Reducing Short Circuits in Multiple-Choice Natural Language Reasoning Models with Data Augmentation

Anonymous ACL submission

Abstract

Statistical biases in the training data may lead to fragility in neural models that makes choices in multiple-choice natural language reasoning problems without referring to the context or premises. To encourage the models to pay more attention to the relations between the premise and the choices, we propose two biologically inspired operations that can generate new training data that "forces" the model to look at the premises and reducing short circuits. They can augment any type of multiple choice reasoning dataset, and can be applied to any supervised learning models. Results show that models trained with the augmented data become more robust against both stress test and original test.

1 Introduction

Large-scale neural networks have been applied extensively to natural language reasoning (NLR) tasks such as causal reasoning (Gordon et al., 2012), story ending prediction (Mostafazadeh et al., 2017), argument reasoning comprehension (Habernal et al., 2017), and reading comprehension (Yu et al., 2020). Many of the current benchmarks of these NLR tasks take the form of multiple-choice questions (MCQs) which are made up of a premise and two or more choices. Below is an example taken from COPA (Gordon et al., 2012), which tests commonsense causal reasoning.

Example 1 *An MCQ from COPA:*

Premise: *The man hurt his back.*
Choice 1: *He stayed in bed for several days.* ✓
Choice 2: *He went to see a psychiatrist.* ✗

Usually, models are trained on the training data and tested with the standard validation-test split paradigm. While accuracy on held-out data is a useful indicator, held-out datasets are often not diverse enough and may contain the same biases in the training data (McCoy et al., 2019). Furthermore, as simple aggregated statistics, accuracy on the test set doesn't show the robustness of the model, or why a question is answered correctly. There has been speculation (Sharma et al., 2018; Zellers et al., 2018) that many models did not really "understand" the semantical and logical connection between the premise and the choices, but do well only due to spurious statistical features in the choices, which means the models are actually fragile.

Such fragility can be observed by both white-box and black-box tests. In a white-box test (Vig, 2019), attention map between the words in the full question from the final encoder layer of the model can reveal the connection, or the lack of one, between the premise and the choices. Figure 1, which is a plot for Example 1, clearly shows that there's virtually no connection between the first choice and the premise (highlighted by the red box) when BERT is processing the full question. While the attention between the words within the first choice remains the same when the model processes only the choices without the premise. We call such a phenomenon "short circuit" in multiple-choice NLR in this paper.

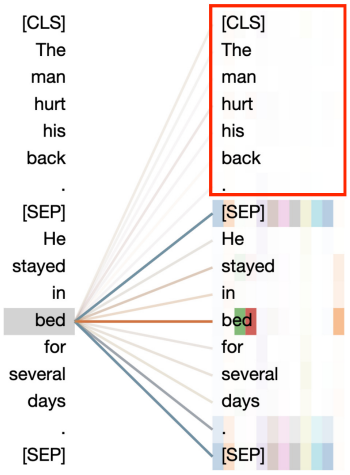


Figure 1: Attention map showing that BERT short-circuits on a COPA question.

Furthermore, two kinds of black-box tests have

been attempted. One is called "ending-only tests" in some literature (Sharma et al., 2018; Bras et al., 2020), which we refer to as "choice-only test" here since our focus is on MCQs. For example, BERT, when fine-tuned on the COPA data, can answer the question in Example 1 correctly. When we remove the premise from the same question and feed it to the model, it still gets the correct answer (Choice 1). This result from the "choice-only" test seems to suggest that the model can make correct predictions without even looking at the premise. The other blackbox test is a kind of stress test (Ribeiro et al., 2020), which tests if the model is short-circuiting toward (or against) certain linguistic features such as named entities, typos, and negations. In this work, we apply many stress test cases of several categories, and observe that many models are fragile with low accuracies. Through the above tests, we are able to confirm that three popular deep models, i.e., BERT, XLNet (Yang et al.) and RoBERTa (Liu et al., 2019c), when applied to multiple-choice NLR, all suffer from the "short-circuit" problem.

One straightforward way to reduce the model short circuits is to train the models with hard cases that look like the stress tests. However, many stress tests are constrained by the way choices are constructed, which limits the quantity of cases to automatically generate, and consequently their ability to serve as general data augmentation methods. Besides, most of the stress tests are feature specific and hard to generalize. To this end, we propose *crossover* and *mutation* operators, which can easily generate abundant data and encourage models to pay more attention to the premise. We apply crossover and mutation to augment the three models on ROC (Mostafazadeh et al., 2017), COPA, ARCT (Habernal et al., 2017), and RECLOR (Yu et al., 2020) and see up to 42% increase in accuracy on the stress tests and 10% increase in the original test data, beating the previous strong baseline back-translation (Xie et al., 2019).

This paper makes the following contributions: i) we propose the crossover and mutation operations to augment training data that teaches the models to pay attention to the premises in questions; ii) experiments show that the augmented models perform substantially better on diverse stress tests while maintaining their accuracies on the original tests, demonstrating reduced short circuits; iii) we provide evidence to show that method indeed reduces short-circuits in these models, thus confirming the validity of our approach.

2 Approach

We first present stress test operators which are used to create stress test cases for measuring model fragility, then propose two novel operations to augment training data to reduce short circuits in models.

2.1 Stress Test Operators

To evaluate the extent of model short circuits, we create these stress test cases using the operators in Table 1. Most of the operators have been proposed previously (Ribeiro et al., 2020), except for PR and PI, which are newly introduced in this work. We create a stress test instance from a specific MCQ by keeping the right choice and creating a **wrong** choice by applying one of the stress operators to the original right choice. This new wrong choice is *grammatically correct* but *logically incorrect* under the particular context. Besides, the new wrong choice contains the same content as the right choice except for the tested feature. These stress test cases with similar choices can evaluate whether models are considering the connection between the premise and choices, or in other words "short-circuiting." For example, if a model doesn't comprehend the consistency of the NER name "Mary" in Table 3 which has been mentioned in both the premise context and the right choice, then the model can be easily confused by a similar wrong choice.

Oper.	Description and Example
Neg+	Add negation (r→w) Input: *They called the police to come to my house.* ✓ Output: *They **didn't** call the police to come to my house.* ✗
Neg-	Remove negation (r→w) Input: *Ben **never** starts working out.* ✓ Output: *Ben starts working out.* ✗
NER	Randomly replace person names (r→w) Input: *A big wave knocked **Mary** down.* ✓ Output: *A big wave knocked **Kia** down.* ✗
PR	Switch pronoun by gender or quantity (r→w) Input: ***She** had a great time.*✓ Output: ***He** had a great time.* ✗
PI	Instantiate pronoun by random person (r→w) Input: ***They** gave Tom a new latte with less ice.* ✓ Output: ***Nathanael** gave Tom a new latte with less ice.* ✗
Voice	Swap subject and object (r→w) Input: ***Kara** asked **the neighbors** not to litter in their yard.* ✓ Output: ***the neighbors** asked **Kara** not to litter in their yard.* ✗

Table 1: Stress test operators considered in this paper. The first line in each cell describes the operation, and the remaining lines in the cell give examples of how the operators work. r→w indicates the operator turns a right choice into a wrong choice.

2.2 Improving Model Robustness by Data Augmentation

To decrease short circuits in models, one natural thought is to generate more challenging training data that forces the models to focus on the relationship between the premise and choices. A straight forward way to do this is to modify existing training data so that the right and wrong choices become similar to each other. For example, we could modify the right choice in an MCQ by changing the named entity in it into another arbitrarily named entity and thus create a new wrong choice that shares most of the features with the right choice. But such an approach has two challenges: i) the modification operations are typically restricted to a *particular linguistic feature*, and hence it is hard to create data with good coverage of diverse linguistic features; and ii) because some of the linguistic features may be sparse in the training data, the number of additional questions generated using these features may be very limited, which means the data augmentation doesn't scale.

To overcome the above issues, we propose two genetically inspired operators, namely *crossover* and *mutation*, which are more scalable and universally applicable for data augmentation of any kind of MCQs. These two operators are not only simple but also not limited to fine-grained features.

2.2.1 Crossover

Crossover is illustrated in Figure 2. It operates on two randomly selected MCQs. We substitute the wrong choice of one MCQ with the right choice from the other MCQ to generate a new MCQ. The substituted choice is almost certainly wrong in the new MCQ. For example, the green choice in the original question B is the right choice, but wrong for the original question A. With this rule, we can get two augmented questions: augmented question A and augmented question B. We only consider swapping the right choices between two questions rather than the wrong ones. This is because if the model was short-circuiting, then it is likely to rely on some spurious features correlated with the true label in the right choices. By substituting these right choices into another question to make them wrong, this operation can disrupt such correlations. Hence, to tell if one choice is better than the other, the model is encouraged to consider the premise.

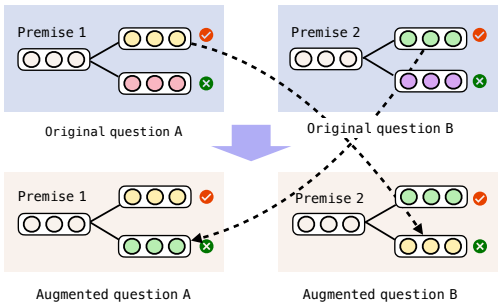


Figure 2: Crossover: the rights choice of both questions are used to replace the wrong choices of these questions to create two new questions. Circles symbolize tokens in the sentences.

2.2.2 Mutation

Mutation is illustrated in Figure 3. It is also designed to teach models to pay more attention to the relationship between the premise and the choices. Different from *crossover* which makes the choices very different, *mutation* makes the two choices of a question very similar except for the order of the words. This forces the model to look to the premise to avoid short-circuit problems.

We reserve the right choice and augment data by changing the wrong choice. *Mutation* operation swaps two consecutive words [1] either in the right choice or wrong choice of the original MCQ, each with 50% probability, to make a new wrong choice. The *mutation* operator should not be confused with the random token swapping (RS) operator (Artetxe et al., 2017; Lample et al., 2017). RS seeks to perturb a choice in the question without changing its label, whereas mutating the right choice (*mutation*) converts it to a wrong choice because the perturbed choice is "less right" than the original one. Mutating the wrong choice has a similar effect as RS. Intuitively, mutation has the potential to reduce short circuits: it not only encourages the model to look into the premise due to its two very similar choices (see the two choices of A1 in Figure 3), but also makes the model more sensitive to the differences in word orders and enhances the model's pre-existing grammatical knowledge (see the two choices in A2).

3 Experiments

We evaluate the effectiveness of the proposed augmentation methods on four popular natural language reasoning tasks. Three transformer-based models are employed as the main targets for our

[1] The words are tokenized with NLTK.

Dataset	Premise	Choices	Training size	Test size
COPA	I pushed the door.	The door opened. ✓ The door locked. ✗	500	500
ROC	Sarah was home alone. She wanted to stay busy. She turned on the TV. She found a reality show to watch.	Sarah then happily watched the show. ✓ Sarah could not find anything to watch. ✗	1871	1871
ARCT	**Reason**: Milk isn't a gateway drug even though most people drink it as children. **Claim**: Marijuana is not a gateway drug.	**Warrant 1**: Milk is similar to marijuana. ✓ **Warrant 2**: Milk is not marijuana.✗	1210	444
RECLOR	**Context**:In a business...to financial prosperity. **Question**:The reasoning in the argument is flawed because the argument	A: ignores the fact that in... the family 's prosperity.✓ B: presumes, without... the family's prosperity.✗ C: ignores the fact... even if they pay high wages.✗ D: presumes, without providing...can succeed.✗	4638	500

Table 2: Examples for all 4 datasets considered in this paper.

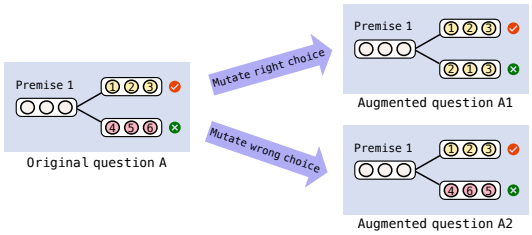


Figure 3: Mutation: the right choice of a question is used to replace the wrong choices of this question to create new questions. Circles symbolizes tokens.

experiments. We first show the experimental setup. Then, we compare different augmentation methods on three models by the end-to-end tests, which contain the stress test and original test of the four datasets, and demonstrate the advantage of crossover and mutation. After that, we apply choice-only tests on the same set of models compared in the last step, to reconfirm that performance gain in the end-to-end tests is due to the reduction of short-circuit problems. Finally, we use a case study to discover the reason for the model improvement by the white-box test.

3.1 Experimental Setup

3.1.1 Datasets

We experiment on 4 datasets from four different tasks:

ROC is a story ending prediction dataset. The task is to identify the correct ending of a four-sentence story premise from two alternative choices.

COPA is a causal reasoning dataset, an example is previously shown in Section 1. Given a premise, COPA requires choosing the more plausible, causally related choice.

ARCT is an argument reasoning comprehension dataset. There may exist an alternative warrant choice in which the reason is connected to the claim.

RECLOR is a reading comprehension dataset that requires logical reasoning.

Examples and statistics of them are shown in Table 2.

3.1.2 Stress Test Cases

Stress	ROC	COPA	ARCT	RECLOR
Neg+	1,797	492	297	375
Neg-	94	2	152	119
NER	362	0	5	0
PR	1,073	328	71	72
PI	861	219	56	91
Voice	1,014	246	174	263
Total	8,943	2,287	1,643	1,920

Table 3: Number of stress test cases generated by different operators for the four datasets.

Different operators generate different but sufficient number of cases as shown in Table 3.

To guarantee the correctness of questions in the stress test, we sample 100 stress cases generated by each operation and annotate whether the cases are correct or not. The correctness of these questions is 100% which indicates the reliability of these stress tests.

We evaluate the effectiveness and short circuits of all data augmentation methods by the accuracy of the stress test set and the original test set.

3.1.3 Models

We investigate three popular pre-trained language models: **BERT**, **RoBERTa**, and **XLNet**. To fine-tune the language models for an MCQ task, we feed LM's final hidden vector to an MLP to compute the probability of the right choice. We conduct all experiments on a server: a GeForce GTX 1080Ti

GPU with 11G RAM and Intel(R) Xeon(R) CPU E5-2630 with 128G of RAM.

Besides the original models (marked as w/o), we also train these three models with four competing data augmentation methods: back-translation (Xie et al., 2019) (B), crossover (C), mutation (M), and the mixture of crossover and mutation with equal proportion (C+M). For each MCQ in the original training set, we create a new question using either one of these 4 methods, yielding 4 augmented training sets the same size as the original one.

We use back-translation as our baseline because it is popularly used in NLU tasks. While there exist promising data augmentation methods (Qu et al., 2020; Chen et al., 2021) that are based on dynamic perturbation of hidden states, back-translation is by far the most effective data augmentation method that operates on the input level. To this end, we generate a new question by conducting a round-trip English-to-French and French-to-English translation over each wrong choice. The translation model we utilized is mBART.

Since crossover and mutation are operators for data augmentation, the modified questions do not need to be strictly correct. We also sampled 100 cases for each operator. 98% and 97% of the cases turned out to be correct for *crossover* and *mutation*.

To ensure fairness, the training data augmented with +B, +C, +M, and +C+M are all the same size. In +C+M, the extra data by +C and +M are equal in size.

3.2 End-to-end Test

In this subsection, we explore the capabilities of models with different data augmentation methods, i.e., back-translation, *crossover*, and *mutation*, from overall and fine-grained perspectives. The overall perspective shows the accuracy results from the stress test set and the overall original tests. Fine-grained perspective shows the stress test accuracy results by different stress operators. We train each model 3 times with different seeds and calculate their average score as the reusults of each test.

3.2.1 Overall results

In Table 4 and Table 5, we can find that vanilla BERT, XLNet, and RoBERTa are mostly not robust on stress tests across all datasets. Compared to the original test data, the accuracy on the stress tests has dropped substantially for models without data augmentation. For example, BERT (w/o) model on ROC task achieves 88.49 % accuracy result (in Table 5) but only achieves 77.48 % (in Table 4) which drops by about 11%. On average, the accuracy drops by 16.21% for BERT (w/o), 12.04% for XLNet (w/o) and 19.54% for RoBERTa (w/o). It confirms that the original models are fragile with short-circuits and can be confused by questions that require a stronger connection between the premise and the choice.

For the stress tests in Table 4, *crossover* (+C), *mutation* (+M), and especially their combination (+C+M) improve the vanilla models substantially. For example, the performance improved by 27.06% for BT+C, 23.25% for BT+M and 30.31% for BT+C+M on RECLOR dataset. Besides, the performance gap between the stress test and the original test all narrows. +C, +M, and +C+M also consistently outperform back-translation (only gains 3.47% on stress test with XL+B). It shows that

Model	ROC	COPA	ARCT	RELOR
BT(w/o)	77.48	62.26	33.07	22.83
BT+B	82.35	77.17	44.75	24.94
BT+C	85.35	76.45	53.87	49.89
BT+M	87.60	81.87	**71.82**	46.08
BT+C+M	**91.31**	**86.38**	70.22	**53.14**
XL(w/o)	73.95	62.18	53.20	24.53
XL+B	75.30	64.85	54.00	33.37
XL+C	85.38	82.13	60.71	48.87
XL+M	88.02	76.21	69.73	54.55
XL+C+M	**92.35**	**90.77**	**73.07**	**56.47**
RB(w/o)	77.58	68.56	49.20	18.15
RB+B	76.17	77.51	53.38	22.03
RB+C	88.46	91.28	56.72	51.91
RB+M	88.55	85.65	73.33	**60.53**
RB+C+M	**94.39**	**93.15**	**74.13**	55.77

Table 4: Overall stress test on 4 models with or without(w/o) data augmentation. All numbers are percentages (%). +B = augmented with back-translation, +C = augmented with crossover, +M = augmented with mutation.

Model	ROC	COPA	ARCT	RELOR
BT(w/o)	88.49	64.60	61.94	45.60
BT+B	88.42	75.4	71.70	48.60
BT+C	87.60	75.73	70.80	47.00
BT+M	87.69	69.53	65.92	46.80
BT+C+M	87.47	73.2	68.54	43.60
XL(w/o)	90.88	63.40	77.85	56.00
XL+B	90.88	64.80	77.70	57.00
XL+C	90.52	74.60	78.60	54.40
XL+M	90.08	66.80	75.45	53.60
XL+C+M	90.40	72.93	76.95	54.20
RB(w/o)	92.16	72.00	77.10	50.40
RB+B	92.16	74.07	80.93	51.00
RB+C	91.68	77.07	79.05	50.40
RB+M	91.91	70.47	78.23	52.00
RB+C+M	92.46	75.67	77.78	48.40

Table 5: Overall original test on 4 models with or without(w/o) data augmentation. The best results for each dataset on each model are highlighted.

these *crossover* and *mutation* are effective for reducing short circuits in the models and improving the generalization of the models. Besides, they can complement each other. We can also observe that models with +M sometimes get the best performance in the stress test, like BERT+M on ARCT and RoBERTa+M on RECLOR. Because *mutation* can enhance grammatical knowledge for models, and voice stress test which accounted for a large proportion in all stress cases for ARCT and RECLOR can also test grammatical capability. We have statistical analysis for the 12 experiments (3 models on 4 datasets) in Table 4: according to t-tests, with $p < 0.05$, +C, +M and +C+M are significantly more accurate than (w/o) and +B in the stress test of all 12 experiments which indicates our improvements are stable.

For the original test in Table 5, +C+M is significantly better than (w/o) in 4 of 12 experiments. For example, we get an 8.6% improvement for BERT on COPA. In 4 of 12 experiments, there are no significant differences between +C+M and (w/o). In the remaining 4 experiments, the performance differences against (w/o) are within 2%. Overall, *crossover* and *mutation* don't hurt the model performance on the original test cases heavily and can even make improvements.

3.2.2 Fine-grained results

We proceed to break down the results in Table 4 into accuracies on stress tests created by different operators. COPA and RECLOR datasets do not show all six operators because some of the operators generate too little data for them, as shown in Table 3. The corresponding results are presented in Figure 4. We observe that the vanilla model in purple and back-translation in green show worse results across different aspects than other lines. The models trained with data augmented by *crossover* and *mutation* (the red lines) are generally more robust than others. Please refer to Appendix A. for complete results.

Since every type of stress tests can evaluate if a model is robust, particularly if it considers the premise by giving it two very similar choices, the above results on the stress tests of all types show that our two methods do reduce short-circuits, and may even encourage the models to look toward the premises. We will provide additional pieces of evidence to confirm this in the next two subsections.

3.3 Choice-only Test

The end-to-end test has shown the success of our data augmentation methods. To further explore the reason behind the performance gain, we also use choice-only test here.

In choice-only test, we only feed choices into a model without a premise which is replaced by an empty string. This way, models cannot utilize the relationship between premise and choices. Under normal conditions, we would expect the model to make arbitrary choices. However, if a model can easily "guess" the "right" choice which normally requires the relationship between premise and choices, one possibility is that this model cheats on evaluation procedure and may be fragile. Thus, the higher score may indicate more use of short-circuits.

In Figure 5, we observe that in choice-only tests, the accuracy of models augmented with *crossover* and *mutation* (red line) drops the most. Sometimes the performances are similar to random selection, e.g., RB+C+M on ARCT (56.38%), which indicates that models are no longer cheating. In other words, models augmented by crossover and mutation are more likely to consider the premises. The results on the choice-only tests provide another perspective for us to re-assure that models augmented with crossover and mutation can reduce short circuits and thus model fragility.

However, one may argue that even if a model can choose or "guess" correctly given only the choices but no premise, it may still have the ability to look at the premise if it's given one, like in the end-to-end test. Therefore, next we conduct an additional case study to show that short-circuit does take place and our augmentation methods alleviate it.

3.4 Case Study

Our case study is a series of white-box tests that demonstrate the change in attention patterns.

We take an example from ROC which is shown in Table 2. We explore BERT-based models by analyzing their attention maps on this question in Figure 6. In this example, the word "show" in the premise is strongly related to the token "reality show" in the right choice from human knowledge. The attention map is visualized via an off-the-shelf tool (Vig, 2019).

There is no positive attention value in front of the fourth sentence, so we intercept it from where it is worth. BERT trained on the original training set

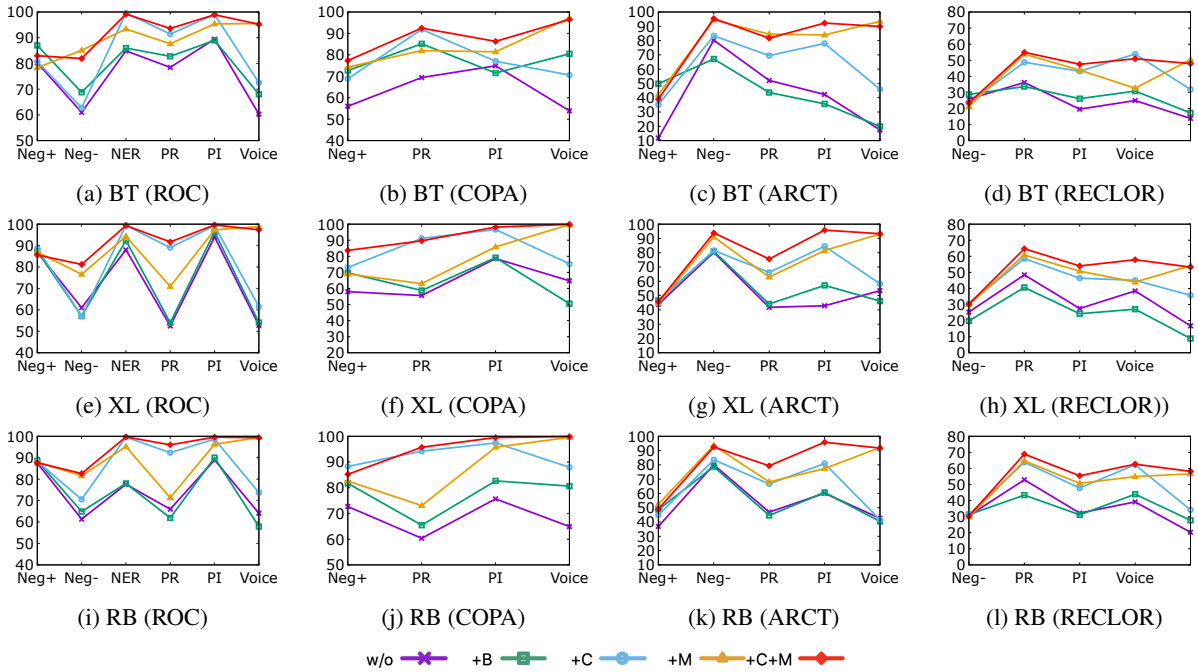


Figure 4: Fine-grained stress test with different aspects on 4 different tasks. The x-axis in the figures indicates different stress test aspects and the y-axis indicates model accuracy in percentage.

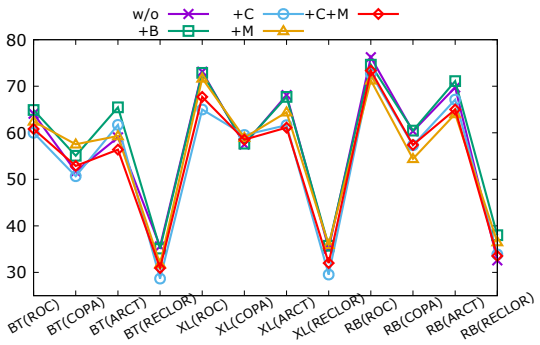
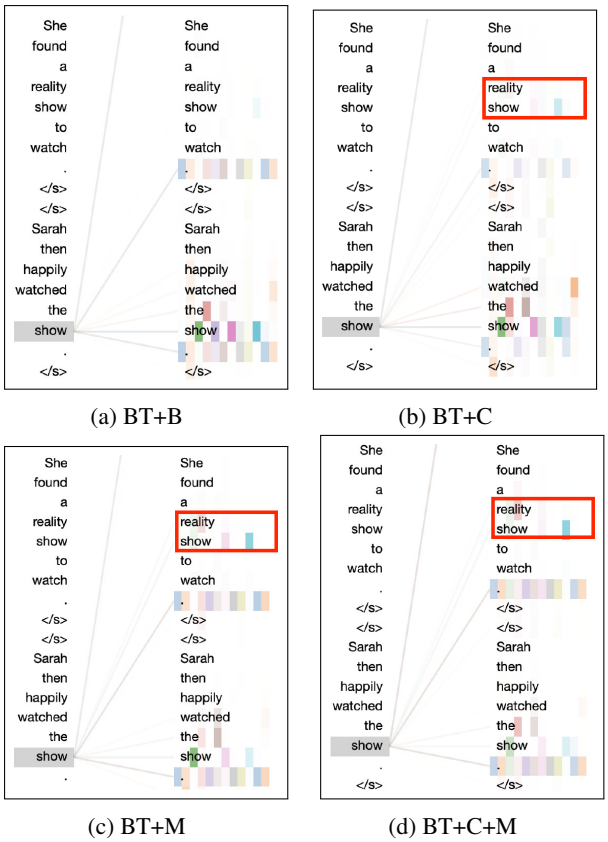


Figure 5: Choice-only test: Accuracies of different data augmentation methods with 3 models on 4 tasks. The detailed numbers are in Appendix B.

fails to pick up the right choice likely due to there being virtually no attention connection between words in the choice and words in the premise. After training with *crossover* data augmentation, the model learns to pay attention to the premise and the relationship between premise and choices. i.e., "show" in this example. Similar trends also exist for the *mutation* operation in Figure 6c and the combination of *crossover* and *mutation* operation in Figure 6d. The rationale behind such a change of attention pattern is that, in an MCQ created by *crossover* operation (Figure 6b), *mutation*(Figure 6c), and the combination of them (Figure 6d), the model needs to combine the information in the premise to effectively distinguish the true "right" choice from the wrong one. However, the light and sparse attention color blocks on the attention map for back-translation in Figure 6a indicate back-translation can not help BERT connect the choice and premise very well in this question. These observations empirically demonstrate the effectiveness of our methods in encouraging the model to pay attention to the premise to reduce short circuits. We provide additional cases in Appendix C.

4 Related Work

Data Augmentation. Data augmentation refers to strategies for increasing the diversity of training examples without explicitly collecting new data. It has received active attention in recent machine learning research such as UDA (Xie et al., 2020), which used back-translation (Sennrich et al., 2016), AutoAugment (Cubuk et al., 2018), RandAugment (Cubuk et al., 2020), and MIXUP (Zhang et al., 2017). These are often first explored in computer vision, and it seems secondary and comparatively underexplored for NLP. It is perhaps due to challenges presented by the discrete nature of language, which rules out continuous noise and makes it more difficult to maintain invariance. To augment more data in NLP tasks, previous work constructed more data with one kind of feature or rule have improved accuracy on that particular case, but didn't



(a) BT+B (b) BT+C

(c) BT+M (d) BT+C+M

Figure 6: Attention map on a ROC example for BERT-based models.

generalize to other cases, suggesting that models overfit to the augmentation set (Iyyer et al., 2018; Liu et al., 2019b). In particular, McCoy et al. found that augmentation with HANS examples may generalize to a different word overlap challenge set, but only for examples similar in length to HANS examples. We reduce the choice-only short circuit inference behavior of models via several simple yet feature-agnostic augmentation methods aiming at teaching models to reason over relations between context and choices.

Model Probing. Ever since the emergence of large pretrained language models, many works have focused on the analysis of their inner workings. As a result, a considerable amount of linguistic properties are shown to be encoded in the contextualized representations and attention heads (Goldberg, 2019; Clark et al., 2019; Liu et al., 2019a; Tenney et al.). In contrast, we are concerned with the model's higher-level reasoning capability. To prob what specific linguistic capabilities models get, one approach is to create challenging datasets. Some work (Belinkov and Glass, 2019) has noted benefits of this approach, such as systematic control over data, as well as drawbacks, such as small scale and lack of resemblance to "real" data. Further, they note that the majority of challenge sets are for Natural Language Inference. Our stress test which can also be called short-circuit test is not aimed to replace the challenge or benchmark datasets, but to complement them to test whether really have the inference capability, in particular the short circuiting behavior. The behavior is reflected in downstream performance through diagnostic stress tests.

Spurious Feature Analysis. Prior studies (Sharma et al., 2018; Zellers et al., 2018) have discovered that NLP models can achieve surprisingly good accuracy on natural language understanding tasks in MCQs form even without looking at the context. Such phenomenon is identified via the so-called "hypothesis-only" test. Sanchez et al. further showed that models sometimes bear insensitivity to certain slight but semantically significant perturbations in the hypothesis, leading to suspicions that the high hypothesis-only performance stems from statistical correlations between spurious cues in the hypothesis and the label. Such spurious cues can be categorized into lexicalized (Naik et al., 2018) and unlexicalized (Bowman et al., 2015): the former mainly contains n-gram and cross-ngram spans that are indicative of certain labels, while the latter involves word overlap, sentence length and BLUE score between the premise and the hypothesis. Instead of unearthing the specific cues in the dataset, we directly diagnose if models are exploiting the short circuit in hypothesis alone and mitigate such reasoning behavior accordingly.

5 Conclusion

We observe that models can select correctly without a premise and pay little attention to the premise on the attention map. Inspired by speculation that models can short circuit the premises on MCQs and become fragile, we propose two data augmentation methods *crossover* and *mutation*. Our experimental results show that, while the proposed methods do not always improve results on the original datasets, they significantly and consistently increase the accuracy on stress tests. They improve the model's robustness and generalization capability. We also confirm the reason for this improvement is the reduction of short-circuits with choice-only tests and case study. We conclude that our data augmentation methods can indeed encourage models to pay more attention to the premise of questions.

Limitations

There are some limitations in this work that could be addressed in future work.

First, though our work can reduce short circuits in multiple choice reasoning models, there are still more opportunities to apply *crossover* or *mutation* to other text generation models for enhancing their performance.

Second, the short circuit is only one kind of weakness for reasoning models. The approach outlined here only attacks short circuits. To tackle other types of model fragility, more clear interpretations of what models learn are required.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*. PMLR.

Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. 2021. HiddenCut: Simple data augmentation for natural language understanding with better generalizability. In *ACL*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *SemEval 2012*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task. *CoRR*, abs/1708.01425.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *NAACL*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *NAACL*.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019b. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *NAACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *ICLR*.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. 2020. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding.

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*.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *NAACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL*, pages 86–96.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *ACL*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *ICLR 2019*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *ACL*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation. *CoRR*, abs/1904.12848.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS 2019*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *ICLR*.

A Details of Stress Tests

Table 6 tells more detailed numbers [2] about stress test results with different aspects in Figure 4. (Section 3.2.2)

B Details of Choice-only test

In Table 7, we show specific numbers for Figure 5 which describe the choice-only results. (Section 3.3)

C Extra Cases

We have shown an example in Section 3.4 for the case study. In this section of the appendix, we provide extra 3 cases for further illustrating that *crossover* and *mutation* encourage models to build contextual reasoning by attending to relevant concepts in the premise.

Example 2 *An MCQ from COPA:*

Premise: *I pushed the door.*
Choice 1: *The door opened.* ✓
Choice 2: *The door locked.* ✗

In Example 2, we explore RoBERTa-based models by analyzing their attention maps on this question in Figure 7. In this example, the word "pushed" in the premise is strongly related with the word "opened" in the right choice from human knowledge. The relationship between these two words is the key to answering this question. We explore different models with the augmentation method with attention map to visualize if these two words have a relationship or not.

In Figure 7, RoBERTa trained on the original training set fails to pick up the relation between "pushed" and "opened". After training with *crossover* data augmentation, the model learns to build contextual reasoning by attending to relevant concepts in the premise. Similar trends also exist for the combination of *crossover* and *mutation* operation in Figure 7d. These observations empirically demonstrate the effectiveness of our methods to encouraging the model to pay attention to the premise so as to improve model robustness. On the contrary, back-translation in Figure 7b seems to have not enhanced such abilities.

Example 3 *An MCQ from COPA:*

[2]The dashes in Table 6 are caused by limited test cases which sizes are too small.

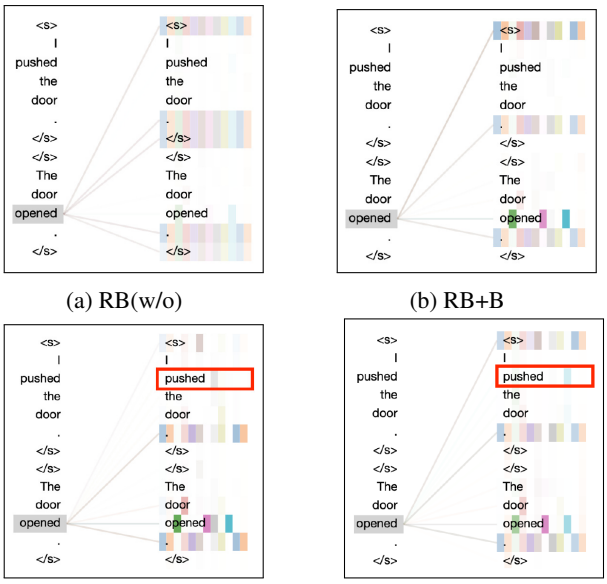


Figure 7: Attention map on a COPA example for models.

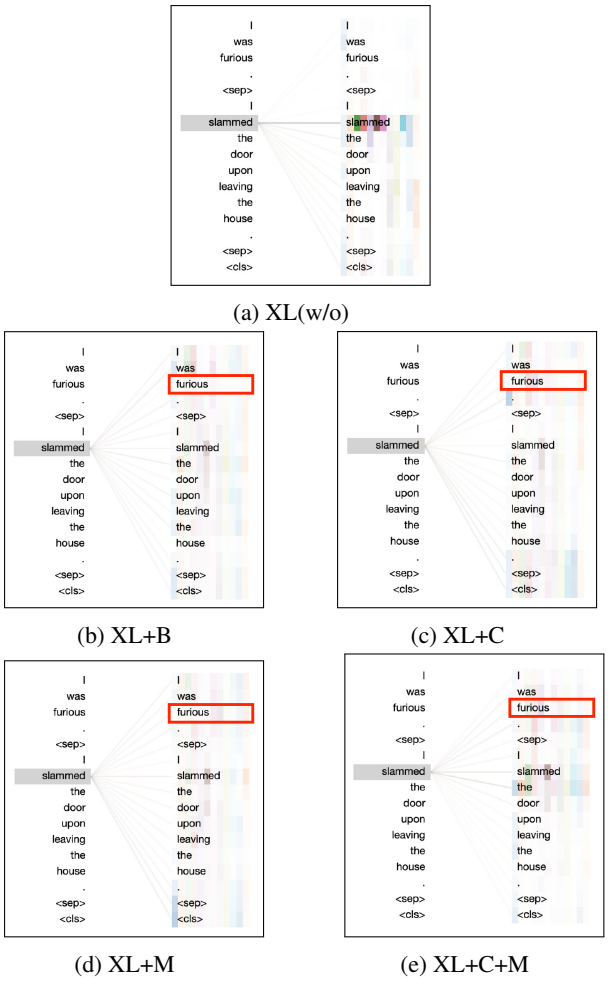
Premise: *I was furious.*
Choice 1: *I slammed the door upon leaving the house.* ✓
Choice 2: *I checked the mailbox upon leaving the house.* ✗

In human cognition, the word "furious" in premise and "slammed" in the right choice have a strong causal relationship in Example 3. However, from the attention map of the vanilla XLNet model in Figure 8, it is difficult to observe that they are related. In Figure 8, we also observe that the ability of XLNet to use relationships has been strengthened by adding augmented data with all methods we mentioned. Back-translation is worse than the other methods with lighter color blocks.

Example 4 *An MCQ from ARCT:*

Premise: *I would be happy to support free community college so those who can't afford it can get educated. College should be free.*
Choice 1: *I would be happy to pay tuition for everyone , even some rich kids.* ✓
Choice 2: *I would not be happy to pay for some rich kids tuition at the same time.* ✗

In Example 4, the claim and reason are "College should be free" and "I would be happy to support ... who can't afford it can get educated" separately. The word "free" is very important for the claim. It should be very related to the information in the correct warrant, such as "tuition" or "pay" from the



(a) XL(w/o)

(b) XL+B

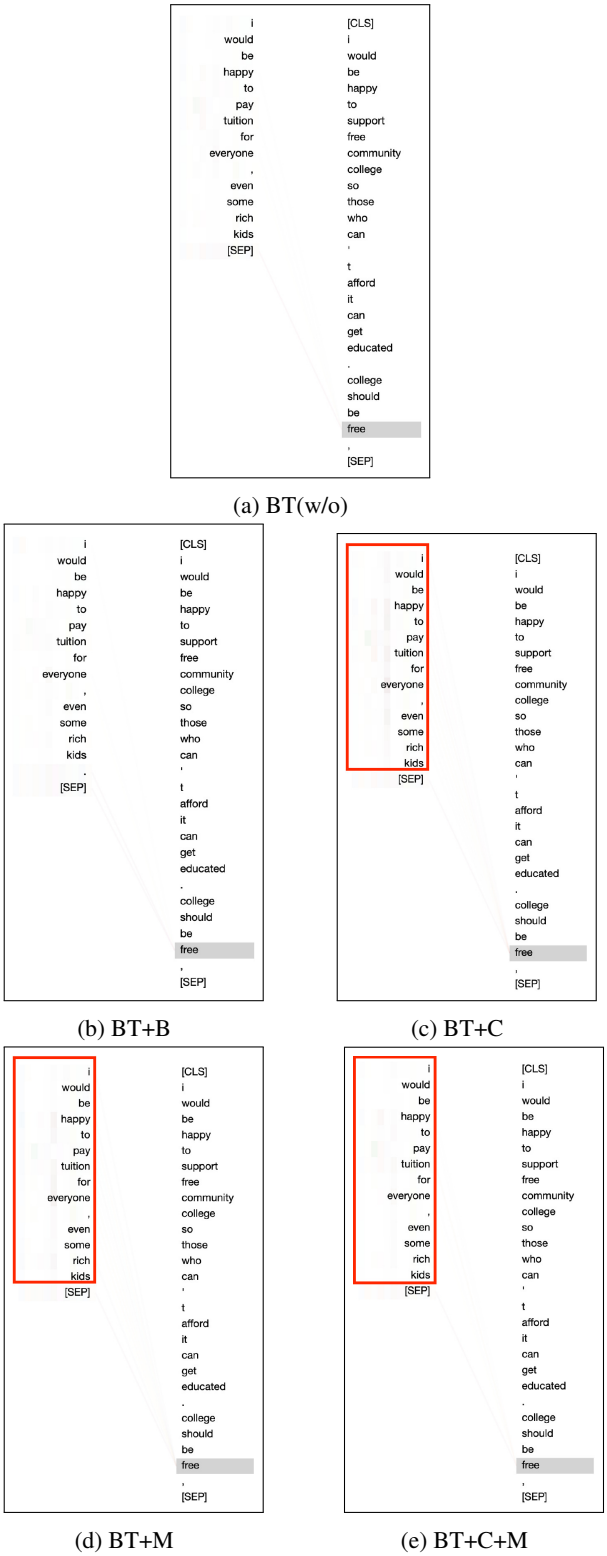
(c) XL+C

(d) XL+M

(e) XL+C+M

Figure 8: Attention map on a COPA example for XLNet-based models.

knowledge of commonsense reasoning. Unfortunately, "free" has little relationship with the warrant in Figure 9a through the vanilla BERT model. Consistent with our previous conclusion, the improvement effect of *crossover* and *mutation* is more obvious than back-translation. Besides, we also observe that the performance of data augmentation methods is not as obvious as the first two examples. One reason may be that analyzing with this white-box method is not completely reliable. The other may be that the ability of these data augmentation methods to reduce short circuits and to improve the stability of the model is limited. We will continue to study the reason in the future.



(a) BT(w/o)

(b) BT+B

(c) BT+C

(d) BT+M

(e) BT+C+M

Figure 9: Attention map on an ARCT example for BERT-based models.

Dataset	Model	Original	Neg+	Neg-	NER	PR	PI	Voice	All
	BT(w/o)	88.49	80.24	60.99	84.90	78.47	89.47	60.26	77.48
	BT+B	88.42	86.99	68.79	86.00	82.76	88.97	68.02	82.35
	BT+C	87.60	80.73	62.76	99.63	91.42	99.03	72.52	85.35
	BT+M	87.69	78.36	85.10	93.37	87.64	95.39	95.50	87.60
	BT+C+M	87.47	82.99	81.91	99.17	93.60	98.92	95.23	91.31
	XL(w/o)	90.88	86.94	60.99	88.03	52.50	94.00	52.76	73.95
	XL+B	90.88	87.39	57.09	92.36	53.99	96.44	54.08	75.30
ROC	XL+C	90.52	88.65	57.09	99.45	89.03	99.30	61.47	85.38
	XL+M	90.08	86.98	76.60	94.29	70.92	97.29	98.88	88.02
	XL+C+M	90.40	85.61	81.21	99.35	91.71	99.62	97.37	92.35
	RB(w/o)	92.16	87.50	61.35	77.62	65.99	88.97	64.10	77.58
	RB+B	92.16	88.56	64.89	77.99	61.91	90.05	57.89	76.17
	RB+C	91.68	88.24	70.57	99.63	92.36	98.68	73.70	88.46
	RB+M	91.91	87.96	81.56	95.21	71.36	96.28	99.48	88.55
	RB+C+M	92.46	87.67	82.62	99.72	96.02	99.61	99.34	94.39
	BT(w/o)	64.60	56.03	-	-	69.41	74.89	53.93	62.26
	BT+B	75.40	72.70	-	-	85.16	71.54	80.49	77.17
	BT+C	75.73	68.83	-	-	91.87	77.02	70.60	76.45
	BT+M	69.53	74.26	-	-	82.01	81.43	97.29	81.87
	BT+C+M	73.20	77.31	-	-	92.48	86.30	96.47	86.38
	XL(w/o)	63.40	57.99	-	-	55.59	78.54	64.77	62.18
	XL+B	64.80	69.71	-	-	58.74	79.15	50.54	64.85
COPA	XL+C	74.60	73.04	-	-	91.16	96.65	75.34	82.13
	XL+M	66.80	68.97	-	-	63.01	85.84	99.73	76.21
	XL+C+M	72.93	83.67	-	-	89.64	98.18	99.86	90.77
	RB(w/o)	72.00	72.69	-	-	60.36	75.65	64.91	68.56
	RB+B	74.07	81.71	-	-	65.44	82.65	80.63	77.51
	RB+C	77.07	88.28	-	-	94.21	97.41	87.94	91.28
	RB+M	70.47	82.59	-	-	73.07	95.74	99.59	85.65
	RB+C+M	75.67	85.23	-	-	95.73	99.54	99.86	93.15
	BT(w/o)	61.94	11.78	80.26	-	52.11	42.26	17.43	33.07
	BT+B	71.70	49.83	67.10	-	43.66	35.72	19.92	44.75
	BT+C	70.80	35.13	83.33	-	69.48	77.98	45.98	53.87
	BT+M	65.92	42.20	94.52	-	84.51	83.93	93.48	71.82
	BT+C+M	68.54	38.94	95.39	-	81.69	92.26	89.85	70.22
	XL(w/o)	77.85	43.88	80.26	-	41.78	42.86	53.45	53.20
	XL+B	77.70	46.57	80.92	-	44.13	57.14	46.17	54.00
ARCT	XL+C	78.60	45.68	81.58	-	66.20	84.53	58.24	60.71
	XL+M	75.45	44.55	91.01	-	62.91	81.55	93.10	69.73
	XL+C+M	76.95	45.68	93.86	-	75.59	95.83	93.29	73.07
	RB(w/o)	77.10	36.92	80.04	-	46.95	60.12	40.61	49.20
	RB+B	80.93	48.71	78.73	-	44.60	60.71	40.42	53.38
	RB+C	79.05	44.89	83.55	-	66.67	80.95	41.57	56.71
	RB+M	78.23	52.41	93.64	-	68.07	77.38	92.14	73.33
	RB+C+M	77.78	49.05	92.54	-	79.34	95.83	91.76	74.13
	BT(w/o)	45.60	25.87	36.13	-	19.56	24.91	13.81	23.08
	BT+B	48.60	28.71	33.61	-	26.09	30.77	17.24	26.06
	BT+C	47.00	23.64	48.74	-	43.12	53.85	31.81	33.94
	BT+M	46.80	21.24	53.78	-	43.84	32.60	50.32	36.81
	BT+C+M	43.60	23.47	54.90	-	47.46	50.92	47.91	39.29
	XL(w/o)	56.00	30.58	52.94	-	32.24	39.19	20.28	31.52
	XL+B	57.00	31.29	43.42	-	31.16	43.95	27.76	33.05
RECLOR	XL+C	54.40	31.47	63.87	-	47.83	62.27	34.22	40.92
	XL+M	53.60	29.78	64.71	-	50.72	54.94	56.65	46.20
	XL+C+M	54.20	30.31	68.91	-	55.43	62.64	58.18	48.58
	RB(w/o)	50.40	25.33	48.46	-	27.53	38.46	16.73	27.34
	RB+B	51.00	19.82	40.62	-	24.27	27.10	8.88	20.53
	RB+C	50.40	30.66	58.54	-	46.38	45.06	35.74	38.54
	RB+M	52.00	29.96	60.78	-	50.73	43.96	54.12	44.01
	RB+C+M	48.40	30.22	64.71	-	53.99	57.88	53.23	46.03

Table 6: Detailed Breakdown of Stress Tests on 4 models with or without(w/o) data augmentation. +B = augmented with backtranslation, +C = augmented with crossover, +M = augmented with mutation. Stress Tests includes the following stress tests: Neg+=negation-add, Neg-=negation-remove, NER, PR=pronoun-replacement, PI=Pronoun-instantiation, Adv=adverbial, Voice, Syn=synonym.

Model	ROC	COPA	ARCT	RECLOR
BT(w/o)	64.10	51.67	59.01	35.60
BT+B	64.90	55.07	65.47	35.13
BT+C	59.99	50.67	61.71	28.67
BT+M	62.44	57.53	59.31	31.80
BT+C+M	60.82	52.87	56.38	30.93
XL(w/o)	73.12	57.47	68.09	35.13
XL+B	72.88	57.67	67.72	35.73
XL+C	65.01	59.53	61.64	29.53
XL+M	71.69	58.93	64.41	35.93
XL+C+M	67.72	58.53	61.11	32.00
RB(w/o)	76.23	60.33	69.75	32.60
RB+B	74.63	60.47	71.10	38.00
RB+C	72.73	57.33	67.12	33.87
RB+M	71.28	54.40	64.04	36.53
RB+C+M	73.35	57.40	65.01	33.53

Table 7: Choice-only test for transformer-based models on 4 datasets. All numbers are percentages (%)