ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230383

Combating Short Circuit Behavior in Natural Language Reasoning: Crossover and Mutation Operations for Enhanced Robustness

Shanshan Huang[a], Siyu Ren[b] and Kenny Q. Zhu[c;*]

[a]Shanghai Jiao Tong University, China
[b]Shanghai Jiao Tong University, China
[c]University of Texas at Arlington, United States

Abstract. In this study, we delve into the "short circuit" phenomenon observed in multiple-choice natural language reasoning tasks, where models tend to make accurate choices without properly considering the context of the question. To better understand this phenomenon, we propose white-box and black-box proxy tests as investigative tools to detect short circuit behavior, confirming its presence in fine-tuned NLU reasoning models. To tackle the short circuit issue, we introduce biologically inspired "crossover" and "mutation" operations. These operations are applied to augment the training data for popular models such as BERT, XLNet, and RoBERTa. Our results demonstrate that these data augmentation techniques effectively enhance the models' robustness and mitigate the short circuit problem.

1 Introduction

Multiple-choice questions (MCQs) are a widely used format for assessing Natural Language Understanding (NLU) tasks, such as causal reasoning [11], story ending prediction [22, 13], argument reasoning comprehension [12], and reading comprehension [33]. These tasks typically consist of a premise followed by two or more choices. For example, the COPA dataset [11] tests commonsense causal reasoning [20] through MCQs, as shown below.

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.* ✗

Recent research has sought to explain the strong performance of advanced neural models on NLU reasoning problems. In particular, there is speculation that many models succeed not by genuinely understanding the semantic and logical connections between the context and the choices, but by exploiting spurious statistical features in the training and test data. This idea is supported by"choice-only tests" (also known as"ending-only tests") [27, 4], where models like BERT can correctly answer questions even when the context is removed.

In this paper, we refer to this phenomenon as"short circuit" in Natural Language reasoning. Although choice-only tests provide some evidence for short circuit behavior, we argue that they have inherent limitations. Just because a model can answer correctly without the premise doesn't necessarily mean it doesn't consider the premise when it is provided. To address this issue, we need a test that works with complete questions, including both premises and choices.

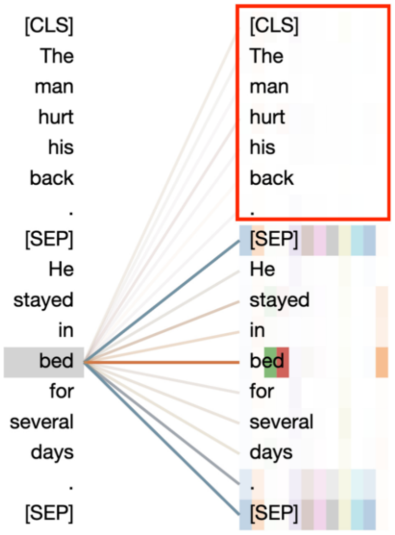


Figure 1: Attention map showing that BERT [8] short-circuits on a COPA question.

Our initial attempt at a more comprehensive short circuit test involves plotting the attention map between the words in a full question from the final encoder layer of the model. We illustrate this approach using an attention map of the example from COPA (Example 1) in Figure 1. The diagram clearly shows that there is virtually no connection between the first choice and the context when the model processes the full question, while the attention between words within the first choice remains the same when the model processes only the choices without the context.

Manually examining the short circuit behavior of a model using attention maps is tedious and costly. To address this issue, we implement a white-box testing algorithm that simulates human visualization using threshold values. However, this approach has limitations: it requires access to the model's code and only works for attention-based models.

To overcome these challenges, we introduce a new operation

[*] Corresponding Author. Email: kenny.zhu@uta.edu.

called“crossover” for MCQ question instances. Crossover exchanges the choices of two MCQs, analogous to how chromosomes swap segments during biological reproduction. This operation poses a unique challenge for models that frequently exploit short circuit behavior and can detect such behavior in real tasks by constructing proxy test cases. By examining crossover tests and other instance-level stress tests, such as named entity replacement, we find evidence of short circuit behavior in three recent, powerful NLU reasoning models, as indicated by notable declines in accuracy on these tests.

Having identified the presence of short circuit behavior, our next goal is to improve model robustness. While generating more training examples using stress tests the model struggles with might be a direct approach, many stress tests impose constraints on choice construction, limiting their effectiveness as general data augmentation methods. However, the crossover operation and its counterpart, “mutation,” offer a suitable solution. These operations not only allow for the detection of short circuit behavior but also serve as effective data augmentation techniques to reduce its occurrence, enhancing the overall robustness of NLU models.

To this end, we apply crossover, mutation, and back-translation [31] to augment BERT, XLNet [32], and RoBERTa [19] on ROC [22], COPA, ARCT [12], and RECLOR [33]. Our experiments show up to a 24% increase in accuracy on stress tests and a 10% increase on the original test data.

This paper makes three main contributions:

1. We propose two approaches for detecting short circuit behavior: a white-box method based on attention weight thresholding and a black-box“crossover” test inspired by molecular biology.
2. We experimentally verify the existence of short circuit behavior in three powerful, fine-tuned NLU reasoning models.
3. We suggest using crossover and mutation operations to augment training data, encouraging models to consider the context of questions. Our experiments confirm the effectiveness of this approach, demonstrating substantial improvements in model robustness, not only on stress tests but also on the original test data.

2 Approach

In this section, we first present our methods for testing short circuits in models, and then modify some of these methods to create training data to address the short circuit problem and enhance model robustness.

2.1 Proxy Test for Short Circuit

Since no existing method can definitively prove if a model is short-circuiting on a question, we propose two types of approaches that serve as proxy tests for short circuits. These approaches reveal the effects of model short-circuiting, though they can’t directly prove the short-circuit itself, similar to dark matter. One approach involves inspecting attention maps in models under a white-box setting, while the other generates new test cases by applying different operations on correct choices under a black-box setting.

White-box Attention Weights (AW)

One intuitive way to detect if an attention-based model is exploiting short circuits is to visualize its attention map. Given a well-trained model and a correctly answered MCQ in the form of *[CLS] premise [SEP] choice [SEP]*, where *[CLS]* and *[SEP]* are model-dependent delimiters and *choice* refers to the correct choice, we first tokenize the input, feed the token sequence into the model, and extract the attention map of all attention heads from the last encoder layer.

The attention maps are visualized through an off-the-shelf tool [29] into a user-friendly demo, as shown in Figure 1. Human annotators are then asked to determine whether there exists strong attention connections from the correct choice to the premise. We consider the MCQ to be solved without short-circuiting only if over half of the annotators label it as having strong attention connections.

Although accurate, such manual annotation is cost-prohibitive to be scaled to larger tests. To remedy this issue, we propose a rule-based procedure to automatically detect the short circuit behavior of a model on MCQ. Specifically, we aggregate the attention maps into one individual map by max-pooling over all attention heads. Then we check if there exists at least one attention score between a token in the choice and a token in the premise higher than threshold t_1, or at least two higher than threshold t_2, excluding special tokens like comma and period. We consider the model to not be short-circuiting on this MCQ if neither of the two conditions is met. In practice, the thresholds t_1 and t_2 are tuned to maximally simulate human annotation. The pseudo-code is shown in Algorithm 1.

Algorithm 1 Attention Weight Thresholding

```
Input: premise $P$, correct choice $C$, model $M$, threshold $t_1$ and $t_2$.
Output: binary 0/1 label $L$.
initialize counters $c_1$ and $c_2$ to 0.
tokenize the formatted input as sequence of tokens $S$.
feed $S$ into $M$ and extract the last layer's attention maps $Attn_{all}$.
aggregate $Attn_{all}$ into $Attn_{max}$ by max-pooling over all attention heads.
for $w_1$ in $C$ do
    for $w_2$ in $P$ do
        if $Attn_{max}(w_1, w_2) > t_1$ then $c_1$ += 1
        if $Attn_{max}(w_1, w_2) > t_2$ then $c_2$ += 1
output 1 if $c_1 > 0$ or $c_2 \geq 2$ and 0 otherwise.
```

Black-box Choice Operator

While attention-based testing methods can detect short circuits within the encoder directly, they don’t directly detect short circuits in the end-to-end MCQ model, which also includes a linear layer above the attention-based pretrained language model. Additionally, these methods are limited to a family of models with inherent attention mechanisms.

A more desirable approach is an automatic end-to-end black-box test that is model-independent. In black-box testing, if a model correctly answers an MCQ, we slightly modify the MCQ by applying a certain“operation” on the original correct choice to produce another wrong choice. The newly generated MCQ must share the same correct choice as the original question. By observing the model’s response to the second MCQ, we can infer whether the model short-circuits on the original MCQ.If the model still selects the correct choice, then we consider it to have passed the test and not short-circuited on the original MCQ. The challenge now is how to construct the new wrong choice by implementing the operation in various ways.

In this paper, we consider the operations listed in Table 1. Some of the operations were mentioned in previous literature, while others are proposed here (marked with *). The first line in each cell describes the operation, and the next two lines provide an example of constructing a false choice from a choice in the original question. An operation may either preserve (p) the truth value (✗ → ✗) or change (c) the truth value of the choice (✓ → ✗).

Oper.	Description and Example
Neg+	Add negation (c) *They called the police to come to my house.* ✓ *They didn't called the police to come to my house.* ✗
Neg-	Remove negation (c) *Ben never starts working out.* ✓ *Ben starts working out.* ✗
NER	Randomly replace person names (c) *A big wave knocked Mary down .* ✓ *A big wave knocked Kia down .* ✗
PR*	Switch pronoun by gender or quantity (c) *She had a great time .*✓ *He had a great time .* ✗
PI*	Instantiate pronoun by randome person (c) *They gave Tom a new latte with less ice .* ✓ *Nathanael gave Tom a new latte with less ice .* ✗
Adv	Add adverbs for emphasis (c) *The ocean was a calm as a bathtub .*✗ *In fact the ocean was a calm as a bathtub .*✗
CO*	Crossover: Swap the true choices between two questions (p) *Josh got sick .* ✓ *She had a great time .*✗
Syn	Replace adj/adv with synonym (p) *Dawn felt happy about getting away with it .* ✗ *Dawn felt glad about getting away with it .* ✗
MT*	Mutate: Swap two consecutive words (c) *Deb said yes to Tim 's marriage proposal.* ✗ *Deb said yes Tim to 's marriage proposal .*✗
Voice	Swap subject and object (c) *Kara asked the neighbors not to litter in their yard .* ✓ *the neighbors asked Kara not to litter in their yard .* ✗

Table 1: A number of operations considered for proxy testing. First line in each cell describes the operation, the next two lines give an example of how to construct a false choice from a choice of the original question. An operation may either preserve (p) the truth value (✓ → ✓ , ✗ → ✗) or change (c) the truth value of the choice (✓ → ✗).

Inspired by boundary testing in software engineering, we can classify these operations into three equivalent classes (three vertical sections in Table 1), depending on the nature of the *false* choice constructed:

1. The syntax and semantics are correct, and the *false* choice appears similar to the *true* choice.
2. The syntax and semantics are correct, and the *false* choice appears distinct from the *true* choice.
3. Either syntax or semantics is incorrect.

The last class is not suitable for testing short circuits because the model may answer the proxy question correctly by eliminating the false choice due to errors in it, not by considering the premise.

We focus on perturbations on negation [24], NER [24], and pronouns in the first class and adverbial [1], crossover, and synonym [24, 1] in the second class.

While most of the operations are self-explanatory, the *crossover* operation is unique and deserves special attention. Inspired by molecular biology, for each MCQ in the dataset that the model answers correctly, we substitute the original false choice with the true choice from another randomly sampled MCQ. The substituted choice remains false in the proxy question. The operation can be visually explained in Figure 2.

Compared to all other operations in classes 1 and 2, the crossover provides a proxy question that is most different from the original one but easier from a human perspective. This is because the two choices may be quite unrelated. If the model does not handle it correctly, it may be more indicative of a short circuit. As a result, the crossover is potentially a better short circuit test than others.

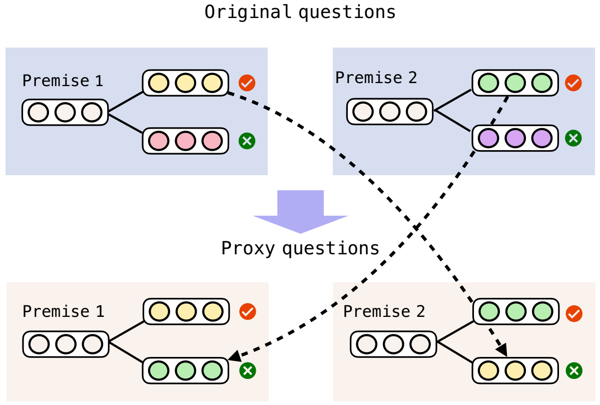


Figure 2: The Crossover Operation: the true choice of both questions are used to replace the false choices of these questions to create two new proxy questions.

Another advantage of the crossover operation is that we can generate multiple false choices for an original question at a low cost, allowing us to test each original question more thoroughly. In contrast, most other operations cannot produce an adequate number of different variants of the original choice.

In summary, the proposed black-box choice operator provides a more generalizable and model-independent method for detecting short circuits in MCQ models. By applying various operations to create proxy questions, we can assess the model's performance and robustness more accurately, contributing to the development of better and more reliable models in the future.

2.2 *Improving Model Robustness by Data Augmentation*

If a model is shown to short-circuit by the proxy tests, its performance may decline, especially when applied to out-of-domain test data. To make models more robust, one natural thought is to generate more data to encourage models to focus on the relation between the premise and choices. While the operations used to generate proxy tests can also be utilized for data augmentation, not all of them are scalable or able to generate enough data for training.

The two operations that can generate a substantial amount of data are crossover and mutation. These operations can be applied to the training data to enhance the model's robustness.

Crossover for Data Augmentation

Crossover is a good option for data augmentation because the two choices were originally true answers in their respective questions and presumably carry spurious features if the model was short-circuiting. By incorporating crossover into the training data, the model is forced to consider the premise in order to determine which choice is better.

Mutation for Data Augmentation

Mutation has two flavors: (1) swap the words only in the true choice; (2) swap the words both in the true and the false choice. Compared to crossover, mutation has the potential to be more effective at improving model robustness. It not only forces the model to look into the premise due to its two very similar choices (same set of tokens), but also makes the model more sensitive to fine differences in word orders and enhances the model's prior grammatical knowledge.

Differentiating between Proxy Test and Data Augmentation

It is essential to differentiate between the use of crossover and mutation operations in proxy tests and data augmentation. In proxy tests, these operations are used to modify the test data to assess the model's short-circuiting behavior. In contrast, when applied for data augmentation, the same operations work on the training data to enhance the model's robustness and generalization capabilities.

In conclusion, data augmentation through crossover and mutation operations can contribute to improving model robustness by encouraging models to focus on the relationship between the premise and choices. By incorporating these operations into the training data, models are forced to consider the premise and become more sensitive to the fine differences in word orders, leading to better performance and reliability in real-world applications.

3 Experiments

First, we show the experimental setup. Second, we compare several test operators for the discovery of short circuit problems. Third, we evaluate robustness and the ability to avoid short circuit for models with different augmentation methods.

3.1 Experimental Setup

In this section, we will present our experimental setup, including the datasets, models, and test operators used in our study.

3.1.1 Datasets

We experiment on four 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. An example is shown in Table 2.

COPA is a causal reasoning dataset, an example of which was previously shown in Section 1. Given a premise context, COPA requires choosing the more plausible, causally related choice. There are 500 instances in the training data and 500 instances for testing.

ARCT is an argument reasoning comprehension dataset. It contains questions where the reason is connected to the claim, and there may exist an alternative warrant choice.

RECLOR is a reading comprehension dataset that requires logical reasoning to answer questions based on provided text passages.

3.1.2 Models

We mainly investigate three popular classifiers based on pre-trained language models. There are several available versions of pre-trained models differing in the number of layers and parameters. We choose to use the base version of each model. We train and test all the models on a server with a GeForce GTX 1080 Ti GPU with 11G RAM and an Intel(R) Xeon(R) CPU E5-2630 with 128G of RAM.

BERT (BT) is a popular attention model that applies the bidirectional training of the Transformer architecture. The base version has 12 Transformer layers, a hidden size of 768, and 12 self-attention heads, totaling 110M parameters. It is fine-tuned for three epochs to predict the relation based on context and choices.

XLNet (XL) is an autoregressive pre-trained language model that combines the strengths of BERT with the permutation-based training approach. It introduces a new technique called Permutation Language Modeling, which enables the model to learn bidirectional context by maximizing the expected likelihood over all possible permutations of the input sequence. XLNet does not use the Next Sentence Prediction (NSP) objective as BERT does.

RoBERTa (RB) is an improved pre-training procedure of BERT that involves training the model on more data, using larger batch sizes, and removing the NSP objective. These changes result in a more robust and better-performing model compared to the original BERT architecture.

3.1.3 Stress Test Cases

Following previous research [24], we will test the effectiveness of different data augmentation methods by looking at the robustness of models against different stress tests. We create these stress test cases using the proxy operations introduced in Table 1. Different operations generate different number of cases, as shown in Table 3. To evaluate the ability to test for short-circuiting, we will use a subset of these test cases in the next section.

3.2 Testing for Short Circuit

In this section, we will select proper testing operators for short circuit testing, and use these operators to detect the extent of model short circuiting.

3.2.1 Selecting Short Circuit Testing Methods

In Section 2.1, we discussed the possibility that both white-box attention-based method (AW [1]) and black-box choice operators in some of the equivalent classes can evaluate short circuits. We now investigate which proxy tests are better suited for short circuit evaluation.

As described in Section 2.1, each test operator generates new test cases by making directional changes to the test cases that the model chooses the right answer. The model is considered not short-circuiting on a case according to a test operator if it still gets the right answer after the operation. Assuming that human attention annotation, attention weight thresholding, and each choice operator are all plausible proxy tests, we can obtain 9 different proxy tests.

Then, we randomly sample 30 MCQs from the test set of ROC that are correctly answered by three models, respectively. Each proxy test will produce a 30-dimensional one-hot vector (proxy vector) for each model, where 1/0 indicates if the model short-circuited on that specific MCQ or not[2]. For each model, we then compute another vector as the ensemble of all proxy tests by majority voting on each of the 30 dimensions.

The smaller Euclidean distance between the individual proxy vector of each test type and the ensemble vector indicates higher reliability. The full results are shown in Table 4. We can find that the results of CO and AW are generally closer to the ensembled results, as reflected by the smaller distances. Thus, we consider that CO and AW are more suitable as proxy tests for short circuit evaluation.

[1] Here, t_1 and t_2 are tuned to 0.14 and 0.13 respectively, using 100 human-labeled cases. These cases are randomly sampled from the training data across the four datasets.

[2] For MCQs where a certain proxy test is not applicable, we randomly label it as 1 or 0.

Dataset	**Premise**	**Choices**	**Training size**	**Test size**
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 three other datasets.

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
Adv	1,850	496	444	500
CO	1,871	500	444	500
Syn	653	25	303	289
MT	1,871	500	444	500
Voice	1,014	246	174	263
Total	11,446	2,808	2,390	2,709

Table 3: Number of stress test cases by different operations for 4 datasets.

Test types	BERT	XLNet	RoBERTa	Ave
Neg+	3.16	3.87	**2.45**	3.16
Neg-	3.74	3.74	4.12	3.87
NER	3.87	3.87	4.12	3.95
PR	4.0	3.61	3.87	3.83
PI	3.74	3.74	3.74	3.74
CO	**2.83**	**2.63**	2.83	**2.76**
AW	**2.45**	3.46	**2.45**	**2.79**
Choice-only	4.0	3.74	3.87	3.87
Human	3.0	**2.55**	3.0	2.85

Table 4: Euclidean distances between proxy vector and the ensemble vector on short circuit test (the smaller the better). Ave is the average score across all models. Top two tests for each model are highlighted.

3.2.2 *Testing Short Circuit Problems*

We test short circuits by observing AW and CO scores, i.e., higher AW/CO scores indicate a lower chance for short-circuiting. We fine-tune the multiple-choice classifiers of BERT, XLNet, and RoBERTa on four datasets. In Table 5, we observe that the original models (in gray color) without data augmentation are most susceptible to short-circuits, as the AW and CO scores are relatively low. For the XLNet model on ROC, the AW score is even lower than 30%, which suggests a high likelihood of short-circuiting on ROC.

3.3 *Improving Overall Model Robustness*

To address the issue of model robustness, we tested the models and proposed data augmentation methods to improve their performance. Our analysis reveals that BERT, XLNet, and RoBERTa models on various datasets are generally not robust under stress tests. To remedy this, we employed data augmentation techniques, such as crossover, mutation, and a combination of both (+C+M), and compared their effectiveness to a back-translation baseline.

3.3.1 *Model Weakness*

As shown in Table 5, BERT, XLNet, and RoBERTa models exhibit a significant performance drop when subjected to stress tests. For instance, the accuracy rate of the XLNet model trained with ROC declines by 11.59%, and the AW short circuit score is 28.8%, suggesting that the model may be susceptible to short circuit issues. Similarly, all three models perform worse on the RECLOR and ARCT datasets, with a performance drop of about 10%, which aligns with the lower CO scores. These results indicate that model instability is a widespread problem, and short circuit is a probable cause.

3.3.2 *Data Augmentation*

To mitigate the weaknesses identified, we trained models using two primary data augmentation methods: crossover and mutation, which were discussed in the previous section. We also combined these two methods (+C+M) by constructing training data that incorporates both techniques. We used back-translation [31] as the baseline for data augmentation, as it has demonstrated universality and effectiveness in previous work. The expanded data volume is consistent with the original data volume.

Table 5 presents the results for the"original test." We observe that the four data augmentation methods do not negatively impact the model's performance on the original dataset and may even help the model achieve better accuracy. For instance, in the ROC dataset, the accuracy of BERT and RoBERTa models trained with crossover augmented data surpasses the base model, ranking first. The crossover method also proves effective on COPA. Although back-translation mostly achieves higher scores on ARCT and RECLOR, +C, +M, and +C+M methods only slightly underperform compared to the base model.

Considering the"Stress" column in Table 5, we find that different methods exhibit varying levels of robustness. Overall, the +C+M method demonstrates the best performance on the stress test, except when training RoBERTa on the RECLOR dataset. This outcome indicates that this type of data can protect models from confusion caused by simple perturbations and enhance model robustness. However, back-translation does not significantly improve model robustness. While the crossover method alone can contribute to robustness under stress tests, it is not as effective as +M and +C+M methods.

Further analysis of the models using the short circuit test reveals that the crossover method consistently achieves the highest CO score and often ranks best in the AW score. This finding suggests that models trained with crossover data augmentation learn to consider the premise to avoid short circuit issues.

	Short circuit Tests		Robustness Tests	
Model	**AW**	**CO**	**Original**	**Stress**
BT(w/o)	98.76	90.80	86.58	81.93
BT+B	99.26	92.54	86.75	82.96
BT+C	**99.69**	**98.47**	**87.07**	84.34
BT+M	99.26	91.47	86.48	86.06
BT+C+M	98.82	97.78	86.75	**88.60**
XL(w/o)	28.08	83.28	**90.81**	79.22
XL+B	19.27	84.4	90.43	82.23
XL+C	**64.58**	**98.81**	89.47	86.23
XL+M	62.77	86.90	90.17	89.47
XL+C+M	60.25	97.10	90.22	**92.64**
RB(w/o)	77.41	88.76	**92.73**	82.33
RB+B	58.15	87.98	92.46	78.50
RB+C	82.71	**99.3**	91.18	88.92
RB+M	71.73	88.06	92.62	90.29
RB+C+M	**93.31**	97.44	91.88	**93.06**

(a) ROC

	Short circuit Tests		Robustness Tests	
Model	**AW**	**CO**	**Original**	**Stress**
BT(w/o)	89.68	68.71	62.00	57.40
BT+B	96.79	85.42	68.60	68.95
BT+C	**98.35**	**97.25**	**72.80**	78.84
BT+M	95.17	90.62	70.40	79.62
BT+C+M	96.69	96.13	72.40	**80.68**
XL(w/o)	93.16	60.26	61.40	57.71
XL+B	91.46	65.51	63.20	61.06
XL+C	45.13	**94.69**	**67.80**	75.42
XL+M	76.85	57.23	62.20	71.10
XL+C+M	**98.51**	83.93	67.20	**81.32**
RB(w/o)	80.89	78.01	76.40	74.85
RB+B	**96.36**	83.64	77.00	80.26
RB+C	89.62	**98.23**	**79.00**	83.31
RB+M	62.26	84.30	72.60	83.53
RB+C+M	61.89	92.70	74.00	**87.30**

(b) COPA

	Short circuit Tests		Robustness Tests	
Model	**AW**	**CO**	**Original**	**Stress**
BT(w/o)	**99.65**	78.52	63.96	58.08
BT+B	99.34	61.18	68.47	56.21
BT+C	98.37	**96.08**	**68.92**	65.73
BT+M	98.67	74.42	67.79	69.65
BT+C+M	98.00	90.0	67.57	**73.71**
XL(w/o)	85.67	59.10	75.45	61.72
XL+B	95.73	60.40	**79.05**	64.78
XL+C	55.59	**92.45**	74.55	69.93
XL+M	**95.74**	59.57	74.10	73.15
XL+C+M	86.26	90.35	77.03	**79.11**
RB(w/o)	99.14	60.29	78.83	66.16
RB+B	97.78	60.94	**81.31**	66.02
RB+C	79.19	92.77	77.93	70.64
RB+M	**100.00**	68.13	77.03	76.64
RB+C+M	71.47	**93.39**	75.00	**78.97**

(c) ARCT

	Short circuit Tests		Robustness Tests	
Model	**AW**	**CO**	**Original**	**Stress**
BT(w/o)	82.46	50.88	45.60	33.91
BT+B	86.01	61.73	**48.60**	35.99
BT+C	80	**96.17**	47.00	47.72
BT+M	82.48	58.55	46.80	50.02
BT+C+M	**96.79**	87.16	43.60	**53.79**
XL(w/o)	79.64	62.86	56.00	39.77
XL+B	81.40	74.04	**57.0**	44.6
XL+C	**87.87**	**98.90**	54.40	51.66
XL+M	72.76	70.15	53.60	56.99
XL+C+M	48.71	88.56	54.2	**58.63**
RB(w/o)	85.88	70.2	51.00	36.76
RB+B	15.69	73.73	51.00	38.71
RB+C	89.68	**96.83**	50.40	50.88
RB+M	**100.00**	80.38	**52.00**	**59.95**
RB+C+M	89.26	88.43	48.40	55.78

(d) RECLOR

Table 5: Short circuit and Robustness Tests on 4 models with or without(w/o) data augmentation. +B = augmented with back-translation, +C = augmented with crossover, +M = augmented with mutation, CO=crossover, AW=attention weight evaluation. Stress includes all cases in Table 3.

3.3.3 Results

In conclusion, our study demonstrates the importance of addressing model robustness and short circuit issues when developing machine learning models for natural language understanding tasks. By investigating the weaknesses of BERT, XLNet, and RoBERTa models across different datasets, we identified that these models are generally not robust under stress tests, and short circuit issues contribute to their instability.

To overcome these challenges, we proposed and evaluated data augmentation methods, including crossover, mutation, and a combination of the two (+C+M), and compared them with the back-translation baseline. Our results revealed that these data augmentation techniques not only maintain or improve the model's performance on the original dataset but also significantly enhance model robustness under stress tests. In particular, the +C+M method demonstrated the best performance for most of the cases.

Additionally, our findings from the short circuit test showed that the crossover method consistently achieves the highest CO score and often ranks best in the AW score, indicating that models trained with crossover data augmentation are more likely to consider the premise and avoid short circuit issues.

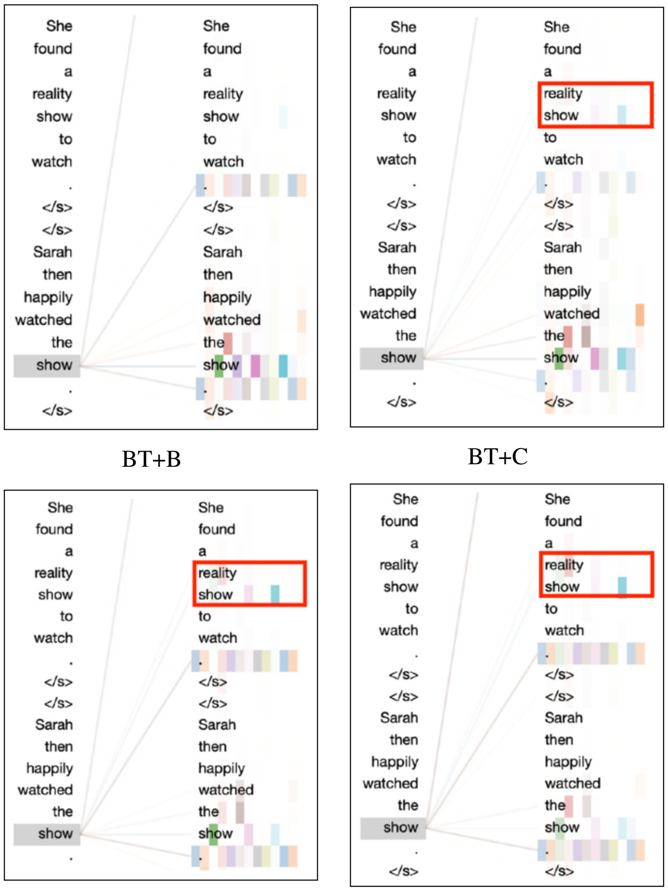
Future work could explore additional data augmentation techniques and their combinations to further enhance model robustness and mitigate short circuit problems. Furthermore, investigating the transferability of these augmentation methods across various natural language understanding tasks and languages could provide valuable insights into the generalizability of these approaches.

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 case in Figure 3. In this example, the word "show" in the premise is strongly related to the token "reality show" in the right choice from human knowledge.

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 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 3 (“BT+M”) and the combination of *crossover* and *mutation* operation in Figure 3 (“BT+C+M”). The rationale behind such a change of attention pattern is that, in an MCQ created by *crossover* operation (“BT+C” in Figure 3), *mutation*(“BT+M” in Figure 3), and the combination of them (“BT+C+M” in Figure 3), 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 3 (“BT+B”) 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.



BT+B BT+C

BT+M BT+C+M

Figure 3: Attention map on a ROC example for BERT-based models.

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 [30], which used back-translation [26], AutoAugment [6], RandAugment [7], and MIXUP [35] which are also mentioned in the survey [9]. 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 generalize to other cases, suggesting that models overfit to the augmentation set [14, 18]. In particular, [21] 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 [10, 5, 17, 28]. 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 [2] 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 [27, 34, 15] 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. [25] 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 [23] and unlexicalized [3, 16]: 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

In this study, we explored the “short circuit” phenomenon in multiple-choice natural language reasoning tasks and proposed white-box and black-box methods to detect such behavior in NLU models. By introducing crossover and mutation operations as data augmentation techniques, we effectively improved model robustness and performance on both stress and original test data, highlighting the importance of refining methodologies to enhance the reliability and robustness of natural language understanding systems.

References

[1] Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard, ‘The sensitivity of language

models and humans to Winograd schema perturbations', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020).

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

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning, 'A large annotated corpus for learning natural language inference', in *EMNLP*, (September 2015).

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

[5] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning, '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*, (2019).

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

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

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *NAACL*, (2019).

[9] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy, 'A survey of data augmentation approaches for nlp', in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968–988, (2021).

[10] Yoav Goldberg, 'Assessing bert's syntactic abilities', *CoRR*, **abs/1901.05287**, (2019).

[11] Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele, 'SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning', in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 394–398, Montréal, Canada, (7-8 June 2012). Association for Computational Linguistics.

[12] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein, 'The argument reasoning comprehension task', *CoRR*, **abs/1708.01425**, (2017).

[13] Shanshan Huang, Kenny Q. Zhu, Qianzi Liao, Libin Shen, and Yinggong Zhao, 'Enhanced story representation by conceptnet for predicting story endings', in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3277–3280, (2020).

[14] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer, 'Adversarial example generation with syntactically controlled paraphrase networks', in *NAACL*, (2018).

[15] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson, 'On feature learning in the presence of spurious correlations', *Advances in Neural Information Processing Systems*, **35**, 38516–38532, (2022).

[16] Nitish Joshi, Xiang Pan, and He He, 'Are all spurious features in natural language alike? an analysis through a causal lens', *arXiv preprint arXiv:2210.14011*, (2022).

[17] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith, 'Linguistic knowledge and transferability of contextual representations', in *NAACL*, (2019).

[18] Nelson F. Liu, Roy Schwartz, and Noah A. Smith, 'Inoculation by fine-tuning: A method for analyzing challenge datasets', in *NAACL*, (2019).

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

[20] Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang, 'Commonsense causal reasoning between short texts', in *Fifteenth international conference on the principles of knowledge representation and reasoning*, (2016).

[21] Tom McCoy, Ellie Pavlick, and Tal Linzen, 'Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference', in *ACL*, (2019).

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

[23] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig, 'Stress test evaluation for natural language inference', in *ICLR*, (August 2018).

[24] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh, 'Beyond accuracy: Behavioral testing of NLP models with CheckList', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020).

[25] Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel, 'Behavior analysis of NLI models: Uncovering the influence of three factors on robustness', in *NAACL*, (2018).

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch, 'Improving neural machine translation models with monolingual data', in *ACL*, pp. 86–96, (2016).

[27] Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh, 'Tackling the story ending biases in the story cloze test', in *ACL*, (2018).

[28] 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*.

[29] Jesse Vig, 'A multiscale visualization of attention in the transformer model', in *ACL*, (2019).

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

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

[32] 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*, eds., Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett.

[33] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng, '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, (2020).

[34] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi, 'SWAG: A large-scale adversarial dataset for grounded commonsense inference', in *EMNLP*, (October-November 2018).

[35] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, 'mixup: Beyond empirical risk minimization', *ICLR*, (2017).