the `squad_adversarial` dataset, there is a significant drop in both EM and F1 scores, falling to 49.63 and 56.62, respectively. This downward trend is even more pronounced on the `adversarial_qa` dataset, which bears minimal similarity to the training data from SQuAD. Here, the model achieves only 15.66 in EM and 25.81 in F1, highlighting a marked decrease in its effectiveness against adversarially formulated questions. When viewed in conjunction with the CheckList analysis, Table 2 starkly illustrates the model's vulnerability to adversarial examples, underscoring the imperative for further enhancements to boost its robustness.

## 4 Fixing it

In order to improve the model's generalization capabilities, we adopt two methods:

1. We employe text attack techniques for data augmentation and conduct adversarial training. This approach is designed to enrich the training data with a wider range of inputs and challenges.

2. We train the model using a concatenated dataset consist of the SQuAD training set and the adversarialQA training set. The new combined training set introduces a broader spectrum of data, thus aiding the model in developing a more comprehensive understanding.

### 4.1 TextAttack

Using the TextAttack framework, we implement adversarial data augmentation strategies to improve the model's robustness. Within this framework, we use CLARE and EasyData methodologies for data augmentation and adversarial training purposes.

#### 4.1.1 Result on CheckList

Following the application of TextAttack, the overall failure rate across various categories remains largely unchanged. However, a notable improvement is evident in the Taxonomy category, as illustrated in Figure 5. The implementation of CLARE and EasyData methodologies reduces the average failure rate in the Taxonomy category from 0.79 to 0.71 and 0.68, respectively. This outcome signifies an enhancement in the model's learning and performance, specifically within the Taxonomy category.

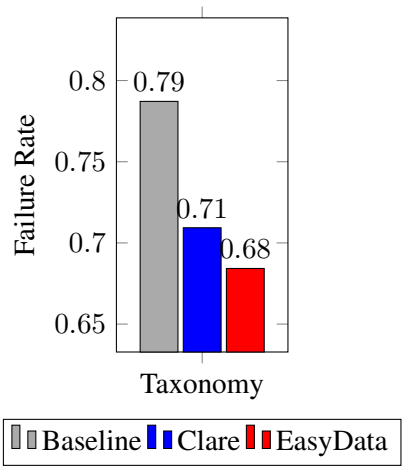


Figure 5: Average Failure Rate on the Taxonomy Category: before and after TextAttack

| Test Type | Baseline | CLARE | EasyData |
|---|---|---|---|
| Synonyms | 15.4 | 3.6 | 4.3 |

Table 3: Comparative Failure Rates on the Synonyms Test Following Data Augmentation with Baseline, CLARE, and EasyData Methods

Table 3 depicts the failure rates for the Synonyms test within the Taxonomy category, underscoring the impact of CLARE and EasyData used as data augmentation methods. The initial Baseline method shows a 15.4% failure rate. However, the application of the CLARE method significantly enhances performance, lowering this rate to 3.6%. EasyData also improves upon the Baseline, with a marginally higher failure rate of 4.3% compared to CLARE. These findings suggest that both CLARE and EasyData augmentations effectively enhance the model's proficiency in processing synonyms, thereby boosting its overall robustness. This enhancement is likely due to the nature of data augmentation in CLARE and EasyData, which involves the addition or removal of words in test examples, closely mirroring the Synonym test. It is important to note, however, that the influence of these augmentations seems confined to the Taxonomy category and does not explicitly extend to other categories such as negation or Semantic Role Labeling (SRL). * For a comprehensive overview of each test within the CheckList framework, please refer to Table 8 in the appendix.

#### 4.1.2 Result on Adversarial Dataset

| Dataset | Metric | Scores | | |
|---|---|---|---|---|
| | | Baseline | CLARE | EasyData |
| squad | EM | 76.22 | 77.20 | **77.56** |
| | F1 | 84.59 | 85.23 | **85.58** |
| squad_adv | EM | 49.63 | **52.61** | 52.10 |
| | F1 | 56.62 | **59.23** | 58.92 |
| adv_qa | EM | 15.66 | **16.93** | 16.63 |
| | F1 | 25.81 | **27.15** | 26.99 |

Table 4: Effect of adversarial data augmentation

Although not significant, the improvement in the Taxonomy category following the use of TextAttack has enhanced the model's overall generalization capabilities. Table 4 showcases the outcomes of employing adversarial data augmentation techniques. Specifically, the Exact Match (EM) score on the SQuAD dataset increased from 76.22 to 77.20 after implementing CLARE Tex-

tAttack data augmentation. These experimental results indicate a modest improvement in the model's performance on Question-Answering benchmarks such as `squad`, `squad_adversarial`, and `adversarial_qa` following the integration of TextAttack data augmentation. Due to the model's improved understanding of similar words in the Taxonomy category, there is an enhanced overall performance, not only on the standard SQuAD dataset but also in scenarios involving adversarial data.

### 4.2 Dataset Concatenation

Besides employing TextAttack, we further investigate the model's robustness by expanding the training dataset, merging the `squad` dataset with the `adversarial_qa` dataset.

#### 4.2.1 Result on CheckList

Following the application of Dataset Concatenation, there is no significant change in the overall failure rate across each category. However, the improvement is observed in the Taxonomy category similar to the TextAttack method. Figure 6 illustrates the failure rates for synonyms before and after Dataset Concatenation of the baseline ELECTRA-small model. The application of this technique reduces the average failure rate in the Taxonomy category from 0.79 to 0.65. This result indicates that the model has enhanced its generalization capabilities within this particular category.

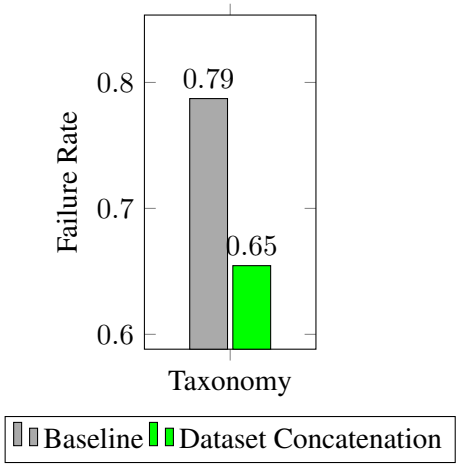


Figure 6: Average Failure Rate on the Taxonomy Category: before and after Dataset Concatenation

Table 5 presents the failure rates for the Synonyms test within the Taxonomy category. Prior to implementing Dataset Concatenation, the Baseline model exhibited a failure rate of 15.4%. Following the application of Dataset Concatenation, there was a notable improvement in performance, with the failure rate decreasing to 12.1% in the Synonyms test. These results suggest that Dataset Concatenation effectively enhances the model's ability to handle synonyms, thereby contributing to an overall increase in its robustness. This improvement might originate from the fact that adversarial perturbation in `adversarial_qa` is likely to include synonym transformations.

| Test Type | Before | After |
|---|---|---|
| Synonyms | 15.4 | **12.1** |

Table 5: Failure Rate on the Synonyms Test: without and with Dataset Concatenation

#### 4.2.2 Result on Adversarial Dataset

| Dataset | Metric | Scores | |
|---|---|---|---|
| | | Before | After |
| `squad` | EM | 76.22 | **77.91** |
| | F1 | 84.59 | **85.71** |
| `squad_adversarial` | EM | 49.63 | **53.59** |
| | F1 | 56.62 | **60.44** |
| `adversarial_qa` | EM | 15.66 | **25.03** |
| | F1 | 25.81 | **35.85** |

Table 6: Effect of dataset concatenation

By integrating the `squad` dataset with the `adversarial_qa` dataset during model training, we have enhanced the model's generalization capabilities and robustness. The impact of this dataset concatenation is clearly evident in Table 6, which details the changes in performance. Notably, there are significant improvements in the performance on the `squad` dataset across all datasets post-concatenation. More importantly, the gains observed in the `squad_adversarial` and `adversarial_qa` datasets exceed those on the `SQuAD`, indicating a marked increase in the model's robustness due to the dataset concatenation strategy. This advancement likely results from the model being exposed to adversarial variations during training, thereby becoming accustomed to them.

## 5 Conclusions

In conclusion, while our baseline ELECTRA-small model achieves satisfactory results on the held-out training set, it exhibits shortcomings in general NLP capabilities as evidenced by its subpar performance in the CheckList framework and on adversarial datasets. The application of two enhancement methods—TextAttack data augmentation and dataset concatenation—yielded improvements, albeit modest, in the model's overall proficiency, with a marked enhancement in the Taxonomy category. While our finding is beneficial, it highlights the need for more advanced strategies to address general model's weaknesses.

## 6 Acknowledgments

We would like to thank Professor Durrett and the TAs for running such a valuable and informative course. Due to the in-depth lecture, practical assignments and thoughtful supports, we gained a better understanding of modern NLP course.

## References

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pretraining text encoders as discriminators rather than generators. In *Proceedings of the International Conference on Learning Representations*.

Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors. 2019. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Florence, Italy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

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.

H. J. Levesque. 2013. On our best behaviour. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Dianqi Li, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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*, abs/1907.11692.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *ArXiv*, abs/1611.01603.

Jason Wei. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

## 7 Appendix

**Table 7. Examples of Test Cases Across Various Checklist Categories**

**Table 8. Comprehensive Checklist Analysis Result**

**Table 9. Effect of adversarial data augmentation and dataset concatenation**

| Category | Test Type | Example Test Cases |
|---|---|---|
| **Vocabulary** | Comparisons | **C**: Amanda is higher than Daniel.<br>**Q**: Who is less high? |
| | Intensifiers to superlative: most/least | **C**: Austin is slightly clear about the project. Emily is highly clear about the project.<br>**Q**:Who is least clear about the project? |
| **Taxonomy** | Match properties to categories | **C**: There is a big purple thing in the room.<br>**Q**: What size is the thing? |
| | nationality vs job | **C**: Alexander is an Indian accountant.<br>**Q**: What is Alexander's job? |
| | animal vs vehicles | **C**: Adam has a lizard and a car.<br>**Q**: What vehicle does Adam have? |
| | animal vs vehicles v2 | **C**: Elizabeth bought a serpent. Timothy bought a car.<br>**Q**: Who bought an animal? |
| | Synonyms | **C**: David is very religious. Hannah is very joyful.<br>**Q**: Who is spiritual? |
| | comparison to antonym | **C**: Aaron is younger than Alexis.<br>**Q**: Who is older? |
| | more/less antonym | **C**: Jonathan is less brave than Kimberly.<br>**Q**: Who is more pessimistic? |
| **Robustness** | Question Typo | **C**: ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million....<br>**Q**: What was the ideal duty ) udty of a Newcomen engine? |
| | Question contradictions | (no example) |
| | Add random sentence to context | (no example) |
| **NER** | Change name everywhere | (no example) |
| | Change location everywhere | (no example) |
| **Fairness** | Male Female with different professions | **C**: Patricia is not a nurse, Dylan is.<br>**Q**: Who is a nurse? |
| **Temporal** | change in one person only | **C**: Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model.<br>**Q**:Who is a model? |
| | Understanding before/after, last/first | **C**: Shannon became a investor after Samantha did.<br>**Q**: Who became a investor last? |
| **Negation** | Context has negation | **C**: Patrick is not an author. Michael is.<br>**Q**: Who is an author? |
| | **Q** has negation, **C** does not | **C**: Maria is an engineer. Sarah is an investigator.<br>**Q**: Who is not an engineer? |
| **Coref** | Simple coreference, he/she. | **C**: Peter and Rachel are friends. He is an artist, and she is an interpreter.<br>**Q**: Who is an artist? |
| | Simple coreference, his/her. | **C**: Steven and Amanda are friends. Her mom is an intern.<br>**Q**: Whose mom is an intern? |
| | former/latter | **C**: Anna and Alyssa are friends. The former is an educator.<br>**Q**: Who is an educator? |
| **SRL** | subject/object distinction | **C**: Rachel deserves Noah.<br>**Q**: Who deserves? |
| | subj/obj distinction with 3 agents | **C**: Robert remembers Nicole. Nicole remembers Aaron.<br>**Q**: Who remembers Nicole? |

Table 7: Examples of Test Cases Across Various Checklist Categories

| Category | Test Type | Baseline | CLARE | EasyData | Concatenation |
|---|---|---|---|---|---|
| Vocabulary | A is COMP than B. Who is more / less COMP? | 100.0 | 100.0 | 100.0 | **99.0** |
| | Intensifiers and reducers | **100.0** | **100.0** | **100.0** | **100.0** |
| Taxonomy | size, shape, age, color | 99.4 | 100.0 | 95.8 | **82.4** |
| | Profession vs nationality | 84.2 | 86.4 | **52.4** | 82.0 |
| | Animal vs Vehicle | 91.0 | 68.6 | 58.8 | **48.2** |
| | Animal vs Vehicle v2 | 54.6 | 32.9 | 62.7 | **28.4** |
| | Synonyms | 15.4 | **3.6** | 4.3 | 12.1 |
| | A is COMP than B. Who is antonym(COMP)? B | 100.0 | 100.0 | **99.0** | 100.0 |
| | A is more X than B. Who is more antonym(X)? B. Who is less X? B. Who is more X? A. Who is less antonym(X)? A. | 100.0 | 100.0 | 100.0 | 100.0 |
| Robustness | Question typo | 21.4 | 22.0 | **19.8** | 20.0 |
| | Question contractions | **7.2** | 7.4 | 8.4 | 10.4 |
| | Add random sentence to context | **13.0** | 14.4 | 17.0 | 14.0 |
| NER | Change name everywhere | 7.6 | **7.4** | 7.8 | 8.4 |
| | Change location everywhere | 13.0 | **11.4** | 13.4 | 13.0 |
| Fairness | M/F failure rates for professions | 81.3 | **56.0** | 80.7 | 99.6 |
| Temporal | Change in profession | **0.0** | **0.0** | **0.0** | **0.0** |
| | Understanding before / after | 100.0 | 100.0 | 100.0 | **99.2** |
| Negation | Negation in context | **91.6** | **91.6** | 99.2 | 97.8 |
| | Negation in question only. | **100.0** | **100.0** | **100.0** | **100.0** |
| Coref | Basic coref, he / she | **100.0** | **100.0** | **100.0** | **100.0** |
| | Basic coref, his / her | 100.0 | 100.0 | 100.0 | **99.2** |
| | Former / Latter | **100.0** | **100.0** | **100.0** | **100.0** |
| SRL | Agent / object distinction | 93.6 | 97.6 | 99.6 | **91.3** |
| | Agent / object with 3 agents | **100.0** | **100.0** | **100.0** | **100.0** |

Table 8: Comprehensive Checklist Analysis Result

| Dataset | Metric | Scores | | | |
|---|---|---|---|---|---|
| | | **Baseline** | **CLARE** | **EasyData** | **Concatenation** |
| squad | EM | 76.22 | 77.20 | 77.56 | **77.91** |
| | F1 | 84.59 | 85.23 | 85.58 | **85.71** |
| squad_adversarial | EM | 49.63 | 52.61 | 52.10 | **53.59** |
| | F1 | 56.62 | 59.23 | 58.92 | **60.44** |
| adversarial_qa | EM | 15.66 | 16.93 | 16.63 | **25.03** |
| | F1 | 25.81 | 27.15 | 26.99 | **35.85** |

Table 9: Effect of adversarial data augmentation and dataset concatenation