

# Appendix for Item Response Theory for Human Efficient Evaluation of Chatbots

## 1 Further Human Evaluation Details

Crowd workers are paid \$0.01 per prompt, and on average it takes 1 minute to evaluate 10 choices with a maximum allowed time of 2 minutes. We used three evaluators per prompt, so, if there are 200 prompts, we have 600 ratings and the net cost of the experiment is \$7.2. We chose 3 annotators since we can generalize enough for IAA and it is cost-effective.

**Rate the Chatbot's Responses** (Click to collapse)

Consider the following exchange between two speakers.

Your task is to decide which response sounds better given the previous things said.

If both responses are equally good, click "It's a tie."

**Example:**  
Speaker A: can i get you something from the cafe?

Speaker B: coffee would be great
Speaker B: I don't know what to say.

In this case, the first response is better as it directly answers Speaker A's question, so you should click the bubble next to it.

You must click the Submit button when you are finished. You must complete every question before you can click Submit.

Figure 1: The instructions seen by AMT workers.

The instructions seen by AMT workers are shown in Figure 1.

We removed workers with a correlation below 0.05 with other annotators. For a worker identified as "bad", all annotations are removed. Including these workers only increases the standard error by 10%.

From the 200 NCM evaluation set prompts,

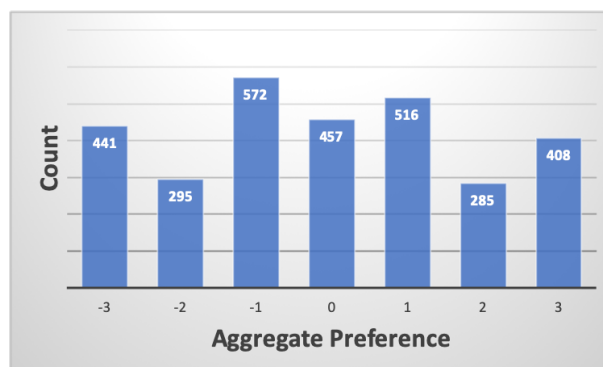


Figure 2: A histogram of aggregated preferences,  $\sum_i \sum_j u_j^i$ , across all prompts and model comparisons by all annotators.

each annotation task has 10 prompts; however, we do not pair the same 3 workers to the 10 prompts; instead we randomize the prompts shown, so worker 1 many compare prompts 1-10, while worker 2 compares prompts 2,3,5,7,9,11,13,17,19,23. As a result, the correlation between one worker and the others is more stable.

A full set of model comparisons on the Neural Conversation Model is available in Table 1.

### 1.1 Rating Distribution

Figure 2 shows a histogram of the grades over all experiments run.

<b>System A</b>	<b>System B</b>	<b>Mean <math>\Delta</math> Ability</b>	<b>Std <math>\Delta</math> Ability</b>
Cakechat	Seq2SeqAttn_Twitter	-0.529*	0.268
Cakechat	OpenNMT_Seq2SeqAttn	0.125	0.262
Seq2SeqAttn_OpenSubtitles	Cakechat	-0.460	0.281
Seq2SeqAttn_OpenSubtitles_without_PTE	Seq2SeqAttn_OpenSubtitles	0.088	0.273
Seq2SeqAttn_Twitter_without_PTE	Seq2SeqAttn_Twitter	0.424	0.273
Cakechat	NCM	1.314*	0.310
Human1	Seq2SeqAttn_Twitter	-1.98*	0.269
Human1	Human2	0.356	0.256
NCM	Cakechat	-0.715*	0.261
NCM	Seq2SeqAttn_Twitter	-1.426*	0.274
NCM	OpenNMT_Seq2SeqAttn	-1.034*	0.287
NCM	Human1	-0.224	0.262
NCM	Human2	0.377	0.324
Seq2SeqAttn_OpenSubtitles	Seq2SeqAttn_OpenSubtitles	0.295	0.274
OpenNMT_Seq2SeqAttn	Seq2SeqAttn_OpenSubtitles	-0.177	0.318
Seq2SeqAttn_OpenSubtitles_Questions	Human2	2.015*	0.265
Seq2SeqAttn_OpenSubtitles	Seq2SeqAttn_Twitter	0.052	0.274
Seq2SeqAttn_Twitter	Human2	2.760*	0.291
NCM	DialoGPT	-0.223	0.245
NCM	Blender (2.7B)	-0.347	0.256

Table 1: Comparison of various models using IRT. Larger positive indicates that System B is superior in terms of rating by human annotators and similarly smaller negative numbers mean that System A is superior. (\* shows significant differences.)