A. Details of Stress Tests

Table 1 tells more detailed numbers [1] about stress test results with different aspects in Figure 4. (Section 3.2)

B. Details of Choice-only test

In Table 2, 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 1. An MCQ from COPA:

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

In Example 1, we explore RoBERTa-based models by analyzing their attention maps on this question in Figure 1. 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 1, 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 1d. 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 1b seems to have not enhanced such abilities.

Example 2. An MCQ from COPA:

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

(a) RB(w/o)

(b) RB+B

(c) RB+C

(d) RB+C+M

Fig. 1: Attention map on a COPA example for models.

In human cognition, the word "furious" in premise and "slammed" in the right choice have a strong causal relationship in Example 2. However, from the attention map of the vanilla XLNet model in Figure 2, it is difficult to observe that they are related. In Figure 2, 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 3. 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.

[1] The dashes in Table 1 are caused by limited test cases which sizes are too small.

✓
Choice 2: I would not be happy to pay for some rich kids tuition at the same time. ✗

In Example 3, 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 knowledge of commonsense reasoning. Unfortunately, "free" has little relationship with the warrant in Figure 3a 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.

D. Details of Stress Test Human Evaluation

We have generated stress test cases with different operators and the size of cases have shown in Section 3.1. To guarantee the correctness of questions in the stress test, we make human annotation by sampling 100 cases from each kind of stress test. If the total number of a stress test is less than 100, we will consider all cases which have been generated. The pass rate for each test is shown in Table 3. Mostly, the test cases are perfect (100 percent pass rate) except for Neg+ stress test cases. To make Neg+ test more convincing, we annotate all generated Neg+ cases and filter the Neg+ stress test cases from 492 cases to 463 cases. All models are tested on this human filtered test set.

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	56.44	-	-	69.41	74.89	53.93	62.55
	BT+B	75.40	73.22	-	-	85.16	71.54	80.49	77.47
	BT+C	75.73	69.69	-	-	91.87	77.02	70.60	76.94
	BT+M	69.53	74.66	-	-	82.01	81.43	97.29	82.19
	BT+C+M	73.20	77.97	-	-	92.48	86.30	96.47	86.83
	XL(w/o)	63.40	58.53	-	-	55.59	78.54	64.77	62.47
	XL+B	64.80	69.90	-	-	58.74	79.15	50.54	64.81
COPA	XL+C	74.60	73.58	-	-	91.16	96.65	75.34	82.54
	XL+M	66.80	69.69	-	-	63.01	85.84	99.73	76.65
	XL+C+M	72.93	84.44	-	-	89.64	98.18	99.86	91.38
	RB(w/o)	72.00	73.87	-	-	60.36	75.65	64.91	68.90
	RB+B	74.07	82.51	-	-	65.44	82.65	80.63	77.71
	RB+C	77.07	88.55	-	-	94.21	97.41	87.94	91.45
	RB+M	70.47	83.37	-	-	73.07	95.74	99.59	86.01
	RB+C+M	75.67	86.03	-	-	95.73	99.54	99.86	93.63
	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

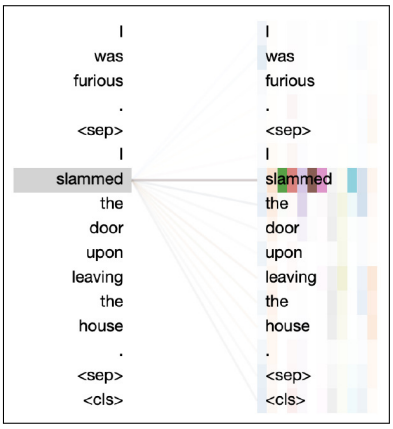
Table 1: Detailed Breakdown of Stress Tests on 4 models with or without(w/o) data augmentation. +B = augmented with backtranslation, +C = augmented with crossover, +M = augmented with mutation. Stress Tests includes the following stress tests: Neg+=negation-add, Neg-=negation-remove, NER, PR=pronoun-replacement, 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

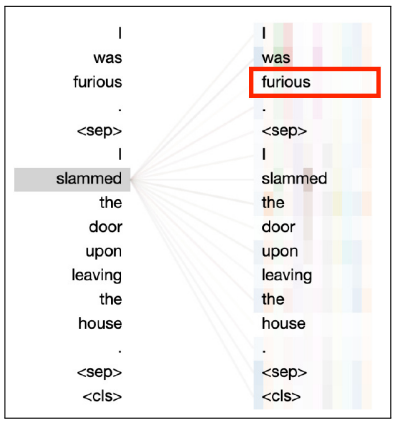
Table 2: Choice-only test for transformer-based models on 4 datasets. All numbers are percentages (%)

Stress	ROC	COPA	ARCT	RECLOR
Neg+	100	94	100	100
Neg-	100	-	100	100
NER	100	-	-	-
PR	100	100	100	100
PI	100	100	100	100
Voice	100	100	100	100

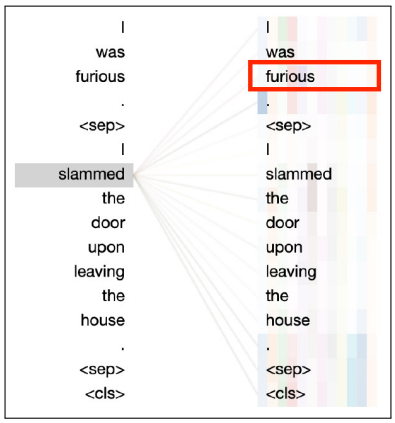
Table 3: Pass rate (%) of stress test cases with human annotation.



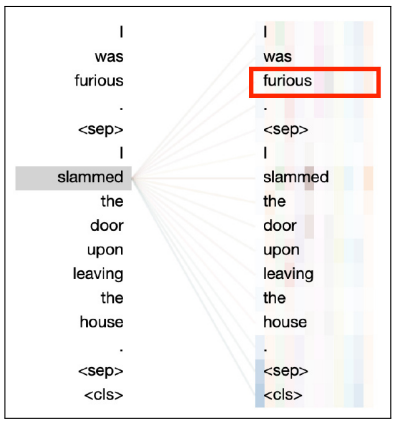
(a) XL(w/o)



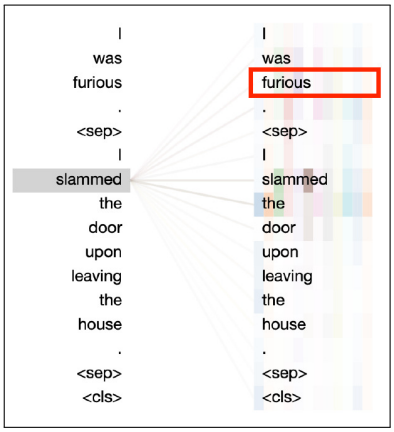
(b) XL+B



(c) XL+C

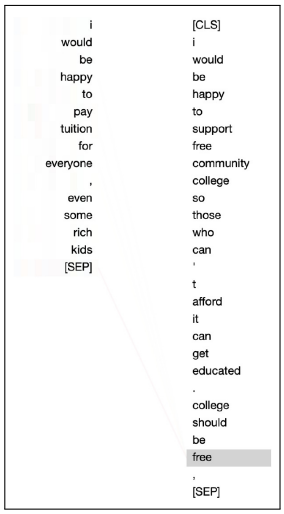


(d) XL+M

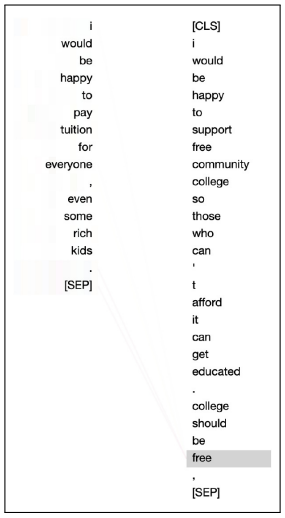


(e) XL+C+M

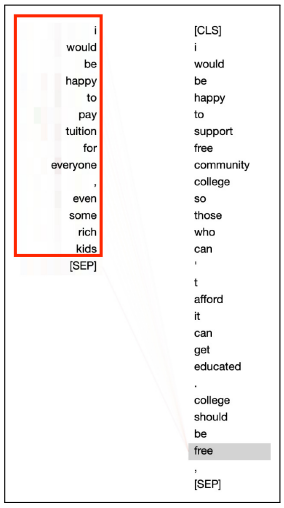
Fig. 2: Attention map on a COPA example for XLNet-based models.



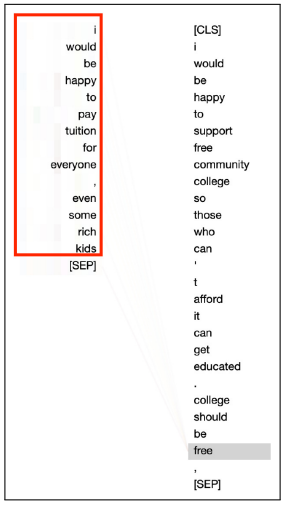
(a) BT(w/o)



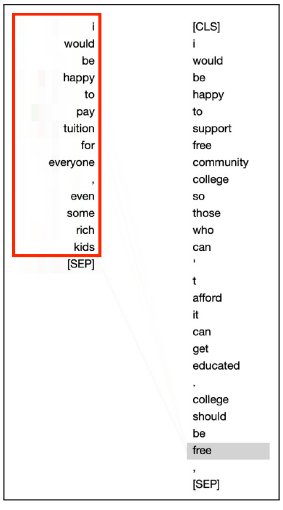
(b) BT+B



(c) BT+C



(d) BT+M



(e) BT+C+M

Fig. 3: Attention map on an ARCT example for BERT-based models.