IMPORTANT NOTE (DO NOT DELETE)
The rebuttal to the reviews **must** be restricted to:

- answering the specific "pressing" questions raised by the reviewers in slot 4 of the reviews; and
- pointing out factual errors in the reviews.

Furthermore, it **must be one page-only**.
The goal of the rebuttal is not to enter a dialogue with the reviewers. Therefore, it is **not allowed to include new results in the rebuttal, provide links to such results, or describe how you might rewrite your paper to address concerns of the reviewers**. It is the current version of your paper that is under review, not possible future versions.
The program committee will be instructed to ignore rebuttals violating these principles.

Rebuttal

Review 1, Question 1: Was a statistical analysis performed and...

Answer: We conducted 12 experiments in Table 4: 3 models on 4 different datasets. According to t-tests, with $p < 0.05$, +C+M is significantly more accurate than (w/o) and +B in the stress test of all 12 experiments. For original tests, with $p < 0.05$, 4 of 12 experiments, +C+M is significantly better than (w/o); and in 2 of 12 experiments, +C+M is significantly better than +B. In all other experiments, there are no significant differences between +C+M and (w/o) or +B.

Review 1, Question 2: Can we interpret a performance in he "choice-only" condition...

Answer: In fact, the "choice-only" test is proposed to test whether the problems in a dataset are easy to solve or not. The performance difference of a certain model with the "choice-only" test on different datasets can partly illustrate the degree of bias contained in the datasets. In Figure 5, we can find the "choice-only" test performance of models drops with our data augmentation methods. It indicates there is less bias in the augmented datasets. We enhance model robustness by encouraging models to pay more attention to the premise with augmented data. As a by-product, we have also alleviated the data bias in the datasets.

Review 1, Question 3: I don't understand why stress test operator PI works as intended...

Answer: While the stress operators may not succeed in invalidating the choice as you pointed out, this seems very unlikely. We sampled 100 cases, and after PI operation, only one case failed to be invalidated. In that particular case, and also in the case you pointed out, the reasoning is obviously weakened after the modification.

And 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. 95% of the cases turned out to be correct.

Review 1, Question 4: I have some general reservations about MCQ for text understanding...

Answer: We totally agree with your concern and it is true that MCQs are prone to data biases. Unfortunately, existing datasets for commonsense reasoning in NLP are almost exclusively in the form of MCQs. We will add some discussion about this and suggest it as future work in the revised version of the paper.

Review 2, Question 1: Why mutation works?
Answer:

1. 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.
2. It also further strengthens the model's grammatical capabilities.

For more details, please refer to Sec 2.2 para 4.

Review 2, Question 2: Can the system learn that grammatically incorrect...

Answer: As you think, the system can learn grammatical knowledge in choices. We also illustrated this in Sec 2.2 (fourth paragraph). In our experiments which didn't show in this paper, we test on grammatical test cases and the system gets great performance. However, this test is an in-domain test for mutation which can not show the reasoning ability of models. Thus, it didn't appear in our paper as a stress test. The main contribution of mutation is to teach models to learn the relation between premise and choices.

Review 3, Question 1: Why not augment the dataset by using the stress test operators...

Answer: Please refer to Sec 2.2, para 1:

1. These stress tests cannot generate a sufficient amount of data for training;
2. The purpose of this paper is to design data augment to promote the model's general ability to avoid short-circuit, while most of the stress operators work on a specific linguistic capability.