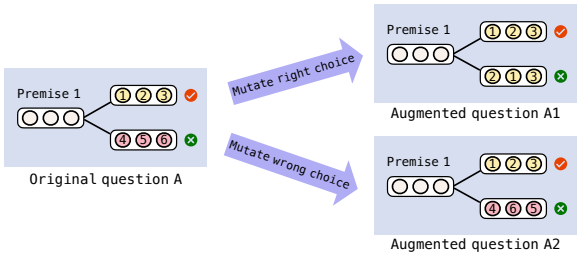
Reviewer #1

Q1: Can you tell me whether this idea has been...
A1: Even though our mutation operator is operationally similar to the random token swapping (RS) operator in previous work (Artetxe et al., 2018; Lample et al., 2018; Wei and Zou, 2019; Miao et al., 2020), the purpose is different. The purpose of RS is to improve models' fault tolerance by perturbing the sentence without changing its meaning; the purpose of our mutation is to encourage the models to look into the premise (avoid short-circuits). Consequently, the way we construct the data is also different which is illustrated in the following figure. We will update Fig 3 in the paper with this figure.



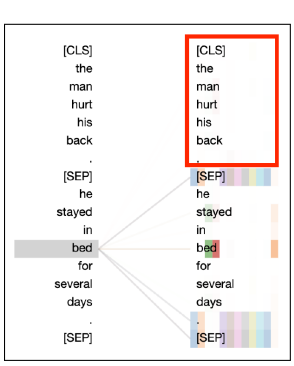
Q2: Can you extend this idea to more settings...
A2: Good suggestion. One of our 4 datasets, RECLOR, is already an MRC dataset. We can include more in the final version.

Reviewer #2

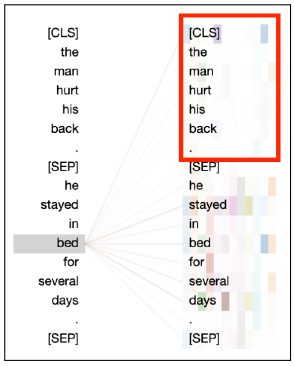
Thank you very much for your good suggestions.
Q1: Mutation in Fig 3 unclear...
A1: Please refer to A1 of Reviewer #1 and the beginning of Sec 2.2 for the motivation of the two operators.
Q2: the dataset size is relatively small...
A2: Please refer to A7 of Reviewer #3.

Reviewer #3

Q1: In Figure 1, is there an impact...
A1: We followed your advice and compare the attention maps of BERT fine-tuned on unlabeled and labeled data in the following figure. Despite some very light color in the premise of Fig (a), BERT+C still shows more obvious effects of "forcing the model to look at premises".



(a) BERT fine-tuned on unlabeled COPA



(b) BERT+C fine-tuned on labeled COPA

Q2: questions and answers are rather single...
A2: Thank you for your good suggestion. What you are suggesting is hierachical representation scheme which is reasonable but uncommon for MCQ type of NL reasoning tasks that this paper is targeting. Moreover, of the four datasets, only two (COPA and ARCT) are single-sentence questions. Hence we chose to experiment on the three more popular encoders in this area. We will consider such hierarchical models in our future work.
Q3: In table 1, about the adverb operator...
A3: We manually selected adverbs ("in fact", "actually", "indeed") as the prefix to the wrong choices. The adverb operator is not used for training but for testing. The purpose of the adv operator is to trick the models into selecting the wrong choices without looking at the premises.
Q4: This may be popular in MCQ papers...
A4: To fine-tune the language models for an MCQ task, we feed LM's final hidden vector to a fully connected layer to compute the probability of the right choice. We will add this bit to the final version.
Q5: In C+M data augmentation scheme...
A5: We noted in Sec 3.1 that "The expanded data volume is equal to the original data volume and the size of new train dataset has doubled." Therefore, 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.
Q6: In Table 4, the average of the 4 datasets...
A6: The purpose of the last two columns is to evaluate the average performance of each model over four different datasets which are equally important for us. Just because a dataset has more test cases doesn't mean that this task is more important than the others. Therefore it is our opinion to not use a weighted average score which is also echoed by other researchers (Liu et al. 2019b; Devlin et al. 2019; Yang et al. 2019). Moreover, these datasets all have a sufficient number of test cases to be statistically significant.
Q7: The baseline seems stronger...
A7: We agree with your observations. You may also notice that in Table 4, there is a wider accuracy gap on the vanilla (w/o) models between original test and stress test for RECLOR than other datasets, which indicates that RECLOR has more bias and is more susceptible to short-circuit. Therefore, the final test results of the vanilla models on the original test data are less reliable. Though our methods slightly underperform the vanilla and +B models in some cases, they have a huge advantage over the vanilla and +B on the stress tests and subtantially reduced the accuracy gap.
Q8: In the introduction, the paragraph starting...
A8: We will revise it to "In contrast, our proposed ..."

Reviewer #4

Thank you for your comments. Unfortunately, there are some misunderstandings in your comments. First, this work targets a wide range of MCQ type tasks in NL reasoning, not just NLI. Second, "crossover" and "mutation" are novel operations we proposed (see A1 of Reviewer #1). Whereas "back-translation" is a strong baseline we compare with. Besides, we also did a detailed analysis in Sec 3.2-3.4.