Model Evaluation Results:

The data collected in different phases of the ablation study is evaluated using RoBERTa, trained on SNLI. The 4 phases are: regular crowdsourcing, traffic signal feedback, AutoFix, and Full System availability.

Data	**ROBERTA (S2)**	**S22**
Crowdsourcing	1	1.000
Traffic Signal	1	0.856
AutoFix	1	0.803
Full System	1	0.748
Full Data (SNLI Dev set)	0.89	--

DQI Evaluation Results:

Parameters:

T1 : Number of unique vocabulary words/Size of Dataset
T2 : Sentence Length Deviation
C1 : DQI component C1: Vocabulary
W1 : Standard Deviation of Word Distribution
W2 : Proportion of words with sufficient inductive bias
Bi T1 : Standard Deviation of Bigram Distribution
Bi T2 : Proportion of bigrams with sufficient inductive bias
Word : Raw count of Words
S : Raw count of Sentences
Adj : Raw count of Adjectives
Adv : Raw count of Adverbs
N : Raw count of Nouns
V : Raw count of Verbs
Bi : Raw count of Bigrams
Tri : Raw count of Trigrams
C3 : DQI Component 3 : Inter-Sample Semantic Textual Similarity
C4 : DQI Component 4: Inter-Sample Word Similarity
WO : Average premise-hypothesis overlap across samples
STS : Average premise-hypothesis similarity across samples

Results for Original SNLI Samples (Sentence 1 and Sentence 2):

Data	T1	T2	C1	W1	W2	Bi T1	Bi T2	Word	S	Adj	Adv	N	V	Bi	Tri	C3
Crowdsourcing	6.07	5.55	11.62	144.93	0.57	621.01	0.88	419	138	99	9	201	88	994	1149	202
Traffic Signal	6.43	5.33	11.76	170.09	0.71	662.55	0.86	444	138	88	17	213	102	1027	1167	202
AutoFix	5.81	5.08	10.89	131.92	0.6	564.54	0.89	401	138	83	12	204	80	938	1041	202
Full	6.62	5.85	12.48	203.01	0.69	922.06	0.96	457	138	90	8	245	89	1055	1156	202

Results for Original SNLI Samples (Sentence 1 and Sentence 2)- Here the last 12 columns represent values for the entailment label:

Data	C4	WO	STS	W1	W2	Bi T1	Bi T2	Word	S	Adj	Adv	N	V	Bi	Tri
Crowdsourcing	0.0008	0.00074	0.0001	108.08	0.36	362.27	0.83	163	46	35	4	69	36	347	369
Traffic Signal	0.0008	0.007	0.0001	108.54	0.43	406.45	0.84	167	46	32	1	83	41	346	363
AutoFix	0.0007	0.008	0.0001	98.14	0.34	477.10	0.89	164	46	37	6	73	37	350	360
Full	0.0008	0.00074	0.0001	134.56	0.35	601.65	0.93	169	46	28	3	91	35	366	370

Results for Original SNLI Samples (Sentence 1 and Sentence 2)- Here the columns represent values for the neutral label:

Data	W1	W2	Bi T1	Bi T2	Word	S	Adj	Adv	N	V	Bi	Tri
Crowdsourcing	126.35	0.36	531.05	0.90	181	46	43	1	89	38	390	407
Traffic Signal	210.84	0.78	668.42	0.94	224	46	45	16	93	54	455	465
AutoFix	136.15	0.48	364.03	0.81	172	46	35	3	87	35	339	350
Full	162.47	0.40	749.45	0.97	178	46	40	3	85	39	385	381

Results for Original SNLI Samples (Sentence 1 and Sentence 2)- Here the columns represent values for the neutral label:

Data	W1	W2	Bi T1	Bi T2	Word	S	Adj	Adv	N	V	Bi	Tri
Crowdsourcing	144.37	0.39	492.82	0.87	194	46	45	3	103	35	401	410
Traffic Signal	131.17	0.47	369.08	0.81	168	46	32	5	85	36	342	369
AutoFix	106.91	**0.33**	390.213	0.87	170	46	27	5	85	36	353	372
Full	170.83	0.57	618.19	1	212	46	53	4	113	34	412	429

Results for VAIDA's Produced Samples (Sentence 1 and Sentence 22):

Data	T1	T2	C1	W1	W2	BiT1	BiT2	Word	S	Adj	Adv	N	V	Bi	Tri	C3
Crowdsourcing	6.2466	5.558	11.883	156.32	0.62	687.91	0.89	431	138	89	6	223	91	1027	1145	202
Traffic Signal	7.339	5.222	12.662	248.08	0.81	919.53	0.88	503	136	109	14	117	237	1117	1219	202
AutoFix	7.008	5.004	12.113	237.01	0.72	886.36	0.91	489	138	89	20	247	104	1030	1071	202
Full	7.334	5.955	13.330	263.46	0.78	1149.68	1	507	138	111	12	247	113	1107	1173	202

Results for VAIDA's Produced Samples (Sentence 1 and Sentence 22)- Here the last 12 columns represent values for the entailment label:

Data	C4	WO	STS	W1	W2	BiT1	BiT2	Word	S	Adj	Adv	N	V	Bi	Tri
Crowdsourcing	0.01	0.08	0.0001	114.09	0.39	446.88	0.92	169	46	37	2	82	38	355	367
Traffic Signal	0.01	0.07	0.001	162.89	0.56	708.30	0.88	200	46	45	8	79	54	418	425
AutoFix	0.001	0.008	0.0001	168.45	0.50	760.25	0.93	201	46	37	11	96	43	396	395
Full	0.	0.	0.	181	0.5	84	0.9	204	46	36	2	11	37	403	403

	01	09	001	.25		3.34	8					5			

Results for VAIDA's Produced Samples (Sentence 1 and Sentence 22)- Here the last 12 columns represent values for the neutral label:

Data	W1	W2	Bi T1	Bi T2	Word	S	Adj	Adv	N	V	Bi	Tri
Crowdsourcing	128.41	0.46	522.37	0.90	173	46	37	2	81	43	361	369
Traffic Signal	251.87	0.63	907.04	0.98	223	46	48	13	95	54	439	442
AutoFix	217.46	0.72	528.88	0.83	198	46	43	7	93	42	355	339
Full	212.53	0.57	1132.03	0.98	190	46	35	6	88	48	391	367

Results for VAIDA's Produced Samples (Sentence 1 and Sentence 22)- Here the last 12 columns represent values for the contradiction label:

Data	W1	W2	Bi T1	Bi T2	Word	S	Adj	Adv	N	V	Bi	Tri
Crowdsourcing	156.21	0.56	485.54	0.89	202	46	37	1	115	36	427	441
Traffic Signal	238.90	0.68	538.99	0.88	209	46	44	6	102	44	392	398
AutoFix	184.01	0.57	602.67	0.88	189	46	36	6	93	40	367	362

Full	214.29	0.71	745.73	0.93	223	46	54	6	114	41	417	422