Achieving Reliable Human Assessment of Open-Domain Dialogue Systems

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu

Science Foundation Ireland in the ADAPT Centre is funded under the SFI Research Centres Programme (Grants 13/RC/2106_P2; 13/RC/2106) co-funded under the European Regional Development Fund.

Dialogue evaluation: existing challenges

- Main evaluation metrics → reference based (BLEU, ROUGE, …) have known issues
  - Unfairly penalize for not corresponding closely with references
  - Ignore dialogue history
  - Weak to no correlation with human evaluation

- Reference-free metrics → Deemed to perform better, according to their correlation with human judgement
  - Issues with results for reference-free metrics
    - Mean correlations are reported but difficult to interpret – correlation coefficients are not additive!!
    - Inter-annotator agreement of expert-based human evaluation may vary ranging from as low as 0.298
    - Such metrics generally require extra resources for training

- Human evaluation: challenges remain
  - Common practice filtering systems via automatic metrics (e.g., ConvAI2 and DSTC6) may inadvertently filter out the best system according to human judgement
  - Live human evaluation is also highly challenging due to lack of method to quality check crowd-sourced human assessors; ConvAI2 live evaluation reported as **senseless** or even **offensive**, and discarded.
  - Many human evaluation methods - data and detailed evaluation techniques are unavailable for the public

Likert Statement & Continuous Rating Scale

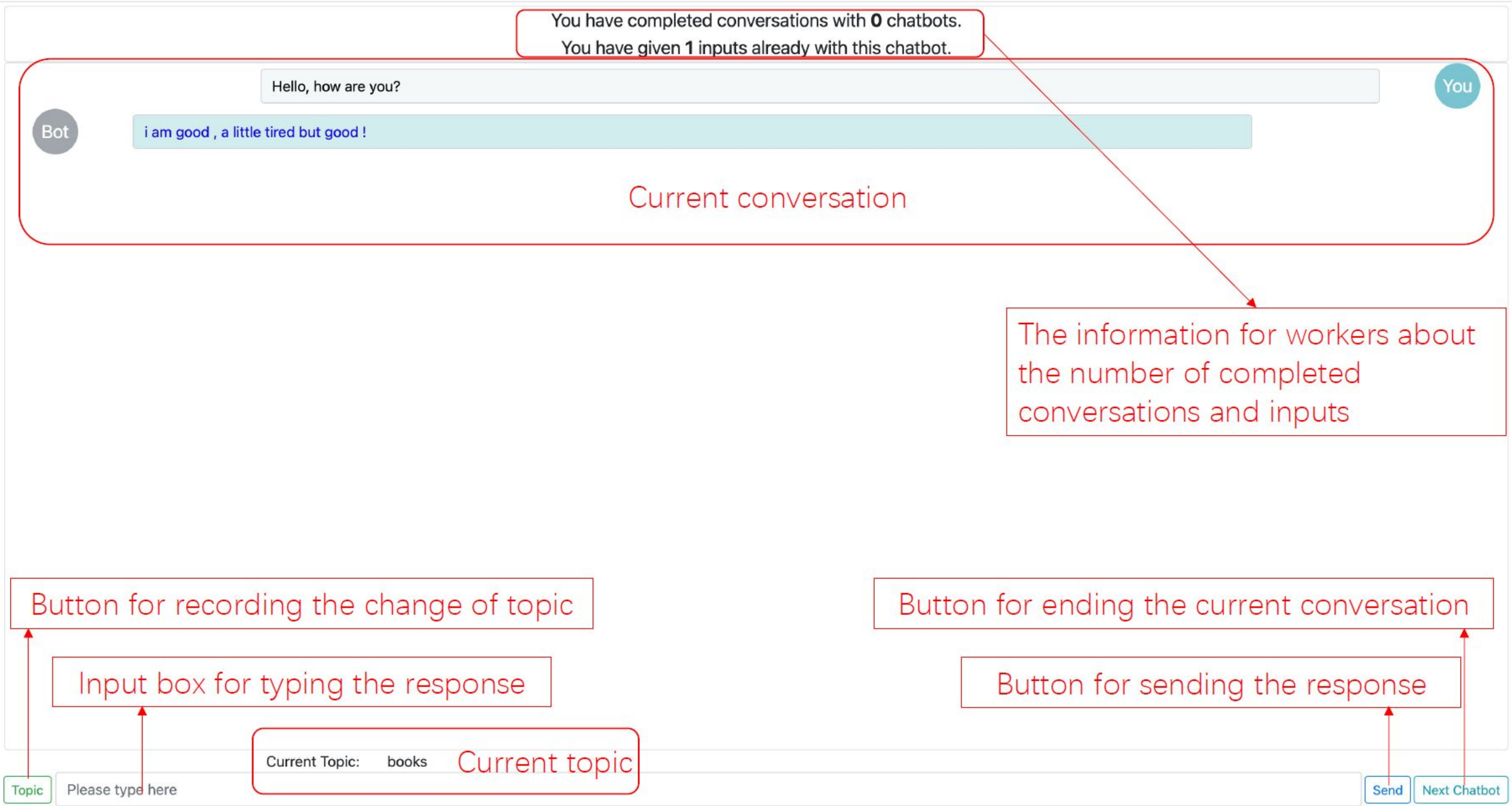
Robotic:	*It was obvious that I was talking to a chatbot as opposed to another human user.*
Interesting:	*The conversation with the chatbot was interesting.*
Fun:	*The conversation with the chatbot was fun/enjoyable.*
Consistent:	*The chatbot was consistent throughout the conversation.*
Fluent:	*The chatbot's English was fluent and natural throughout the conversation.*
Repetitive:	*I felt that the chatbot kept being repetitive during the conversation.*
Topic:	*The chatbot stays on topic.*

Likert Statement
- Adjectival scale labels shown to introduce bias
- Instead use Likert declarative statement
- Workers are asked to rate agreement with statement

Continuous Rating Scale
- Reduce bias by score standardization
- Standard significance tests to score distributions
- Accurate quality control of crowd-sourced workers

Live Dialogue Evaluation
- Direct Assessment by the user
- User chosen topic – genuinely open domain
- Switch topic possible

User interface – interact with a model



Quality-control Live Dialogue Evaluation

Deploy models that have known distinct performance levels in each Human Intelligence Task (HIT)

- 5 (genuine) dialogue models and a quality-control model
- Quality-control model only returns a degraded random response of which a random substring is replaced by another random string
- The model order is shuffled and invisible - blind human evaluation

Given a HIT that has six models, a crowd-sourced worker is asked to take following steps:

1. Converse with a model (at least 10 turns)
2. Rate the quality of current conversation.
3. Repeat step 1 and 2 until all six models are rated.

Statistical significance tests are then applied score distributions of workers for the ratings they attributed to genuine models, relative to the quality-control model.

- Any worker with $p < 0.05$ is retained

The computation of system-level scores

After quality control, system-level scores computed

- Scores for negative attributes reversed (i.e., robotic and repetitive) 100 − the original rating
- Each worker's mean and standard deviation computed
- Raw scores are then **standardized** according to worker's mean and standard deviation to remove bias from overly harsh or lenient judges
- **Average** standardized scores for each criteria are calculated
- The **overall score** is calculated as the average of all measurement criteria.

Dialogue models in this experiment

We employ following 5 models from ParlAI that are pre-trained on the ConvAI2 dataset

- Poly-encoder Transformer
- Bi-encoder Transformer
- Sequence to sequence
- Key-value memory network
- LSTM-based

Each model is with a persona (approximately five textual statements), and we additionally include a version of each of the above models without any persona, resulting in 10 models.

- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. CoRR, abs/1905.01969.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. CoRR, abs/1811.01241.
- Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir- Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. CoRR, abs/1606.03126.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

User interface – rating after conversation

Please say how much you agree with each of the following statements:

It was obvious that I was talking to a chatbot as opposed to another human user.

strongly disagree *strongly agree*

The conversation with the chatbot was interesting.

strongly disagree *strongly agree*

The conversation with the chatbot was fun/enjoyable.

strongly disagree *strongly agree*

The chatbot was consistent throughout the conversation.

strongly disagree *strongly agree*

The chatbot's English was fluent and natural throughout the conversation.

strongly disagree *strongly agree*

I felt that the chatbot kept being repetitive during the conversation.

strongly disagree *strongly agree*

The chatbot stays on the topic.

strongly disagree *strongly agree*

NEXT

Experiment – Choice of topic

Two settings of experiments with regard to topic

- Workers can choose a topic freely before a conversation (Free)
- A topic is given to workers before a conversation (Ice-breaker)

Additionally, a second run of Free Topic is employed as the self-replication experiment.

Topic	Workers			Ave. Duration (min)			Dialogues		
	Total	Passed	Pass Rate	Passed	Failed	All	Total	Passed	Pass Rate
Free Run 1	249	173	69.5%	6.53	7.04	6.68	1,525	1,075	70.5%
Free Run 2	248	139	56.0%	6.87	7.58	7.18	1,480	838	56.6%
Ice-breaker	248	171	69.0%	6.60	6.70	6.63	1,450	1,030	71.0%

Table 1: Numbers of workers, average time taken per dialogue, and total number of dialogues

Experiment – User Chosen Topic

	Model	n	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
Free Run 1	A	798	0.534	0.564	0.602	0.711	0.863	0.964	−0.038	0.069
	B	798	0.419	0.474	0.481	0.614	0.875	0.994	−0.431	−0.075
	A_p	707	0.318	0.399	0.372	0.443	0.821	0.404	−0.330	0.116
	C	791	0.262	0.491	0.379	0.028	0.636	−0.066	−0.316	0.680
	C_p	714	0.189	0.409	0.373	0.159	0.672	−0.114	−0.521	0.349
	B_p	707	0.173	0.230	0.197	0.369	0.673	0.320	−0.395	−0.187
	D	707	−0.087	−0.190	−0.208	0.166	0.311	0.401	−0.637	−0.449
	D_p	798	−0.201	−0.308	−0.234	0.092	0.312	0.025	−0.625	−0.669
	E_p	763	−0.217	−0.181	−0.201	−0.196	0.380	−0.455	−0.605	−0.264
	E	742	−0.243	−0.165	−0.160	−0.142	0.329	−0.407	−0.745	−0.411
	r	–	0.969	0.952	0.927	0.899	0.960	0.951	0.646	0.936

Average standardized scores for models in initial data collection run; workers were free to choose the topic of conversation (Free run 1); the correlation (r) between systems in this and a second data collection run distinct data collection runs; where A=Bi-Encoder Transformer, B=Poly-Encoder Transformer, C=Key-Value Memory Network, D=Sequence to Sequence, and E=LSTM-based Model; models with p models with a the persona; score for robotic and repetitive have been reversed; n is number of ratings; models ordered by overall average score.

Experiment – Ice-breaker Topic Prescribed

	Model	n	Overall	Interesting	Fun	Consistent	Fluent	Topic	Robotic	Repetitive
Ice-breaker	A	721	0.552	0.565	0.527	0.873	1.018	1.011	−0.287	0.156
	A_p	742	0.422	0.589	0.560	0.518	0.718	0.527	0.009	0.034
	B	721	0.376	0.379	0.340	0.634	0.769	0.820	−0.221	−0.087
	C	784	0.322	0.615	0.537	0.190	0.631	0.061	−0.344	0.565
	B_p	658	0.273	0.406	0.340	0.414	0.633	0.423	−0.369	0.063
	C_p	700	0.222	0.402	0.337	0.089	0.654	−0.068	−0.376	0.514
	D	728	−0.139	−0.277	−0.204	0.123	0.349	0.295	−0.638	−0.620
	E_p	714	−0.198	−0.172	−0.203	−0.054	0.316	−0.343	−0.533	−0.396
	E	721	−0.240	−0.125	−0.161	−0.196	0.318	−0.393	−0.631	−0.489
	D_p	721	−0.267	−0.426	−0.402	−0.011	0.234	0.000	−0.628	−0.636
	r	–	0.984	0.967	0.944	0.958	0.951	0.981	0.715	0.950

Average standardized scores for models in human evaluation where workers were prescribed an ice-breaker topic of conversation sampled from the persona of the model; the correlation (r) between these scores and Free run 1 in Table 3; models are consistent with Table 3; n is number of ratings; models without p did not have a persona (ice-breaker statement was subsequently unknown to these models).

Experiment – Human Assessor Consistency

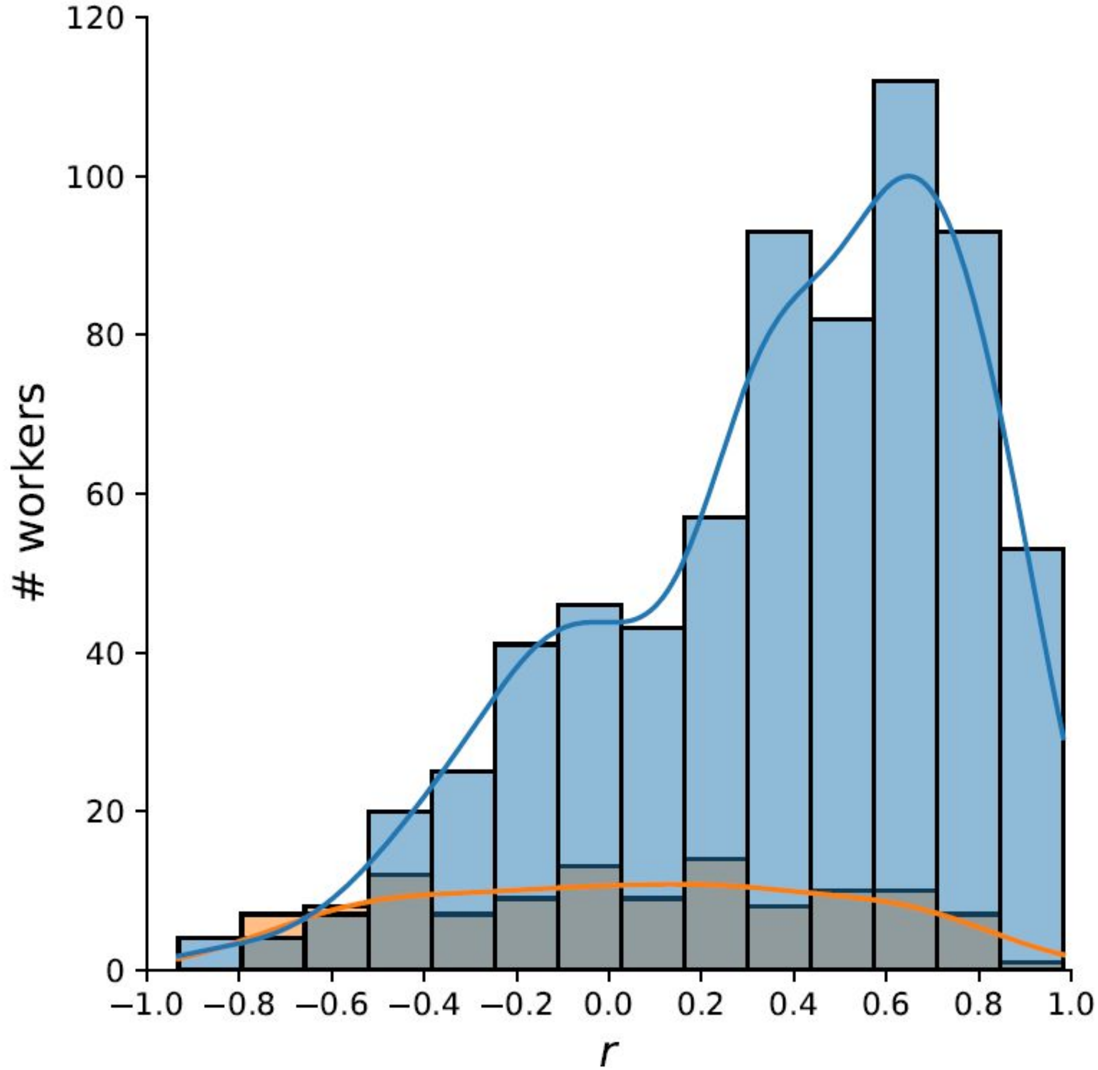


Figure 1: Agreement between pairs of human assessors as measured by the Pearson correlation (r) of ratings provided by workers who passed (blue) and failed quality control (orange).

Experiment – comparison with automatic evaluation metrics

- Word-overlap-based Metrics: BLEU, ROUGE-L, METEOR, GLEU

Metric	r
BLEU-4	-0.883
BLEU-1	-0.707
ROUGE-L	-0.799
METEOR	-0.321
GLEU	-0.816

Table 5: Pearson correlation (r) of word-overlap metric scores and human evaluation.

Experiment – comparison with automatic evaluation metrics

- Word-overlap-based Metrics: BLEU, ROUGE-L, METEOR, GLEU
- Severe lack of correlation with human assessment!! (but not surprising)

Metric	r
BLEU-4	−0.883
BLEU-1	−0.707
ROUGE-L	−0.799
METEOR	−0.321
GLEU	−0.816

Table 5: Pearson correlation (r) of word-overlap metric scores and human evaluation.

Experiment – comparison with automatic evaluation metrics

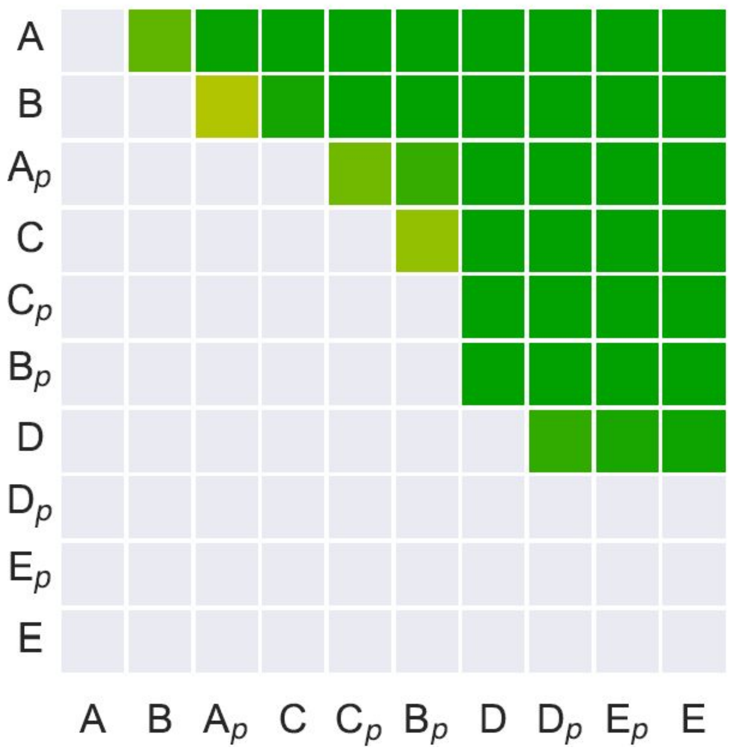
- Reference-free Metrics: FED, USR

	FED_m	FED_l	USR	USR-MLM	USR-DR(c)	USR-DR(f)
Overall	0.590	0.530	−0.230	−0.419	0.046	0.205
Interesting	0.028	−0.042	−0.451	−0.235	−0.238	−0.081
Fun	−0.339	0.115	−0.378	−0.319	−0.131	0.032
Consistent	0.236	0.227	0.214	−0.620	0.518	0.652
Fluent	−0.138	−0.054	−0.227	−0.374	0.028	0.151
Robotic	0.528	0.461	−0.070	−0.290	0.106	0.191
Repetitive	0.841	0.752	−0.713	0.182	−0.690	−0.568
Topic	0.046	0.004	0.222	−0.754	0.606	0.746

Table 6: Pearson correlation (r) of reference free metric scores and human evaluation, where FED_m and FED_l respectively use medium and large DialoGPT, USR is the overall USR score computed according to three sub-metrics: USR-MLM, USR-DR(c) and USR-DR(f).

Persona Contribution Experiment

- Investigate persona contribution to conversation quality
- Conclusion: persona *diminishes* conversation quality in general



- Systems with _p denote same model with persona
- Green cell denotes significant win of model in that row over model in a given column

Conclusion

Overcome previous challenges and provide a new human evaluation methodology that has the following advantages:

- New method **highly consistent** with results for models correlating at ***r* = *0.969*** in two separate data collection runs;
- It has a highly accurate means of quality-control of crowd-sourced workers – ***first dialogue human evaluation to be scalable and repeatable while making data and code public***
- Irons out differences in scoring strategies via score standardization
- It has applicability of standard significance testing while increasing the reliability of results

If you want to use this evaluation, please let us know, we can help!

Thanks and questions …

Topic Change

Topic Change

What is happening to the conversation topic?

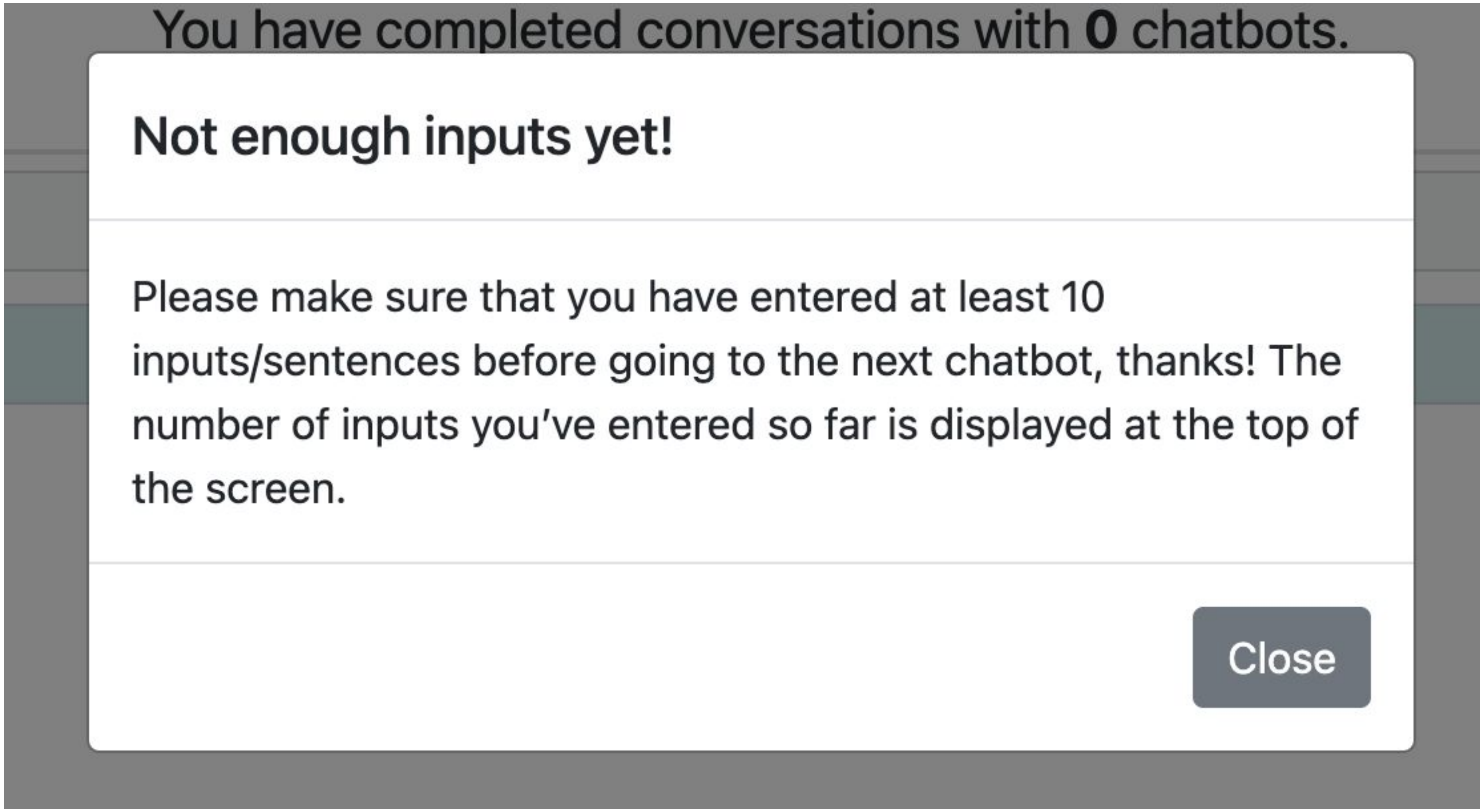
- (selected) The chatbot just changed the topic.
- I will change the topic in my next input
- I changed the topic in my last input.
- No change.

According to the chatbot statements about topic books, what do you think the chatbot's overall feeling about it was?

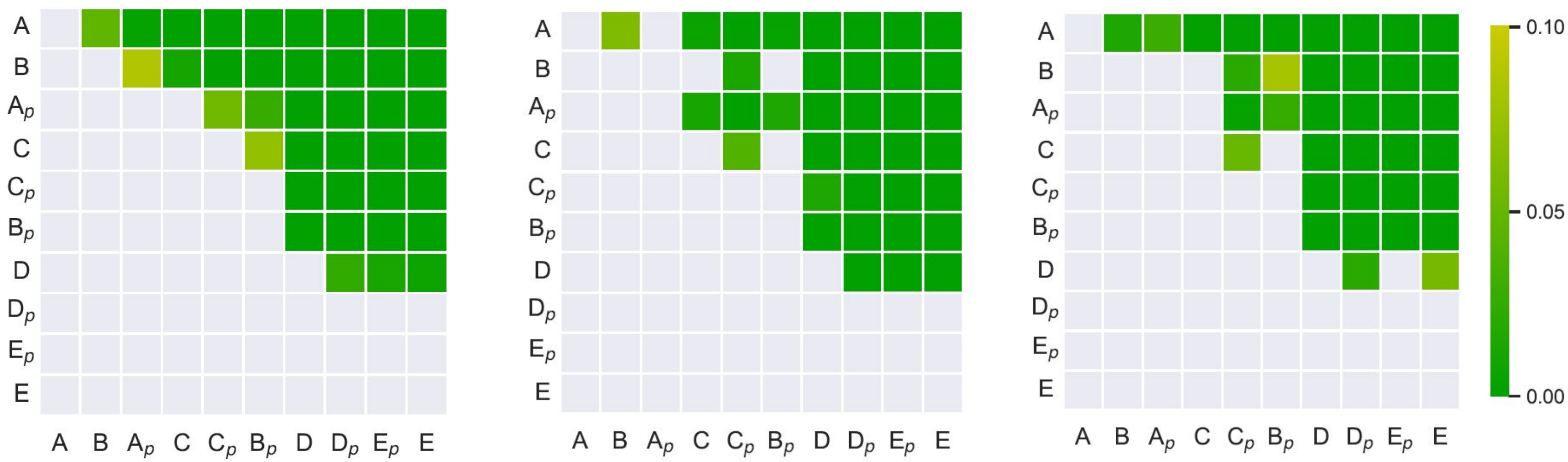
- The chatbot persona likes it.
- The chatbot persona dislikes it.
- (selected) The chatbot persona is ambivalent about it.

Submit Cancel

User interface – warning of insufficient turns



Experiment – significance test

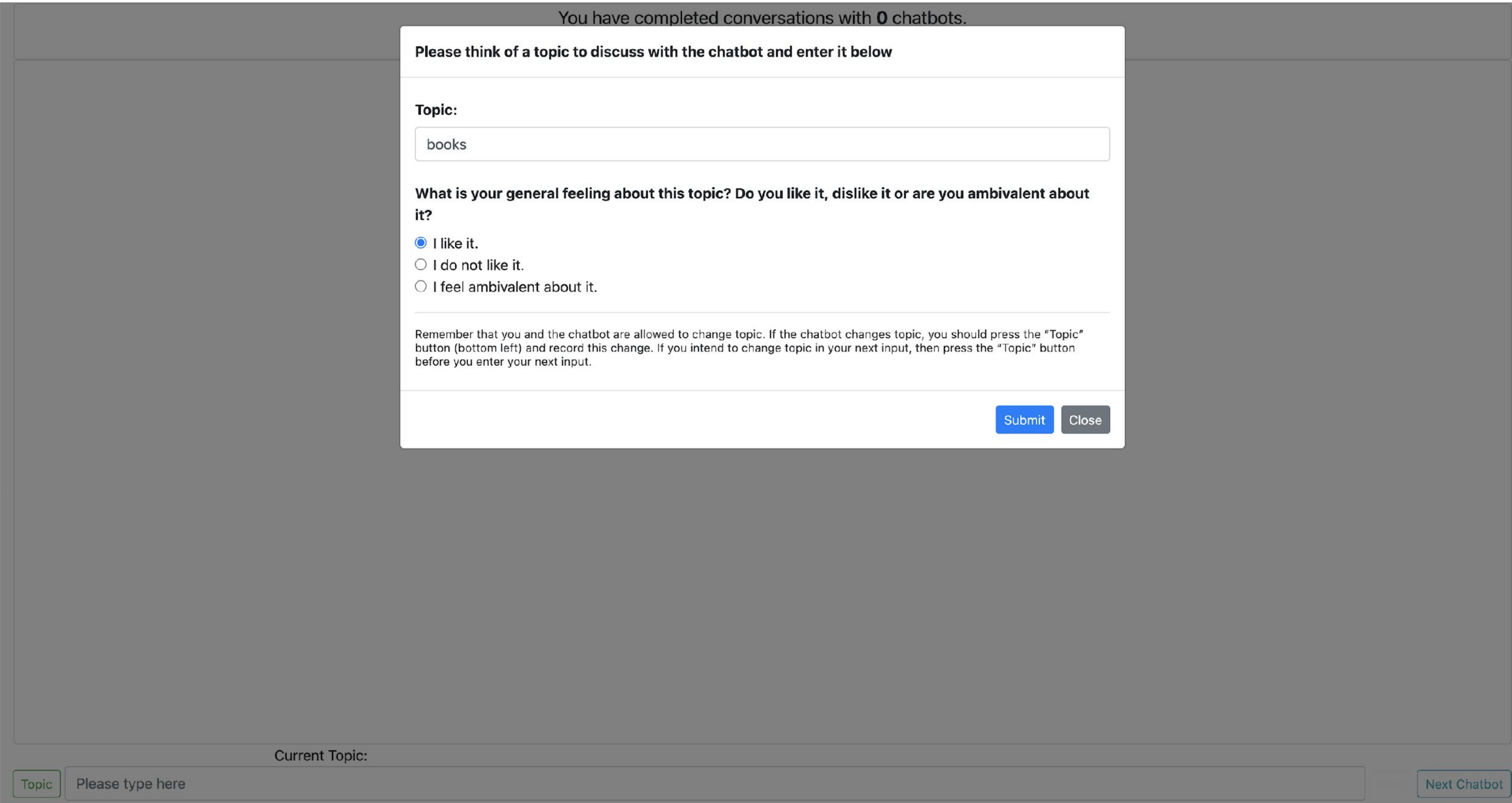


Free Run1 Free Run2 Ice-breaker

Figure 2. Results of pairwise significance test where a colored cell indicates that the system in that row significantly outperformed the system in that column.

User interface – beginning of a conversation

www.adaptcentre.ie



Experiment – general feelings of topics

	Free run 1		Free run 2	
	Pass	Fail	Pass	Fail
Like	83.9	88.6	86.4	93.8
Ambivalent	7.4	3.8	6.2	2.3
Dislike	8.7	7.7	7.4	3.9

Table 2: Proportions (%) of topics that are reported as liked, ambivalent about or disliked by workers who passed and failed quality control.