Exploring the Impact of Human Evaluator Group on Chat-Oriented Dialogue Evaluation

Sarah E. Finch, James D. Finch, and Jinho D. Choi
Department of Computer Science
Emory University
Atlanta, GA, USA
{sfillwo, jdfinch, jinho.choi}@emory.edu

Abstract

Human evaluation has been widely accepted as the standard for evaluating chat-oriented dialogue systems. However, there is a significant variation in previous work regarding who gets recruited as evaluators. Evaluator groups such as domain experts, university students, and crowdworkers have been used to assess and compare dialogue systems, although it is unclear to what extent the choice of an evaluator group can affect results. This paper analyzes the evaluator group impact on dialogue system evaluation by testing 4 state-of-the-art dialogue systems using 4 distinct evaluator groups. Our analysis reveals a robustness towards evaluator groups for Likert evaluations that is not seen for Pairwise, with only minor differences observed when changing evaluator groups. Furthermore, two notable limitations to this robustness are observed, which reveal discrepancies between evaluators with different levels of chatbot expertise and indicate that evaluator objectivity is beneficial for certain dialogue metrics.

Keywords: chatbot, dialogue, conversational ai, evaluation

1. Introduction

It is common for chat-oriented dialogue modeling to rely on human evaluation for comparing the performance of different dialogue systems, as automated metrics have been shown to be insufficient (Liu et al., 2016; Deriu et al., 2022). There is not a standard pool of human evaluators used across all works, though, as works tend to recruit their evaluators from many sources, with the most common being students (Zhou et al., 2021; Gu et al., 2022), crowd workers (Liu et al., 2021; Kim et al., 2021), annotation companies (Song et al., 2021; Cao et al., 2022), and experts (Varshney et al., 2022).

The impact of varying evaluator characteristics has been thoroughly studied for other applications, with many indicating that evaluator characteristics have a substantial impact on the provided judgments (Dundes and Rajapaksa, 2001; Vella and Mills, 2017), although not for every application (Young et al., 2009). The current degree to which dialogue evaluation metrics are consistent across people of varying backgrounds is unclear due to limited work exploring this effect, which is concerning from the perspective of comparing results across works using different evaluator groups.

Our motivating hypothesis is that dialogue evaluation is vulnerable to changing the background characteristics and the point-of-view of the evaluators themselves, leading to inconsistent judgements on how different dialogue systems compare to one another. This is an important question because if it is true that it is possible to substantially change the evaluation outcomes of dialogue systems by simply changing who is doing the evaluation, then the choice of evaluator group becomes a critical decision that needs to be carefully considered moving forward.

Towards this goal, this paper investigates the impact of different evaluator groups on multi-turn human evaluation of chat-oriented dialogues. The two most popular evaluation methods for chat-oriented dialogue are used: Likert ratings and Pairwise selections (Finch and Choi, 2020). For our experiments, 4 groups of evaluators are invited to provide evaluations on the same dialogue dataset: chatbot developers, professional crowdworkers, university students with interactive point-of-view (POV), and university students with external POV. Their evaluations are then compared via dialogue-level and bot-level measures.

Our work illustrates that certain evaluation methods are impacted by the choice of evaluator group, although the degree of impact varies. In particular, we contribute 3 main findings that can guide future dialogue evaluations:

1. Likert ratings are more stable than Pairwise selections across evaluator groups

2. Chatbot developers produce dissimilar evaluations to less-experienced groups

3. Objective dialogue metrics achieve better consistency among external evaluators

Our data, raw evaluation results, and analysis scripts can be accessed through our open-source project: `https://github.com/sfillwo/DialogueEval-AnnotatorImpact`.

LREC-COLING 2024, pages 6966–6973

2. Related Work

There have been a handful of works that compare the evaluation agreement between different evaluator groups. Ram et al. (2018) observed a high ($r = 0.93$) correlation between the 15 bot rankings of average Likert quality from Alexa Prize users and the average Likert engagingness and coherence from Amazon employees. Venkatesh et al. (2018) claimed that they observed low agreement between their internal Amazon employee raters and Alexa Prize users for Alexa Prize chatbot conversations without providing their numerical results. Kulikov et al. (2019) examined the differences in the average ratings provided by individual annotators for a greedy-search dialogue model and found that these averages can be dramatically different.

A few works have focused specifically on the difference between expert evaluators and non-experts. Finch and Choi (2020) demonstrate that experts and non-experts do not agree on Likert quality ratings. Similarly, Higashinaka et al. (2021) find that experts are more consistent on utterance error labeling than non-experts, although they do not compare the agreement between experts and non-experts.

3. Data

For this work, we use ABC-Eval from Finch et al. (2023), a dialogue dataset containing 400 30-turn human-bot dialogues from 4 chatbots: Blender2 (Weston and Shuster, 2021), Emora (Finch et al., 2020), Blender-Decode (Nie et al., 2021), and BART-FiD-RAG (Shuster et al., 2021). ABC-Eval provides evaluations using two of the most common dialogue evaluation procedures as identified in Finch and Choi (2020):

- **Likert rating**: human annotators provide a rating from 1 to 5 for how well the chatbot's responses fits a metric definition.
- **Pairwise selection**: human annotators are shown 2 dialogues and select the dialogue for which the chatbot's responses better fit the metric definition.

The 8 included metrics are shown in Table 1.

ABC-Eval provides such evaluations from two groups: Stu_i and Sur_x, covering two common evaluation groups of interactive students and crowdworkers, respectively, for dialogue evaluation. To cover additional common groups, we also collect evaluations from chatbot development experts (Dev_x) and external students (Stu_x). All groups are composed of native English speakers.

By design, the important conditions for dialogue evaluation are standardized between the evaluator groups in our experiments; namely, all evaluator groups are given the same instructions, annotation format, time-to-complete restrictions, and dialogue dataset to evaluate. The differences between the evaluator groups that are described in the remainder of this section are intentional, aiming to represent different ways that dialogue evaluation is performed. These differences include the background of the group (crowdworkers/students/developers) and the point-of-view of the group (external/interactive), which are realized through different recruitment methods and whether the evaluators in a group are both conversing and rating the subsequent conversation or viewing an existing dialogue and rating it. By constructing the evaluator groups in this way, we are able to compare whether there is a significant impact on the evaluation results between groups that vary on these different characteristics under study.

`Gra`	Responses are free of grammatical and semantic errors
`Rel`	Responses are on-topic with the immediate dialogue history
`Inf`	Responses produce unique and non-generic information that is specific to the dialogue context
`Emo`	Responses indicate an understanding of the user's current emotional state and provide an appropriate emotional reaction based on the current dialogue context
`Eng`	Responses are engaging to user and fulfill the conversational goals implied by the user
`Con`	Responses do not produce information that contradicts other information in the dialogue
`Pro`	Responses actively and appropriately move the conversation along different topics
`Qua`	The overall quality of and satisfaction with the dialogue

Table 1: The 8 dialogue evaluation metrics and their definitions; adapted from Finch and Choi (2020).

3.1. Evaluator Groups

All evaluator groups use the same underlying annotation platform to perform their annotations, which is a web-based platform based on ParlAI (`https://parl.ai/`), thus they are shown the same instructions and interface for annotation. Stu_i, Stu_x, and Dev_x are given URL links through email to access the web platform. Sur_x are given URL links to access the web platform through job posts that we uploaded to the Surge crowdsourcing platform (`https://www.surgehq.ai`). In the end, all evaluators are redirected to the same annotation platform to perform their evaluation tasks; it is just how they are given the URL links to the platform that differs. Next, the details of each evaluator group are discussed.

Interactive Students ($\mathbf{Stu_i}$) Undergraduate students are recruited through email advertisements. The hired students receive links to the online evaluation platform to complete their assigned sessions. For each session, they are paired with random dialogue systems and converse back-and-forth for 30 turns. At the conclusion of each conversation, they perform the Likert evaluation tasks on the 8 metrics *on the conversation that they just had*. At the conclusion of every two conversations, they perform the Pairwise evaluation tasks on the 8 metrics.

Contrary to interactive evaluators (Stu_i), external evaluators (Sur_x, Dev_x, and Stu_x) are provided with a static dialogue and asked to perform the evaluation task on it. This static dialogue is taken from the dialogues collected using the interactive evaluators. The important distinction is that external evaluators are *not involved* in the dialogues which they are evaluating, unlike interactive evaluators. External evaluation of dialogue systems has historically been more common, but interactive evaluation is growing in popularity recently. We choose to represent this dichotomy in our evaluator groups to explore whether they produce differences in dialogue evaluation, since they are used interchangeably in the field these days.

SurgeHQ Crowdworkers ($\mathbf{Sur_x}$) The annotation company SurgeHQ provides a pool of historically high-performing crowdworkers (hereafter, referred to as Surgers) for the ABC-Eval project. URL links to the evaluation tasks are posted as jobs on the SurgeHQ crowdworking platform. A single Likert task consists of providing Likert ratings on 8 metrics for one dialogue. A single Pairwise task consists of providing pairwise selections between two dialogues on the 8 metrics.

Chatbot Developers ($\mathbf{Dev_x}$) A group of chatbot developers involved in the development of a university chatbot are recruited from an ongoing project. None of the chatbot developers were involved in the development of any of the 4 chatbots being evaluated in this work. URL links are sent via email and their tasks follow the same format as Sur_x.

External Students ($\mathbf{Stu_x}$) A group of undergraduates are recruited from the same university as Stu_i. URL links are sent via email and their tasks follow the same format as Dev_x and Sur_x.

Table 2 provides statistics on the evaluations from each group. A subset of the dialogues are doubly annotated per group (except for Stu_i due to its interactive setup), meaning that 2 human annotators provided evaluations for those dialogues. It should be noted that the number of evaluators in Stu_x and Dev_x is much smaller than that of Stu_i and Sur_x. Our multiple attempts to recruit participants for Stu_x were met with little interest from the student population, whereas the recruitment for Stu_i in Finch et al. (2023) did not seem to suffer from such disinterest based on their success in obtaining such a large number of willing participants. One likely explanation for this is that the task of *conversing* with dialogue models is much more compelling to human participants compared to just *evaluating* human-bot dialogues. Similarly, the specialization criteria severely reduces the population from which the evaluators can be drawn from for Dev_x. Due to their smaller sizes, it was challenging to collect full evaluations on the dialogue dataset even over several months for both Stu_x and Dev_x.

Group	Stu_i	Stu_x	Sur_x	Dev_x
Evaluators	46	8	32	3
Likert **Payment**	400 (0) †	228 (37) \$0.50	400 (108) \$0.60	177 (25) \$0.50
Pairwise **Payment**	200 (0) \$1.67	193 (19) \$1.00	192 (54) \$1.43	72 (11) \$1.00

Table 2: Statistics on the number of evaluators, number of evaluated dialogues (# of doubly annotated conversations in parentheses), and compensation amount for each group. †: Due to the interactive setup, Stu_i received compensation covering Likert and Pairwise work together and only produced singly annotated dialogues.

4. Dialogue Score Agreement

One aspect of evaluation robustness is whether the same dialogue is given the same score by different evaluators. For this, between-group interannotator agreement (IAA) acts as a measure of the impact of changing evaluator group. Higher agreement between groups signals that their dialogue-level evaluation decisions are more similar.

Following Finch et al. (2023), we use Krippendorff's alpha (α) to measure IAA. The between-group IAA is measured for each evaluation metric by aligning the evaluation annotations between two evaluator groups such that annotations made for the same dialogue are paired. The within-group IAA is calculated from the doubly annotated dialogues per group. Table 3 shows the results.

Overall Low Agreement Across the metrics, the between-group agreements are rather low, rarely surpassing $\alpha = 0.5$, showing that the dialogue-level judgements are rarely matched between groups. In addition, there is not an obvious difference between the agreements observed for Likert and Pairwise evaluations; thus neither evaluation procedure

	Con	Emo	Eng	Gra	Inf	Pro	Qua	Rel
Dev$_x$ / Stu$_i$	0.27	0.15	0.30	**0.38**	0.05	0.20	0.34	0.34
Dev$_x$ / Stu$_x$	**0.51**	**0.40**	**0.42**	0.27	-0.30	0.23	**0.51**	**0.45**
Dev$_x$ / Sur$_x$	0.24	0.33	0.19	0.20	**0.16**	0.12	0.33	0.14
Stu$_i$ / Stu$_x$	0.30	0.12	0.19	0.08	-0.00	**0.32**	0.28	0.16
Stu$_i$ / Sur$_x$	0.20	0.24	0.24	0.05	0.12	0.21	0.23	0.28
Stu$_x$ / Sur$_x$	0.30	0.17	0.16	0.19	0.06	0.25	0.27	0.10
Dev$_x$	0.48	0.59	0.60	0.16	0.41	0.69	0.61	0.44
Stu$_x$	0.45	0.17	0.12	0.42	0.26	0.13	0.46	0.05
Sur$_x$	0.36	0.26	0.26	0.13	0.41	0.24	0.29	0.30
Dev$_x$ / Stu$_i$	**0.38**	0.54	0.37	0.28	0.16	0.21	0.35	0.36
Dev$_x$ / Stu$_x$	0.19	**0.57**	**0.69**	**0.32**	**0.40**	**0.66**	**0.40**	**0.40**
Dev$_x$ / Sur$_x$	0.16	0.11	0.23	0.13	0.25	0.27	0.22	0.14
Stu$_i$ / Stu$_x$	0.13	0.20	0.24	0.22	0.09	0.40	0.32	0.27
Stu$_i$ / Sur$_x$	0.01	0.13	0.10	0.12	0.06	0.13	0.16	0.11
Stu$_x$ / Sur$_x$	0.28	0.21	0.25	-0.04	0.26	0.25	0.38	0.17
Dev$_x$	0.82	0.64	1.00	0.90	0.32	0.64	1.00	0.64
Stu$_x$	0.27	0.01	0.07	0.38	0.07	0.48	-0.15	0.12
Sur$_x$	0.51	0.20	0.10	0.22	0.19	0.25	0.47	0.44

Table 3: α for Likert (top) and Pairwise (bottom). Gray row color indicates within-group α's. **Bold** indicates highest between-group α.

seems to be more robust to changing the evaluator group on the dialogue-level.

High Agreement Developers For both Likert and Pairwise, the highest agreement between any two groups is achieved when one of the groups is Dev_x. The developer group seems to have the most in common with the other groups in terms of the specific dialogue-level judgements.

5. Bot Performance Analysis

Although dialogue-level judgements are one outcome of dialogue evaluations, the ultimate goal of dialogue evaluation is to compare various bots to one another. Indeed, it has been seen that low dialogue-level agreement does not necessarily result in low agreement on relative bot performances when assessing evaluation strategies (Lee et al., 2020; Ji et al., 2022). As a result, looking at the dialogue-level judgements alone is not enough to understand the effect of changing evaluator groups and it is crucial to directly assess the relative bot performances between evaluator groups.

We first considered testing for differences in the aggregate scores for each bot produced by different evaluator groups. However, such results are not useful towards the goal of comparing the relative bot performances produced by different evaluator groups as they consider the evaluation of a bot in isolation. For example, suppose evaluator groups A and B score bot X at 3.0 and 4.0 respectively, and they score bot Y at 3.5 and 4.5 respectively. Given equal variance in their scoring distributions and sufficient sample size, testing with respect to each bot in isolation would detect that the scores produced by A and B are different when rating a particular bot, even though their evaluation of the 2 bots' relative performance was highly analogous.

Consequently, we need a more appropriate metric that measures agreement among evaluator groups on bot-pair comparisons. First, we look to quantify bot-pair comparisons produced by an evaluator group through their evaluations. This is achieved by calculating the effect size, which quantifies the magnitude of difference between two conditions (in this context, between two bots). Following this, for each bot-pair under consideration, the effect size for each evaluation metric is computed for every evaluator group. Second, a mechanism for comparing these effect sizes across different evaluator groups is required. This comparison is facilitated by examining the numerical difference in the effect sizes for identical bot-pairs and evaluation metrics between two evaluator groups. A diminishing numerical difference denotes a closer alignment in the evaluations conducted by the two groups. Finally, an approach is needed to consolidate these comparisons into a singular score that reflects the evaluation concordance among a pair of evaluator groups for each evaluation metric. This is accomplished by averaging the absolute values of the effect size differences observed between the evaluator groups across all bot-pairs. Thus, the formal calculation of the degree of agreement in evaluations between two evaluator groups is:

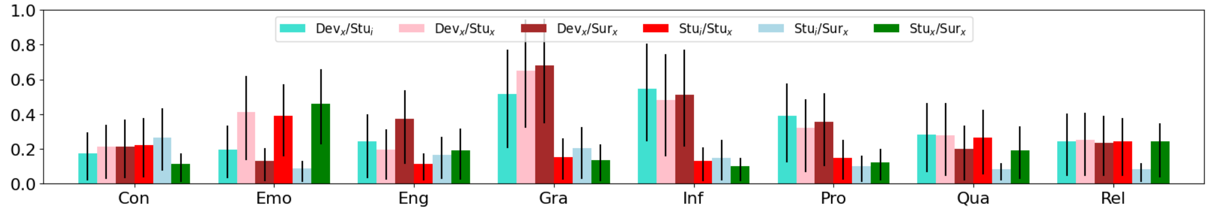
$$\frac{1}{N}\sum_{(b,b')\in B}|E(b_{g1},b'_{g1})-E(b_{g2},b'_{g2})|$$

where B is the set of bot-pairs, N is the size of B, and E is the effect size function (Cohen's d for Likert ratings and Cohen's h for Pairwise bot-vs-all win proportions). $b/b'_{g1/g2}$ refers to the evaluations for one of the bots (b, b') by one of the evaluator groups ($g1$, $g2$). $|\cdot|$ denotes absolute value.

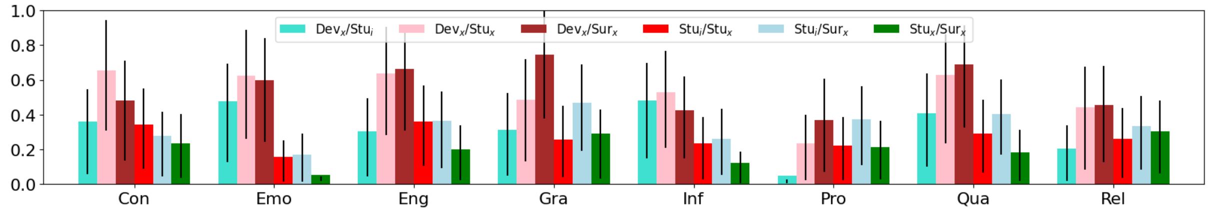
This calculation is done for each evaluation metric under study. Figure 1 shows the effect size differences observed between each evaluator group.

Developer Dissimilarity Dev_x often showcases the greatest difference relative to other evaluator groups, surpassing an effect-size difference of 0.4 for many metrics. This effect is more consistently observed for the Pairwise evaluations, although it is observed for several Likert metrics as well.

Even though Dev_x achieved some of the best agreement with other groups on the *dialogue*-level (Section 4), Dev_x frequently produced low similarity on the *bot*-level effect-sizes compared with those groups. Although this may seem contradictory, it is possible for both high dialogue-level agreement and low bot-level effect-size similarity to occur. One possible explanation is that the ratings provided by Dev_x in the cases of disagreement were more extreme than those produced by the other groups, resulting in larger observed effect-sizes for Dev_x evaluations and, thus, larger effect-size differences



(a) Likert



(b) Pairwise

Figure 1: Bot-pair effect size differences between groups (**smaller** is better). Error bars: 95% confidence intervals using bias-corrected and accelerated bootstrap confidence intervals with $k = 10,000$ Monte Carlo case resamples.

between Dev_x and other groups. In other words, if Dev_x produced more extreme ratings on the disagreed upon instances than the other groups, then Dev_x would identify larger differences between the bots overall. Indeed, this explanation is corroborated when we examine the bot-pair effect sizes observed for each group (Table 4) which shows that Dev_x consistently identifies higher bot-pair effect sizes than any other group. Based on the consistency of this effect, it is likely that an evaluator's level of expertise has more of an impact on bot evaluation than other evaluator characteristics.

	Dev$_x$	**Stu$_i$**	**Stu$_x$**	**Sur$_x$**
`Con`	0.31	0.13	0.32	**0.39**
`Emo`	**0.52**	0.35	0.34	0.44
`Eng`	**0.48**	0.40	0.38	0.38
`Gra`	**1.10**	0.58	0.44	0.43
`Inf`	**0.73**	0.19	0.25	0.27
`Pro`	**0.87**	0.48	0.63	0.53
`Qua`	**0.43**	0.36	0.39	0.36
`Rel`	**0.47**	0.30	0.30	0.31
`Con`	**0.63**	0.28	0.25	0.35
`Emo`	**0.65**	0.29	0.22	0.19
`Eng`	**0.72**	0.44	0.15	0.32
`Gra`	**0.73**	0.50	0.43	0.14
`Inf`	**0.77**	0.33	0.35	0.43
`Pro`	**0.63**	0.58	0.49	0.40
`Qua`	**0.69**	0.36	0.21	0.14
`Rel`	**0.56**	0.35	0.20	0.17

Table 4: Raw effect sizes between bot-pairs as produced by each evaluator group (averaged across bot-pairs). **Bold**: largest observed effect size.

Likert Cross-group Robustness Overall, there is a lower difference in effect sizes between groups for the Likert evaluations compared to Pairwise evaluations. This result indicates that Likert evaluations are more robust to group changes. Furthermore, the observed effect-size differences for Likert metrics are often small, rarely surpassing 0.2, between the evaluator groups (excluding Dev_x). As a result, the evaluation outcomes observed between these different evaluator groups are likely to manifest as only minor changes in practice.

Objectivity Favors External Point-of-View The smallest effect-size difference for Con is seen between Stu_x/Sur_x, for both Likert and Pairwise evaluations. This is also observed for Inf, although the impact is more profound for Pairwise than for Likert. It is likely that these metrics with their more objective foci benefit from having external evaluators who are divorced from the conversation itself, thus reducing bias in their judgements from their subjective experience or emotional reaction to the conversation.

6. Conclusion

The analyses presented in this work provide insight into the effects that switching evaluator groups can have on the results of dialogue model evaluations. The results support a recommendation for utilizing Likert ratings due to their higher stability across evaluator groups. Furthermore, if the ultimate goal is to understand how laypeople would evaluate a particular bot, we would discourage the use of chatbot developers for evaluation based on the observed dissimilarities and encourage the use of external evaluators for those metrics that are more objective.

7. Limitations

This study was limited to 4 chatbots and 4 evaluator groups, 2 of which did not provide complete evaluations for the 400 total dialogues (Section 3). In addition, the Dev_x evaluator group was only composed of 3 annotators, which is quite small. We acknowledge that in an ideal situation we would have been able to recruit more developers and ensure that each evaluator group annotated the same number of dialogues. However, even with just the data we collected, the discoveries made in this study are useful for highlighting the need for further work in this area and encouraging further consideration of best practices for choosing evaluator groups for future dialogue works, especially since this is the first study of its kind to do this degree of analysis for dialogue evaluation across multiple evaluation metrics and several evaluator groups. Further work on additional group characteristics and dialogue models will aid in gaining a more complete picture of the impact of different evaluator groups on dialogue evaluation results.

It should also be mentioned that there is a small amount of work on optimizing the within-group agreement for evaluations of dialogue models. For instance, Li et al. (2019) present a small-scale study of the impact of different phrasings for dialogue metrics, noticing that it is possible to increase IAA through surface-level modifications of the evaluation questions. The evaluation setup that we follow from Finch et al. (2023) does not perform any such optimization. Thus, it is possible that even the cross-evaluator-group agreements could be improved beyond that observed in this work through comprehensive investigation of such optimization.

8. Ethics Statement

This work presents findings that are useful for improving the quality of human evaluation of chat-oriented dialogue models. In particular, we explore the effect that changing evaluator group has on dialogue evaluation results. Our analyses will allow researchers to make more informed decisions and interpretations of dialogue model evaluations.

This work relies on the evaluation of real human-bot conversations; however, we do not collect new dialogue data and instead leverage an anonymized, publicly available dialogue dataset from the Finch et al. (2023) study. Finch et al. (2023) screened this data for any personally identifiable information in their work, alleviating any concerns regarding privacy of the original human participants.

When evaluating human-bot conversations, there is also a small risk of conversations containing triggering or offensive content to the evaluator from either the utterances of the human participants or of the dialogue model. Finch et al. (2023) explicitly instructed the human participants to refrain from such antisocial behavior and also reported dramatically low rates of antisocial bot utterances (near 0% for all conversations). As a result, the utilized dataset does not carry this risk.

The human evaluators used in this work were USA-based and paid at a rate above the federal USA minimum wage of $10/hour for Stu_x and Dev_x. The per-task payment rate was based on the results of pilot studies which afforded time estimates for the completion of the different evaluation tasks. Participation by human evaluators was entirely voluntary. Evaluators were able to opt-out of the study at any time. Evaluators were given a clear description of their responsibilities and the compensation structure before being given any work. These measures ensure that the participation of our human evaluators was fair, transparent, and benefited their interests.

9. Acknowledgements

We gratefully acknowledge the support of the Amazon Alexa AI grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Amazon.

10. Bibliographical References

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002, Dublin, Ireland. Association for Computational Linguistics.

Jan Deriu, Don Tuggener, Pius Von Däniken, and Mark Cieliebak. 2022. Probing the Robustness of Trained Metrics for Conversational Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 750–761, Dublin, Ireland. Association for Computational Linguistics.

Lauren Dundes and Sushama Rajapaksa. 2001. Scores and grades: a sampling of how college students and food safety professionals interpret restaurant inspection results. *Journal of Environmental Health*, 64(5).

Sarah E. Finch and Jinho D. Choi. 2020. Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Sarah E. Finch, James D. Finch, Ali Ahmadvand, Ingyu, Choi, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, and Jinho D. Choi. 2020. Emora: An Inquisitive Social Chatbot Who Cares For You.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don't forget your abc's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022. HeterMPC: A heterogeneous graph neural network for response generation in multi-party conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5086–5097, Dublin, Ireland. Association for Computational Linguistics.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online. Association for Computational Linguistics.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving Reliable Human Assessment of Open-Domain Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Seolhwa Lee, Heuiseok Lim, and João Sedoc. 2020. An Evaluation Protocol for Generative Conversational Systems. ArXiv:2010.12741 [cs].

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: The Science Behind the Alexa Prize.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

Deeksha Varshney, Akshara Prabhakar, and Asif Ekbal. 2022. Commonsense and named entity aware knowledge grounded dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1335, Seattle, United States. Association for Computational Linguistics.

Elizabeth J Vella and Gregory Mills. 2017. Personality, uses of music, and music preference: The influence of openness to experience and extraversion. *Psychology of Music*, 45(3):338–354.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On Evaluating and Comparing Open Domain Dialog Systems. In *Proceedings of Conversational AI workshop at NIPS 2017*. arXiv. ArXiv:1801.03625 [cs].

Jason Weston and Kurt Shuster. 2021. Blender Bot 2.0: An open source chatbot that builds long-term memory and searches the internet.

Suzanne Young, Leslie Rush, and Dale Shaw. 2009. Evaluating gender bias in ratings of university instructors' teaching effectiveness. *International Journal for the Scholarship of Teaching and Learning*, 3(2):n2.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2021. Learning from perturbations: Diverse and informative dialogue generation with inverse adversarial training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 694–703, Online. Association for Computational Linguistics.