A Details of Robustness Tests

Dataset	Model	Original	Neg+	Neg-	NER	PR	PI	CO_{rt}	Adv	MT	Voice	Syn	All
ROC	BT(w/o)	86.58	82.19	59.57	78.18	76.61	90.48	86.8	**83.73**	78.19	70.22	80.55	81.93
	BT+B	86.75	**86.14**	60.64	86.46	78.66	94.31	87.97	83.03	76.06	70.22	81.78	82.96
	BT+C	**87.07**	80.08	62.77	**97.79**	88.07	95.93	**96.85**	83.41	73.76	70.12	81.62	84.34
	BT+M	86.48	81.86	**79.79**	93.09	85.0	96.75	88.56	70.22	**96.53**	96.35	72.74	86.06
	BT+C+M	86.75	83.36	77.66	97.24	**93.2**	97.79	94.92	73.89	96.21	**97.53**	**73.35**	**88.6**
	XL(w/o)	**90.81**	88.65	55.32	86.74	50.89	93.5	82.79	**89.68**	66.27	51.58	87.29	79.22
	XL+B	90.43	87.09	60.64	90.88	65.24	94.54	83.97	89.46	73.22	61.24	**88.36**	82.23
	XL+C	89.47	85.7	60.64	**99.17**	91.61	99.65	**98.34**	87.35	71.83	64.6	85.45	86.23
	XL+M	90.17	**87.37**	80.85	96.69	74.09	98.84	83.11	87.68	**99.36**	**98.62**	84.07	89.47
	XL+C+M	90.22	85.53	**81.91**	**99.17**	**93.38**	**99.88**	96.85	86.0	98.61	98.22	87.29	**92.64**
	RB(w/o)	**92.73**	85.7	69.15	75.97	67.94	87.34	86.8	91.62	71.99	60.36	**90.35**	82.33
	RB+B	92.46	**88.26**	62.77	65.19	58.62	77.93	86.37	**91.73**	64.08	43.79	89.89	78.5
	RB+C	91.18	87.76	74.47	99.17	**90.49**	96.75	**98.77**	90.27	78.19	75.64	88.51	88.92
	RB+M	92.62	86.7	80.85	86.46	76.14	93.73	86.37	90.76	**99.36**	**99.51**	85.3	90.29
	RB+C+M	91.88	84.97	**86.17**	**99.45**	88.35	**98.95**	98.02	88.86	99.04	99.21	87.14	**93.06**
COPA	BT(w/o)	62.0	51.42	-	-	55.79	63.47	52.4	55.65	64.4	56.91	52.0	57.41
	BT+B	68.6	64.02	-	-	71.95	69.86	71.8	66.53	69.4	72.36	**72.0**	68.95
	BT+C	**72.8**	69.72	-	-	93.6	77.17	**92.0**	**69.76**	76.4	89.43	68.0	78.84
	BT+M	70.4	72.15	-	-	83.23	**80.37**	81.0	63.1	**99.0**	**99.59**	**72.0**	79.62
	BT+C+M	72.4	**74.8**	-	-	**87.5**	79.91	87.8	63.51	93.8	95.12	**72.0**	**80.68**
	XL(w/o)	61.4	34.15	-	-	54.88	60.27	56.2	57.26	68.8	79.67	60.0	57.71
	XL+B	63.2	**88.62**	-	-	55.49	64.84	61.0	62.5	50.4	24.8	64.0	61.06
	XL+C	**67.8**	60.16	-	-	**83.84**	**97.26**	**93.8**	**69.56**	69.0	79.67	68.0	75.42
	XL+M	62.2	59.96	-	-	56.4	94.52	56.2	62.5	**99.8**	**100.0**	**72.0**	71.1
	XL+C+M	67.2	82.32	-	-	76.83	98.17	80.8	68.75	94.8	**100.0**	68.0	**81.32**
	RB(w/o)	76.4	80.69	-	-	71.04	76.71	76.0	73.59	72.8	67.07	80.0	74.85
	RB+B	77.0	79.67	-	-	82.62	90.87	82.4	72.78	84.4	77.64	80.0	80.26
	RB+C	**79.0**	79.47	-	-	88.41	97.72	**95.0**	**77.82**	78.8	76.83	**88.0**	83.31
	RB+M	72.6	78.46	-	-	86.89	98.63	80.2	71.37	**99.8**	**100.0**	44.0	83.53
	RB+C+M	74.0	**87.2**	-	-	**93.9**	**100.0**	90.4	70.36	99.2	99.59	72.0	**87.3**
ARCT	BT(w/o)	63.96	31.65	88.16	80.0	53.52	60.71	65.77	68.24	47.52	36.78	62.05	58.08
	BT+B	68.47	**37.04**	83.55	60.0	40.85	48.21	58.56	62.39	44.37	29.31	**68.65**	56.21
	BT+C	**68.92**	36.7	85.53	**100.0**	70.42	76.79	**87.84**	**71.17**	47.07	50.57	71.62	65.74
	BT+M	67.79	32.32	91.45	80.0	74.65	82.14	67.34	59.68	**94.82**	**91.95**	62.71	69.65
	BT+C+M	67.57	36.36	**94.08**	**100.0**	**85.92**	**83.93**	86.49	66.89	88.51	91.38	63.37	**73.71**
	XL(w/o)	75.45	39.39	72.37	20.0	30.99	51.79	56.98	**76.13**	56.53	38.51	74.59	61.72
	XL+B	**79.05**	**45.12**	80.26	40.0	64.79	57.14	57.43	68.02	63.06	46.55	**76.24**	64.78
	XL+C	74.55	39.73	84.21	**60.0**	69.01	82.14	**92.34**	72.97	57.66	54.6	73.27	69.94
	XL+M	74.1	41.75	92.11	40.0	70.42	80.36	55.41	72.75	**96.62**	95.4	72.28	73.15
	XL+C+M	77.03	45.12	**95.39**	**60.0**	**85.92**	**92.86**	86.49	72.75	93.24	**95.98**	71.62	**79.11**
	RB(w/o)	78.83	48.82	78.29	60.0	46.48	42.86	57.43	**79.05**	63.29	44.83	77.89	66.16
	RB+B	**81.31**	48.82	76.32	**60.0**	47.89	58.93	62.16	77.7	54.95	44.25	**78.22**	66.02
	RB+C	77.93	45.12	79.61	**60.0**	64.79	69.64	**93.24**	78.15	58.78	38.51	73.93	70.64
	RB+M	77.03	**56.57**	88.16	40.0	**78.87**	85.71	62.39	74.1	**96.4**	**96.55**	72.61	76.64
	RB+C+M	75.0	42.42	**93.42**	**60.0**	74.65	**87.5**	87.61	75.23	93.92	93.68	75.58	**78.97**
RECLOR	BT(w/o)	45.6	21.87	39.5	-	17.39	42.22	47.0	**45.8**	13.8	11.45	44.64	33.91
	BT+B	**48.6**	26.93	42.86	-	15.22	42.22	50.6	45.2	14.0	10.69	**49.48**	35.99
	BT+C	47.0	**27.47**	**56.3**	-	63.04	**95.56**	**93.8**	45.2	26.6	59.54	44.64	47.72
	BT+M	46.8	26.13	53.78	-	56.52	55.56	47.8	39.4	84.6	64.89	38.06	50.02
	BT+C+M	43.6	22.13	55.46	-	**78.26**	77.78	72.0	41.8	**88.8**	**80.92**	41.87	**53.79**
	XL(w/o)	56.0	25.07	44.54	-	30.43	26.67	51.4	**53.2**	18.8	13.74	51.56	39.77
	XL+B	**57.0**	**39.2**	39.5	-	28.26	42.22	64.0	**53.2**	20.8	20.61	**53.98**	44.6
	XL+C	54.4	28.53	60.5	-	73.91	**91.11**	**95.0**	53.0	28.8	51.15	51.9	51.66
	XL+M	53.6	29.33	**66.39**	-	63.04	68.89	61.2	45.8	**92.6**	**77.86**	42.56	56.99
	XL+C+M	54.2	33.07	64.71	-	**65.22**	75.56	73.6	45.6	88.2	77.1	43.6	**58.63**
	RB(w/o)	50.4	**33.07**	60.5	-	**89.13**	**82.22**	95.8	49.6	31.0	54.2	50.52	36.76
	RB+B	51.0	22.13	39.5	-	28.26	28.89	60.2	51.4	19.6	9.92	47.75	38.71
	RB+C	50.4	**33.07**	60.5	-	**89.13**	**82.22**	95.8	49.6	31.0	54.2	50.52	50.88
	RB+M	**52.0**	31.47	**68.91**	-	73.91	80.0	71.6	40.0	96.6	87.79	39.79	59.95
	RB+C+M	48.4	26.4	67.23	-	67.39	77.78	72.8	45.4	89.0	81.68	39.79	55.78.08

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