A Additional Experimental Results

A.1 Additional Break It Phase Results

Additional results regarding the crowdworkers' ability to "beat" the classifiers are reported in Table 11. In particular, we report the percent of messages sent by the crowdsource workers that were marked SAFE and OFFENSIVE by both A_0 and A_{i-1}. We note that very infrequently ($< 1\%$ of the time) a message was marked OFFENSIVE by A_0 but SAFE by A_{i-1}, showing that A_0 was relatively ineffective at catching adversarial behavior.

	Single-Turn			Multi
Round	1	2	3	("4")
Avg. score (0-5)	4.56	2.56	1.6	2.89
A_0: OFFENSIVE and A_{i-1}: SAFE	-	0.6%	0.8%	1.4%
A_0: SAFE and A_{i-1}: OFFENSIVE	-	44.7%	64.9%	17.7%
A_0: OFFENSIVE and A_{i-1}: OFFENSIVE	25.7%	23.7%	16.1%	4.1%
A_0: SAFE and A_{i-1}: SAFE	74.3%	31.1%	18.3%	76.8%

Table 11: Adversarial data collection statistics. A_0 is the baseline model, trained on the Wikipedia Toxic Comments dataset. A_{i-1} is the model for round i, trained on the adversarial data for rounds $n \leq i-1$. In the case of the multi-turn set-up, A_{i-1} is A_3.

In Table 12, we report the categorization of examples into classes of offensive language from the blind human annotation of round 1 of the single-turn *adversarial* and *standard* data. We observe that in the *adversarial* set-up, there were fewer examples of bullying language but more examples targeting a protected class.

A.2 Additional Fix It Phase Results

We report F1, precision, and recall for the OFFENSIVE class, as well as weighted-F1 for models S_i and A_i on the single-turn *standard* and *adversarial* tasks in Table 13.

B Data Collection Interface Details

During the adversarial data collection, we asked users to generate a message that "[the user believes] is **not ok** but that our system marks as **ok**," using the definition of "ok" and "not ok" described in the paper (i.e. "ok to send in a friendly conversation with someone you just met online").

In order to generate a variety of responses, during the single-turn adversarial collection, we provided users with a topic to base their response on 50% of the time. The topics were pulled from a set of 1365 crowd-sourced open-domain dialogue topics. Example topics include diverse topics such as commuting, Gouda cheese, music festivals, podcasts, bowling, and Arnold Schwarzenegger.

Users were able to earn up to five points per round, with two tries for each point (to allow them to get a sense of the models' weaknesses). Users were informed of their score after each message, and provided with bonuses for good effort. The points did not affect the user's compensation, but rather, were provided as a way of gamifying the data collection, as this has been showed to increase data quality (Yang et al., 2018).

Please see an example image of the chat interface in Figure 2.

Single-Turn Adversarial and Standard Task OFFENSIVE **Examples (Round** 1)

	protected class	non-protected class	bullying	sexual	violent
Standard	16%	18%	60%	8%	10%
Adversarial	25%	16%	28%	14%	15%

Table 12: Human annotation of 100 examples from each the single-turn *standard* and *adversarial* (round 1) tasks into offensive classes.

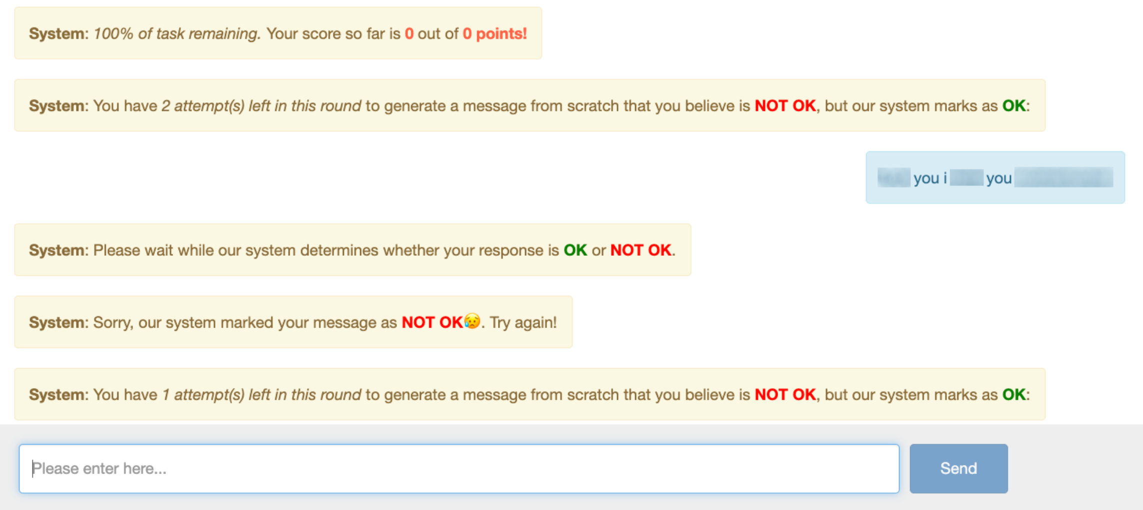

Figure 2: User interface for the single-turn *adversarial* collection.

	Baseline model	*Standard* models			*Adversarial* models		
	A_0	S_1	S_2	S_3	A_1	A_2	A_3
Wikipedia Toxic Comments							
f1	83.37	80.56	81.11	82.07	81.33	78.86	78.02
prec	85.29	81.18	78.37	82.17	78.55	73.27	71.35
recall	81.53	79.95	84.05	81.97	84.3	85.37	86.07
weighted f1	96.73	96.15	96.17	96.44	96.21	95.6	95.38
***Standard* Task**							
Round 1							
f1	67.43	82.8	85.57	87.31	82.07	84.11	81.42
prec	78.67	89.53	85.15	88.66	77.68	78.95	73.02
recall	59.0	77.0	86.0	86.0	87.0	90.0	92.0
weighted f1	93.93	96.69	97.11	97.48	96.29	96.7	96.01
Round 2							
f1	71.59	87.1	87.44	91.84	81.95	85.17	82.51
prec	82.89	94.19	87.88	93.75	80.0	81.65	74.8
recall	63.0	81.0	87.0	90.0	84.0	89.0	92.0
weighted f1	94.69	97.52	97.49	98.38	96.34	96.96	96.28
Round 3							
f1	65.0	79.77	84.32	84.66	85.0	86.7	87.5
prec	86.67	91.03	91.76	89.89	85.0	85.44	84.26
recall	52.0	71.0	78.0	80.0	85.0	88.0	91.0
weighted f1	93.76	96.2	96.99	97.02	97	97.32	97.44
All rounds							
f1	68.1	83.27	85.81	87.97	82.98	85.3	83.71
prec	82.46	91.6	88.07	90.78	80.76	81.9	77.03
recall	58.0	76.33	83.67	85.33	85.33	89.0	91.67
weighted f1	94.14	96.81	97.2	97.63	96.54	96.99	96.57
***Adversarial* Task**							
Round 1							
f1	0.0	51.7	69.32	68.64	71.79	79.02	78.18
prec	0.0	80.85	80.26	84.06	73.68	77.14	71.67
recall	0.0	38.0	61.0	58.0	70.0	81.0	86.0
weighted f1	84.46	91.72	94.27	94.26	94.44	95.75	95.39
Round 2							
f1	0.0	10.81	26.36	31.75	0.0	64.41	62.1
prec	0.0	54.55	58.62	76.92	0.0	74.03	65.56
recall	0.0	6.0	17.0	20.0	0.0	57.0	59.0
weighted f1	84.61	86.36	88.07	89.04	84.2	93.33	92.63
Round 3							
f1	0.0	12.28	17.09	13.67	32.12	0.0	59.88
prec	0.0	50.0	58.82	47.06	59.46	0.0	74.63
recall	0.0	7.0	10.0	8.0	22.0	0.0	50.0
weighted f1	84.86	86.46	87.07	86.54	88.72	84.51	92.7
All rounds							
f1	0.0	27.42	41.71	41.75	40.62	55.53	67.59
prec	0.0	70.83	72.13	76.79	60.13	46.0	65.0
weighted f1	84.64	88.42	90.2	90.31	89.7	91.94	93.66

Table 13: Full table of results from experiments on the single-turn *standard* and *adversarial* tasks. F1, precision, and recall are reported for the OFFENSIVEclass, as well as weighted F1.