Human Evaluation of Conversations is an Open Problem: *comparing the sensitivity of various methods for evaluating dialogue agents*

Eric Michael Smith[1] **Orion Hsu**[2] **Rebecca Qian**[1]
Stephen Roller[1] **Y-Lan Boureau**[1] **Jason Weston**[1]

[1]Facebook AI Research [2]Duke University

Abstract

At the heart of improving conversational AI is the open problem of how to evaluate conversations. Issues with automatic metrics are well known (Liu et al., 2016), with human evaluations still considered the gold standard. Unfortunately, how to perform human evaluations is also an open problem: differing data collection methods have varying levels of human agreement and statistical sensitivity, resulting in differing amounts of human annotation hours and labor costs. In this work we compare five different crowdworker-based human evaluation methods and find that different methods are best depending on the types of models compared, with no clear winner across the board. While this highlights the open problems in the area, our analysis leads to advice of when to use which one, and possible future directions.

1 Introduction

Any comprehensive analysis of the performance of an open-domain conversational model must include human evaluations: automatic metrics can capture certain aspects of model performance but are no replacement for having human raters judge how adept models are at realistic and interesting conversation (Deriu et al., 2021; Liu et al., 2016; Dinan et al., 2019b). Unfortunately, human evaluations themselves must be carefully constructed in order to capture all the aspects desired of a good conversationalist. Any evaluation technique must evaluate over many turns of a conversation in order to detect emergent faults such as repetitiveness or contradiction, while techniques that rely solely on a single evaluation at the end of a conversation may fail to take into account changes in model performance over its span. Further, techniques that rate model performance on a Likert scale may suffer from inconsistencies in subjective numerical ratings across evaluations of different models (Li et al., 2019). When comparing various human evaluation methods to assess which works best, we find that each has success and failure cases, leading us to conclude that human evaluation is still an open problem.

In this work, we analyze a representative set of human evaluation techniques. First, we compare *per-turn* evaluations, where ratings are given after every model response, and *per-dialogue* evaluations, where ratings are collected solely at the end of the conversation. Per-turn evaluations have the advantage of being more fine-grained, encouraging annotators to focus on small differences; however, the quality of a conversation is more than the sum of its parts, and global per-dialogue evaluations can capture this better. Second, we consider *pairwise methods*, where two models are compared directly by an annotator, to *single-model* methods, where the annotator sees and rates only one model at a time. Both approaches can be either per-turn or per-dialogue. For example, in Pairwise Per-Turn evaluation, a crowdworker chats with a dialogue agent, and after each of the worker's messages, they must choose between two possible responses from the agent, one from each of two different models. The pairwise approach can spot subtle differences apparent when comparing responses, and it can mitigate problems with distribution shift that occur in absolute scoring. Single-model approaches, however, can work well when direct comparison is not paramount.

We compare all of these different techniques for evaluating dialogue models in three different settings, and we contrast their individual strengths. We find that:

- Pairwise per-turn evaluations are adept at measuring changes in model performance throughout a conversation. This technique tends to work well when pairs of models clearly differ in how appropriate their responses are in the context of the previous lines of dialogue, for example, when comparing two models that are trained on different datasets.

- Pairwise per-dialogue evaluations tend to perform best when differences between models only emerge after several conversation turns, such as when these differences are very subtle, or when noticing patterns in responses that emerge globally across the entire conversation, for example the average length of responses.

- Single-model evaluations, performed both per conversation turn and at the end of a conversation, tend to not do as well in the two previously described settings, but do perform well when comparing models that differ only slightly in quality but are otherwise similar, for example two models with different numbers of parameters.

These findings, while highlighting the difficulty of human evaluation, also provide guidance on which method might be best to use in these different circumstances, as well as possible future work. In particular, investigating the best way to merge pairwise and single-model, per-turn and per-dialogue benefits into a single method could be a fruitful direction. We also analyze the interpretability of these approaches when collecting human written explanations. We have released code for these evaluation techniques in the ParlAI framework.[1]

2 Existing work

This work concerns itself with evaluation of open-domain dialogue, which, unlike more restricted domains such as question-answering and goal-oriented conversations, may not have a precise goal, and no widely accepted evaluation technique for it currently exists (Deriu et al., 2021; Huang et al., 2020; Roller et al., 2020). Automatic metrics are relatively fast, efficient, and reproducible, but many of them have been shown to "correlate very weakly with human judgement" (Liu et al., 2016, see also Dinan et al. (2019b)), and the best way to create a reliable automatic metric is still up for debate (Deriu et al., 2021). In this work we focus on human evaluation, and in particular on employing crowdworkers, which has an advantage over utilizing trained experts (Deriu et al., 2021) or deployment (Gabriel et al., 2020; Shuster et al., 2020) of allowing for a larger pool of evaluators and for ensuring alignment with research goals, respectively. However, the use of crowdworkers itself has a number of pitfalls to avoid as well (Huynh et al., 2021).

[1] https://parl.ai/projects/humaneval

Particular instruction wording choices to crowdworkers have a large effect on the quality of conversations and resulting evaluations (Huynh et al., 2021). Wording can direct workers to evaluate specific facets, such as general "get to know each other" chitchat (Zhang et al., 2018), getting a bot to generate unsafe utterances (Xu et al., 2020), and instructing crowdworkers to be adversarial vs. not (Dinan et al., 2019a). One can also pick from a variety of specific questions when asking crowdworkers to rate conversations, including asking about interestingness, making sense, fluency (See et al., 2019), sensibleness, specificity (Adiwardana et al., 2020), toxicity, and bias (Xu et al., 2020), and the exact phrasing of these questions can have a large impact on sensitivity (Li et al., 2019). Standard evaluation protocols have a single human both converse with a model and rate that conversation in the same task, but other methods have a rater rate pre-existing conversations between a human and a model or between a pair of models (Li et al., 2019; Deriu et al., 2020). These latter techniques allow for efficient reuse of existing conversational data, and have shown to be useful experimentally (Li et al., 2019; Roller et al., 2021), but it may be harder for evaluators to rate conversations that they have not been involved in.

Another choice when designing evaluation protocols is whether conversations are rated individually, e.g., with Likert-score ratings (Ashwin et al., 2017; Venkatesh et al., 2018, see more in Appendix A), or pairwise by comparing models (Li et al., 2019; Liang et al., 2020, etc.). Likert scoring suffers from weaknesses such as potential per-annotator bias (Kulikov et al., 2019) and drift in the distribution of errors over time (See et al., 2019), but is more efficient than pairwise comparisons in that new models' ratings can be compared to those of older models without having to re-collect those older models' ratings.

Lastly, evaluation techniques differ in whether they collect ratings on each turn of the conversation (Adiwardana et al., 2020; Komeili et al., 2021) or only at the end of the conversation, as in Acute-Eval (Li et al., 2019). Whole-conversation techniques can work well if the quality of a conversation is assumed to be more than just the sum of its parts, but could perhaps suffer due to the *primacy effect* and *recency effect* that appear when more weight is given to information presented at the start and end of the rating session, respectively (Asch, 1946; Anderson, 1965; Murdock Jr, 1962; Postman and Phillips, 1965).

See Appendix A for a more thorough assessment of related works.

3 Methods

3.1 Evaluation techniques

We investigate several human evaluation techniques, spanning a cross-section of the different methods discussed in existing work. Specifically:

- Single-model per-turn evaluations
- Single-model per-dialogue evaluations
- Pairwise per-turn evaluations
- Pairwise per-dialogue evaluations
- Pairwise per-dialogue *self-chat* evaluations

We thus compare the spectrum of single vs. pairwise and per-turn vs. per-dialogue variations, as well as trying a self-chat method compared to conventional human-bot conversation ratings. Figure 1 summarizes the methods. In the following, we will describe our exact methodology for each. See Appendix B.3 for details on quality checks used when performing these evaluations.

3.1.1 Conversational setting

Our human-bot evaluations consist of a set of conversations. Each conversation consists of a human worker crowdsourced from Amazon Mechanical Turk[2] (the "Human Speaker") paired with a conversational model (the "Bot Speaker"). The Human Speaker will speak naturally in the conversation, and they will be role-playing as a certain persona with the help of two provided *persona sentences* given to them at the start of the conversation: see Figure 2 (left) for an example.

The Human Speaker's first message in the conversation is fixed to *"Hi!"*, following the convention of Adiwardana et al. (2020). The conversation ends after the Human Speaker and Bot Speaker have both spoken for 6 turns each, to roughly match the conversation lengths used for BlenderBot evaluations in Roller et al. (2021). We test three different evaluation metrics, *preference*, *humanness* and *interestingness*, with exact wordings described in the following subsections.

[2]Our crowdsourcing task pays workers well above minimum wage, and the task does not request any personal information from workers.

3.1.2 Pairwise per-turn evaluations

The Pairwise Per-Turn evaluation (PW-Turn) technique provides annotations for every turn of conversation by asking for the crowdworker to choose from a pair of model responses after every sent message. Hence, in this setting the Human Speaker speaks to a Bot Speaker, the latter of which actually represents the two models to be compared. The Human Speaker will speak naturally in the conversation. Every time that it is the Bot Speaker's turn to speak, the crowdworker will first be presented with two options as possible responses: each response will come from one of the two models being compared, similarly to Clark and Smith (2021). We randomize the ordering of these model responses. The worker must choose the better response for the given evaluation metric. The wordings we use for the three metrics are adapted from Li et al. (2019):

- **Preference**: *"Which next response from your partner would you prefer in a long conversation?"*
- **Humanness**: *"Which next response from your partner sounds more human?"*
- **Interestingness**: *"If you had to say one of these responses is interesting and one is boring, which would you say is more interesting?"*

The worker must give a free-text justification for their choice of response. The response that they choose is set to be the actual response given by the Bot Speaker, and the conversation continues from there. Figure 2 provides a screenshot example of the UI. A description of quality checks performed when onboarding workers for this evaluation technique is given in Appendix B.2. In our experiments we consider win rates based on simply averaging over turns, as well as nonlinear combinations of per-turn results over entire dialogues (e.g., winner-takes-all voting) in order to measure their impact.

3.1.3 Pairwise per-dialogue evaluations

The Pairwise Per-Dialogue evaluation (PW-Dialog) technique we introduce asks evaluators to choose between two models by presenting a pair of conversations. The technique we employ is identical to the Acute-Eval method (Li et al., 2019), but for consistency with the names of other techniques, we refer to it here as PW-Dialog evaluations. For each of the model pairs and evaluation metrics used, we

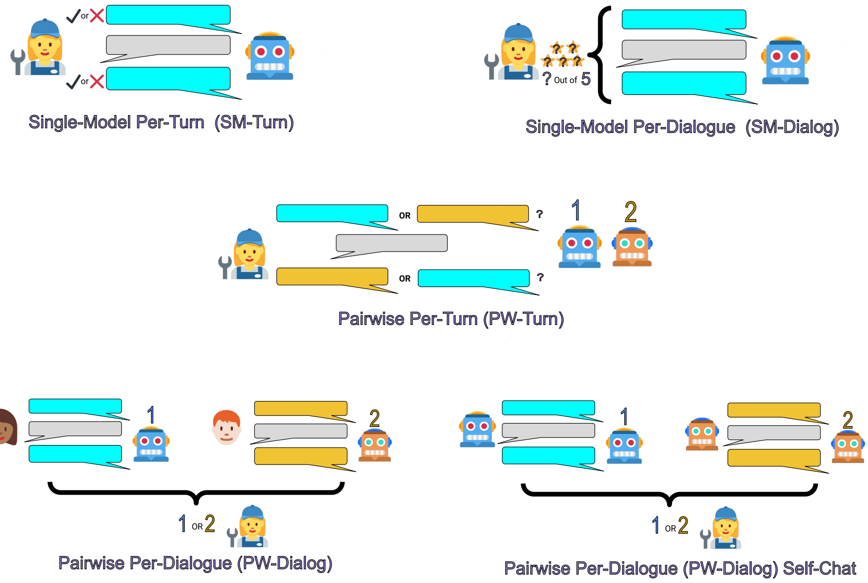

Figure 1: The human evaluation methods we compare in this work. SM-Turn rates each bot response during the conversation, while SM-Dialog rates the entire conversation. PW-Turn compares two different bots' responses at every turn in the conversation, while PW-Dialog compares two entire conversations with two different bots. PW-Dialog *self-chat* compares two conversations which only involve the two bots talking to themselves (self-chat).

collect evaluations on (1) conversations conducted between a crowdworker and a model agent; and (2) self-chat conversations conducted between two conversational agents of the same model (the *self-chat* variant). The wordings we use (from Li et al. (2019)) are almost identical to the PW-Turn versions, but phrased for the per-dialogue, rather than per-turn, case:

- **Preference**: *"Who would you prefer to talk to for a long conversation?"*
- **Humanness**: *"Which speaker sounds more human?"*
- **Interestingness**: *"If you had to say one of these speakers is interesting and one is boring, who would you say is more interesting?"*

Figure 3 provides a screenshot example of the UI.

3.1.4 Single-model evaluations

In our single-model evaluation experiments, we combine per-turn and per-dialogue into the same UI (see Figure 4 for a screenshot).[3] A crowdworker chats with a conversational agent backed by a single model, and for each response of that model the worker must annotate whether it is engaging, human-like, and/or interesting, with wording provided in the screenshot. At the end of the conversation, again consisting of 6 messages per speaker, the worker must rate their partner on a Likert scale of 1 to 5 for each of the three evaluation metrics listed in Section 3.1.2. We refer to the per-turn annotations of model responses from this task as Single-Model Per-Turn evaluations (SM-Turn) and the end-of-conversation Likert scores as Single-Model Per-Dialogue evaluations (SM-Dialog).

Empirically, we find that SM-Turn success rates and SM-Dialog Likert scores are highly dependent on the particular day that the evaluations are collected: this is perhaps due to day-to-day variability in the pool of crowdworkers. To counteract this, we run these evaluations on all four of the models discussed in this work (Section 3.2) simultaneously.[4]

3.2 Models

We analyze the relative performance of these five human evaluation techniques, SM-Turn, SM-Dialog, PW-Turn, PW-Dialog, and PW-Dialog *self-chat*, on four different well-performing but relatively similar dialogue models from Roller et al. (2021):

[3] This may have undesirable effects in correlating their results, but nonetheless they do appear to perform quite differently in evaluations.

[4] For the pairwise evaluation techniques PW-Turn and PW-Dialog, we collect evaluations over several days across multiple weeks for each of the three model pairs evaluated. This helps to smooth out variability among days.

- **BlenderBot3B**: The version of BlenderBot with 2.7 billion parameters, pretrained on a previously existing Reddit dataset (extracted and obtained by a third party and made available on pushshift.io (Baumgartner et al., 2020)) and then fine-tuned on several purpose-built dialogue datasets.

- **BlenderBot3B-M0**: BlenderBot3B uses a minimum generation length of 20 tokens to ensure relatively long, interesting responses. We also compare to exactly the same model but without a minimum generation length, referring to it with **-M0** postfix.

- **BlenderBot90M**: The variant of BlenderBot with 90 million parameters, trained on the same datasets as BlenderBot3B.

- **Reddit3B**: BlenderBot3B, but only pretrained on the third-party Reddit dump and not fine-tuned on dialogue datasets.

For all models, we use the same generation settings as in Roller et al. (2021), apart from the **-M0** adaptation. We choose these relatively similar models in our experiments as a difficult challenge for evaluation techniques to tell which one is best, but we *a priori* expect from previous Acute-Eval (PW-Dialog) *self-chat* measurements in Roller et al. (2021) that BlenderBot3B may perform as well as or better than the other three models.

For the three pairwise evaluation techniques, we specifically perform comparisons between three pairs of models, each of which differ in a characteristic way:

- **Length comparison**: Comparing BlenderBot3B to BlenderBot3B-M0: these models differ only in the length of their generations.

- **Size comparison**: Comparing two models with different numbers of parameters, BlenderBot3B and BlenderBot90M.

- **Fine-tuning comparison**: Comparing the fine-tuned BlenderBot3B to the pretrained-only Reddit3B (both with the same number of parameters).

4 Results

4.1 Model win rates from pairwise per-turn evaluations

We compute the win rates of BlenderBot3B over other models in Table 1 for the pairwise evaluation technique PW-Turn. We expect BlenderBot3B to be better, hence values closer to 100% are deemed more preferable. We display the win rates of four different variants: including all 6 conversation turns from the Bot Speaker, excluding the Bot Speaker's first turn from the evaluations, and computing a nonlinear function of the turns: either calculating squared or winner-takes-all win rates for each conversation and then averaging those scores across all conversations. We generally find that PW-Turn win rates are higher when dropping the first turn of the Bot Speaker, as discussed further in Appendix C.2. Win rates are typically even higher by aggregating over conversations in a winner-takes-all fashion, which has the effect of reducing the turn-by-turn variability of which model's response is chosen by the crowdworker.

		All turns	Turns 2 to 6		
Comp.	Metric	Lin	Lin	Sqr	WTA
Length	Pref	63%	67%	72%	**74%**
	Human	63%	68%	75%	**79%**
	Inter	68%	70%	77%	**84%**
Size	Pref	48%	52%	**53%**	49%
	Human	51%	56%	**58%**	54%
	Inter	49%	52%	54%	**55%**
FT	Pref	80%	82%	88%	**93%**
	Human	81%	84%	88%	**93%**
	Inter	71%	75%	80%	**85%**

Table 1: PW-Turn win rates of BlenderBot3B vs. BlenderBot3B-M0 ("Length"), vs. BlenderBot90M ("Size"), and vs. the base pretrained model, Reddit3B ("FT"), across three different evaluation metrics, Preference, Humanness, and Interestingness. Win rates are computed both across all turns and across only the last 5 turns from the Bot Speaker ("Turns 2 to 6"). **Lin**: the linear win rate $x/(x+y)$ of BlenderBot3B, given x wins of BlenderBot3B and y wins of the comparison model. **Sqr**: the "squared" win rate $x^2/(x^2+y^2)$, calculated per-conversation and then averaged across all conversations. **WTA**: the winner-takes-all win rate, defined as the percentage of all conversations for which BlenderBot3B wins on more turns, or equivalently $x^\infty/(x^\infty+y^\infty)$ as calculated per-conversation. Winner-takes-all scores are generally highest (highest values bolded).

We find that, in general, win rates of BlenderBot3B do not vary much as a function of the evaluation question used when asking workers to choose one model response over the other. It is unclear *a priori* whether this results from an ambiguity in the precise definitions of these questions/metrics when interpreted by the workers, correlations in how well models perform on some metrics vs. others, or some other reason.

4.2 Model scores from single-model evaluations

		SM-Turn			SM-Dialog
		All	Turns 3 to 6		
Met.	Model	Lin	Lin	WTA	
Preference	BB3B	70%	**71%**	**73%**	**4.19**
	BB3B-M0	**71%**	70%	70%	4.02
	BB90M	65%	64%	65%	3.97
	Reddit3B	55%	50%	50%	3.30
Human	BB3B	**70%**	**72%**	**73%**	**4.49**
	BB3B-M0	67%	66%	70%	4.22
	BB90M	65%	66%	70%	3.94
	Reddit3B	56%	54%	53%	3.50
Interesting	BB3B	**44%**	**45%**	**47%**	**4.22**
	BB3B-M0	35%	35%	36%	3.76
	BB90M	39%	40%	42%	3.83
	Reddit3B	39%	39%	37%	3.30

Table 2: Performance of BlenderBot3B (BB3B), BlenderBot3B-M0 (BB3B-M0), BlenderBot90M (BB90M), and Reddit3B on SM-Turn and SM-Dialog evaluations. SM-Turn mean success rates are calculated across all turns ("All") or across only the last 4 turns from the Bot Speaker ("Turns 3 to 6"). Scores represent the overall fraction of model responses marked as successful on the given evaluation metric ("Lin") or the number of conversations for which at least half of the model responses are marked as successful (winner-takes-all, "WTA"). SM-Dialog evaluations are Likert scores (with standard deviations in the range of 0.8 to 1.3). Highest scores across models are bolded.

Table 2 provides the per-turn success rates (SM-Turn) and end-of-conversation Likert scores (SM-Dialog) over all models. As with the pairwise evaluations of Section 4.1 and Roller et al. (2021), BlenderBot3B generally outperforms the other models using the SM-Turn and SM-Dialog methods as well. Table 5 (in the Appendix) shows success rates from the SM-Turn technique as a function of conversation turn (rather than aggregated). BlenderBot3B scores are generally stable across conversation turn but are slightly lower on the first two turns of the Bot Speaker, echoing similar findings with PW-Turn in Appendix C.2. We thus also consider removing SM-Turn scores from the first two turns in order to maximize the performance of BlenderBot3B relative to the other models. As with PW-Turn, we find that calculating the winner-takes-all score per conversation allows for an even bigger separation in performance between BlenderBot3B and the other models.

Unlike PW-Turn for which win rates are similar across all three evaluation metrics (Section 4.1), single-model success rates on the Interestingness metric are generally lower than those on the other two, especially for SM-Turn. We hypothesize that the juxtaposition of all three evaluation questions side-by-side in the UI of the SM-Turn and SM-Dialog crowdworker task (Figure 4) may aid workers in distinguishing among these three metrics and rating models differently on them.

See Appendix C.4 for an exploration of which turns of the conversation contribute most strongly to workers' final Likert-scale ratings.

4.3 Direct comparison of all evaluation techniques

In this section we directly compare all the pairwise and single-model evaluation techniques to each other to discern their relative strengths. See Appendix C.1 for details on the number of evaluations performed and number of crowdworker hours spent per technique.

4.3.1 Computing win rates across all techniques

In order to directly compare the performance of SM-Turn and SM-Dialog with that of the pairwise techniques, we calculate effective win rates for the two single-model techniques by bootstrapping samples of ratings from different models and then calculating how often SM-Turn success rates and SM-Dialog Likert scores from one model are higher than those of another. Following the analysis of best performing methods from Sections 4.1 and 4.2, we consider only Bot Speaker turns 2 through 6 for PW-Turn and turns 3 through 6 for SM-Turn in winner-takes-all (WTA) mode, in order to maximize the ability of these techniques to distinguish different models' performances.

Table 3 compares the win rates produced by all evaluation techniques. Overall, we find that a different technique performs best for each of the three model comparison types:

Length comparison The pairwise evaluation techniques PW-Dialog and PW-Turn perform much better than the single-model ones. BlenderBot3B responses tend to contain many more words on average than those of BlenderBot3B-M0, and so we hypothesize that this difference in sensitivity among the techniques may be due to the fact that

Comparison	Metric	PW-Turn Turns 2–6, WTA	PW-Dialog Human	 Self	PW combo	SM-Turn Turns 3–6, WTA	SM-Dialog
Length	Pref	74%	77%	**82%**	80%	55%	58%
	Human	79%	77%	**83%**	81%	52%	59%
	Inter	84%	**85%**	73%	73%	60%	65%
Size	Pref	49%	56%	55%	54%	59%	**60%**
	Human	54%	61%	55%	55%	52%	**66%**
	Inter	55%	59%	57%	56%	55%	**64%**
Fine-tuning	Pref	**93%**	70%	66%	69%	64%	71%
	Human	**93%**	54%	61%	65%	62%	73%
	Inter	**85%**	59%	64%	66%	60%	70%

Table 3: Win rates of BlenderBot3B vs. other models, for all evaluation techniques. For the per-turn techniques PW-Turn and SM-Turn, only the specified Bot Speaker turns are used to compute winner-takes-all scores, as in Tables 1 and 2. We show PW-Dialog win rates as measured on conversations between a crowdworker and a model ("Human") as well as from model self-chats ("Self"). "PW combo" represents the win rate when sampling ratings from PW-Turn (turns 2–6) and PW-Dialog (on model self-chats) at a ratio of 1:5. PW-Turn, PW-Dialog, and SM-Dialog are each found to be most sensitive at measuring model performance for one of the three model comparisons tested (highest win rates bolded). See Appendix C.1 for the number of evaluations and the estimated total number of worker-hours per technique.

viewing responses from both models side-by-side makes the length differences between them much more evident, especially when comparing two entire conversations as in PW-Dialog. Thus, if crowdworkers tend to prefer longer responses on average, the side-by-side comparison of model responses might aid in their ability to choose BlenderBot3B responses over those of BlenderBot3B-M0.

Size comparison The differences among the techniques here are smaller than for the Length comparison, with the full-dialogue techniques PW-Dialog and SM-Dialog slightly outperforming the per-turn ones. As shown by Roller et al. (2021), BlenderBot3B and BlenderBot90M do not perform statistically significantly differently on Acute-Evals (i.e. PW-Dialog) on self-chat conversations. Thus, it may make sense that any small differences in performance between these models are more evident on the level of whole conversations.

Fine-tuning comparison In this comparison, PW-Turn performs best out of all techniques. Because the Reddit3B model was not fine-tuned on conversational dialogue datasets, its responses to its partner generally make less sense in context than those of BlenderBot3B. We hypothesize that these more nonsensical responses may be very obvious to workers who are in the middle of having a conversation with the Bot Speaker during the PW-Turn evaluation. However, these responses may be less obvious to workers reading whole conversations in the PW-Dialog evaluation who have not interacted with the models directly, as well as to workers in SM-Turn and SM-Dialog evaluations who cannot directly compare Reddit3B responses to those of a model that has been fine-tuned on dialogue.

Explainability in experiments: analysis of crowdworker reasons During the crowdworker evaluation tasks, we also ask for reasons for the crowdworker's judgments. These reasons can give interpretability to the results. A full analysis is given in Appendix C.5. Overall, we find justifications that make sense in each of the three model comparisons, e.g. in the Length comparison we see keywords like *"information"* and *"detailed"* appearing often. For the Fine-tuning comparison, we often find keywords like *"flows"*, *"personal"* and *"contradicts"*, which implies that the fine-tuning conversational datasets like Persona-Chat provide for more personal, less contradictory, and flowing conversations.

Repeatability of experiments We provide an analysis in Appendix C.6 of the variability of model win rates over time for each of the evaluation techniques. Overall, we find that PW-Turn, PW-Dialog, and SM-Turn vary least across chunked experiments, with SM-Dialog having more variability. This makes the use of SM-Dialog less compelling.

4.3.2 Overall findings

The results of these three model comparisons hint that perhaps a per-turn evaluation technique may be more suitable for pairs of models that differ in their ability to reply sensibly in a way that is easily detectable by their partner (i.e. BlenderBot3B vs. Reddit3B), but that a whole-conversation technique may be preferable when differences between models are more sensitive. However, evaluations on many more pairs of models would be needed to sufficiently support such a broad hypothesis. We also find that single-model techniques perform competitively to pairwise ones except for when model generations differ by average length: in this case, comparing the responses of both models side-by-side may make the differences between them more apparent than just viewing them separately.

Combining techniques Given how much the relative sensitivities of different evaluation techniques vary across different pairs of models, we also explore whether combining results from multiple techniques together may allow for a compromise technique that performs reasonably well in all cases. We thus include in Table 3 the win rate ("PW combo") when sampling ratings from the PW-Turn and PW-Dialog techniques together at a ratio of 1:5. This sampling retains most of the ability of PW-Dialog to quickly compare BlenderBot3B to BlenderBot3B-M0 and BlenderBot90M (the Length and Size comparisons), and it also gains some of PW-Turn's superior strength at measuring the performance of BlenderBot3B over Reddit3B (the Fine-tuning comparison).

By contrast, since ratings for the two single-model techniques SM-Turn and SM-Dialog are collected simultaneously, ratings from both techniques on a given conversation can be averaged together to achieve slightly finer sensitivity than either technique individually. Figures 11, 12, and 13 show that, with the proper weighting, such averaging can produce a statistically significant difference between models a bit faster than with only SM-Dialog and dramatically faster than with only SM-Turn (Appendix C.8).

Beyond win rates, another way to directly compare the relative usefulness of our various evaluation techniques is to estimate the amount of person-hours that must be spent on evaluations by crowdworkers in order to achieve a statistically significant result. These results (Figures 8, 9, and 10) roughly follow the patterns found by win rates (Section 4.3.1). See Appendix C.7 for a discussion of the assumptions made when producing these time estimates.

5 Conclusion

In this work we compare the extent to which different evaluation techniques are able to measure performance differences between dialogue models, and we show instances in which the performance varies between per-turn techniques and per-dialogue techniques, and between pairwise techniques and single-model techniques. A completely exhaustive analysis of the cases in which each technique is most appropriate would require measurement on many more pairs of models than the three studied here, and would likely require a dramatic scaling-up of labor for crowdworkers.

Nevertheless, the results shown here demonstrate the difficulty in anointing one evaluation technique as superior to all others regardless of the models being compared, and they suggest that a combination of techniques, or else a different technique entirely, may be necessary for optimal measurement of differences among models. A more universally ideal technique would likely need to investigate model performance per-turn but still be able to give an overall judgment of model quality across a conversation in order to capture elements of performance that manifest clearest in a single response vs. in aggregate. We demonstrate that combining evaluation scores from per-turn and per-dialogue techniques can bridge the gap in the performance differences between the two, but that this does not outperform either individual technique in all cases, at least in the way that we combined them.

Future improvements may also come from exploring other ways to amplify the weak signal from models with only slight performance differences such as BlenderBot3B and BlenderBot90M, perhaps by training workers to select responses based on general measures of conversational quality, as opposed to content that appeals to their personal interests. Improving sensitivity to roughly equivalent pairs of models such as these should in turn enable the comparison of models whose performance differences are smaller still.

While this work has concentrated on evaluating techniques that enable *differentiability* (one can differentiate between models) with efficiency (with less annotator hours), there are other desirable qualities as well. Some of these in particular are *diversity* of conversations (Hashimoto et al., 2019), *repeatability* of experiments, and *explainability* of results (Deriu et al., 2021). While there is some discussion of the latter two topics in our experiments, these topics are fully deserving of a more thorough analysis than is provided here.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2021. Reason first, then respond: Modular generation for knowledge-infused dialogue. *arXiv preprint arXiv:2111.05204*.

Joshua Albrecht and Rebecca Hwa. 2007. A re-examination of machine learning approaches for sentence-level mt evaluation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 880–887.

Norman H Anderson. 1965. Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of personality and social psychology*, 2(1):1.

SE Asch. 1946. Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3):258–290.

Ram Ashwin, Prasad Rohit, Khatri Chandra, Venkatesh Anu, Gabriel Raefer, Liu Qing, Nunn Jeff, Hedayatnia Behnam, Cheng Ming, Nagar Ashish, King Eric, Bland Kate, Wartick Amanda, Pan Yi, Song Han, Jayadevan Sk, Hwang Gene, and Pettigrue Art. 2017. Conversational AI: The science behind the Adlexa Prize. In *Proceedings of Workshop on Conversational AI*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the International Conference on Learning Representations*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124.

Elizabeth Clark and Noah A Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3566–3575.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Jan Milan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Álvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019b. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 207–219. ACL.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

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.

Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, et al. 2020. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Alexa Prize Proceedings*.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Àgata

Lapedriza, and Rosalind W. Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. *Advances in Neural Information Processing Systems*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701.

Helen Hastie. 2012. Metrics and evaluation of spoken dialogue systems. In Oliver Lemon and Olivier Pietquin, editors, *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, pages 131–150. Springer.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 263–272.

Stephen J Hoch. 1984. Availability and interference in predictive judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(4):649.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Jessica Huynh, Jeffrey Bigham, and Maxine Eskenazi. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not. *arXiv preprint arXiv:2111.05241*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.

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 preprint arXiv:2010.12741*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003. ACL.

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*.

Weixin Liang, James Zou, and Zhou Yu. 2020. Beyond user self-reported likert scale ratings: A comparison model for automatic dialog evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1363–1374.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad 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.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1116–1126. ACL.

Norman Miller and Donald T Campbell. 1959. Recency and primacy in persuasion as a function of the timing of speeches and measurements. *The Journal of Abnormal and Social Psychology*, 59(1):1.

Bennet B Murdock Jr. 1962. The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482.

Leo Postman and Laura W Phillips. 1965. Short-term temporal changes in free recall. *Quarterly journal of experimental psychology*, 17(2):132–138.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication*, 74:12–36.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1702–1723. ACL.

Kurt Shuster, Jack Urbanek, Emily Dinan, Arthur Szlam, and Jason Weston. 2020. Deploying lifelong open-domain dialogue learning. *arXiv preprint arXiv:2008.08076*.

Neil Stewart, Gordon DA Brown, and Nick Chater. 2005. Absolute identification by relative judgment. *Psychological review*, 112(4):881.

Alan M Turing and J Haugeland. 1950. *Computing machinery and intelligence*. MIT Press Cambridge, MA.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*, 4:60–68.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of the 31st International Conference on Machine Learning, Deep Learning Workshop*, Lille, France.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 438–449. ACL.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7281–7288.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

A Additional existing work

Open-domain versus specific domain Our work concentrates on the open-domain setting. In specific conversational domains, such as question answering (QA), evaluation can be simpler and is often reduced to measuring overlap or exact match with the correct answer (Chen et al., 2019). However, this no longer as easily suffices for free-form, conversational and long-form QA where answers are more open-ended (Fan et al., 2019; Adolphs et al., 2021). Similarly, for certain types of goal-oriented conversations more targeted evaluations can take place, for example evaluation of state tracking (Williams et al., 2016), interaction quality (Schmitt and Ultes, 2015), and task completion (Hastie, 2012; Henderson et al., 2014; Bordes et al., 2017; El Asri et al., 2017; Wen et al., 2017). Open-domain dialogue potentially covers all these other cases as special cases, but also covers conversations that are more free-form or do not have a precise goal. Hence, finding a reliable evaluation technique is more difficult, and there is currently no single standard method that is agreed upon (Deriu et al., 2021; Huang et al., 2020; Roller et al., 2020). Different techniques that have been proposed will be described in the following paragraphs.

Automatic metrics Automatic metrics are the most convenient for fast, efficient and reproducible research with a quick turn-around and development cycle, hence they are frequently used. Unfortunately, many of them, such as BLEU, METEOR and ROUGE, have been shown to only "correlate very weakly with human judgement" (Liu et al., 2016). A central problem is that, due to the open-ended nature of conversations, there are many possible responses in a given dialogue, and, while having multiple references can help, there is typically only one gold label available (Gupta et al., 2019). Perplexity (computing the predicted probability of the given gold utterances) has been argued to correlate with human judgments (Adiwardana et al., 2020), however this has also been shown to not always be the case (Dinan et al., 2019b), and moreover does not actually evaluate the generations themselves produced by a decoder architecture. Hence, changing the behavior of the generation method can dramatically change human evaluations, while maintaining identical or near-identical perplexity (See et al., 2019; Welleck et al., 2020; Adiwardana et al., 2020; Roller et al., 2021). An alternative recent trend is to employ trainable metrics, whereby a neural network model is used to score the conversational model (typically also another neural network), see e.g. Lowe et al. (2017); Ghandeharioun et al. (2019). Such systems provide a promise of improved speed of research and development of dialogue agents, but so far have not been met with wide adoption. Some issues are that they may not generalize as well to data beyond that which they are trained (overfit) and also may be biased and gameable (Wu et al., 2019; Albrecht and Hwa, 2007). For a comprehensive comparison of automatic metrics – both standard and learned metrics – see Yeh et al. (2021). In general, creating a reliable automatic metric is still considered an open problem (Deriu et al., 2021).

Crowdworkers versus experts versus organic users While utilizing human evaluations in research is the current standard, we contend that choosing exactly which kind of human evaluation is also still an open question. In this work we concentrate on the study of crowdworker human evaluations, however there are several alternative paradigms. Utilizing trained experts, such as a group of researchers in the same institution, is one alternative (Deriu et al., 2021). Compared to employing crowdworkers, while model comparison results can agree between the two types of annotators, there can be vastly different sensitivity and win rates (Welleck et al., 2020), with the experts having more agreement and higher resulting sensitivity. On the other hand, it is harder to recruit and employ experts, limiting reproducibility. In both the crowdworker and expert annotator cases, neither of those groups are necessarily the intended target audience of a given system. If it is possible to deploy a model to people who genuinely want to talk to it (e.g., without being paid), conversations may be more natural and evaluations will be in line with genuine interests. Evaluation by deployment can be successful (Gabriel et al., 2020; Shuster et al., 2020), where behavioral metrics such as the amount of conversation time per user or retention rate can serve as a proxy for interestingness and engagingness metrics. Model deployment however also has its issues. First, user desires may not necessarily be aligned with the goals of the research itself, meaning researchers may have to develop features and improvements towards the goals of the product rather than towards long-term research. Further, experiments are difficult to set up and may be difficult to reproduce by other groups. Crowdworker tasks can be more reproducible especially when code is made available to reproduce experiments, but there are also many pitfalls when constructing the tasks, see e.g. Huynh et al. (2021).

Conversation instructions to raters When utilizing evaluators in a evaluator-model conversational setup, the precise instructions on how to go about the conversation will clearly have large effects. Such instructions can control the topic, e.g. "get to know each other" as in the Persona-Chat task (Zhang et al., 2018), versus "have a knowledgeable conversation" in Wizard of Wikipedia (Dinan et al., 2018). Instructions can also orient workers towards a more fruitful strategy for a desired dataset, for example orienting them towards open questions on sensitive topics rather than profanity to get a bot to generate unsafe utterances (Xu et al., 2020). The length of the conversation will also play a role in the performance of models, for example, short conversations do not test the ability of models to retain knowledge in the long-term (Xu et al., 2021). Overall, the style of conversation has large effects (even if the topic is unchanged) for example when instructing crowdworkers to be adversarial vs. non-adversarial (Dinan et al., 2019a), which relates to the classic Turing Test (Turing and Haugeland, 1950). Further, particular instruction wording choices will change the quality of conversations, as they will change how well crowdworkers understand the task (Huynh et al., 2021).

Evaluation question phrasing for raters Besides how the conversation is carried out, one also needs to choose the precise question (or questions) being asked to crowdworkers in order for them to rate conversations. In open-domain conversation there are a variety of qualities one could expect from a good conversationalist, and potentially one could ask about any of them individually, as well as asking for overall performance. For example, See et al. (2019) asks evaluators for ratings of interestingness, making sense, fluency, avoiding repetition, listening ability and inquisitiveness as intermediate conversational aspects, and humanness and engagingness questions to measure overall quality. Adiwardana et al. (2020) asks questions based on sensibleness and specificity. Responsibility, toxicity and bias can also be measured (Xu et al., 2020). Even after settling on the exact question(s) to be asked, their exact phrasing also has impact on sensitivity, as shown in Li et al. (2019). In that work, the authors optimized the question phrasing by running evaluations with alternative phrasings, and choosing the one with the highest agreement.

Rating existing versus own conversations The standard setup is for a human to have a conversation with a model, and rate that conversation. Some evaluation protocols deviate from this setup, and ask evaluators to rate conversations they did not participate in. One simple approach of that kind is to present model completions of a dialogue from the fixed test set of a given task, and ask for their evaluation, with hence no human taking part in the actual conversation (Vinyals and Le, 2015; Li et al., 2016). In the Acute-Eval method (Li et al., 2019) raters are asked to compare two existing conversation logs, and the authors consider both the case of human-model chat logs, and model-model (self-chat) logs, where the former are actually a different set of human conversationalists compared to the final raters. Deriu et al. (2020) considers chat logs between pairs of models, again with no humans taking part in the conversations. These techniques allow efficient reuse of existing conversational data and have some reproducibility gains: conversations collected in previous trials and by other systems can be directly compared with a new system, without having to recollect additional data. This can significantly reduce the resources needed by a new evaluation, and ensure that multiple papers are comparing to prior work consistently. On the other hand, it may be harder for evaluators to rate conversations that they have not been involved in (Finch and Choi, 2020). Conversations that do not even involve humans should be treated with some scepticism, as there is no human to guide conversation and hence evaluate interactive quality. Nevertheless, such approaches do appear to be useful experimentally (Li et al., 2019; Roller et al., 2021).

Pairwise versus single-model ratings Conversations are often either rated individually, e.g. with Likert-score ratings (Ashwin et al., 2017; Venkatesh et al., 2018; Zhang et al., 2018; Rashkin et al., 2019; See et al., 2019; Dinan et al., 2019b, 2018), or pairwise by comparing models (Li et al., 2019; Liang et al., 2020; Vinyals and Le, 2015; Li et al., 2016; Lee et al., 2020). Likert scoring relies on absolute identification rather than relative discrimination, which is less reliable in humans (Stewart et al., 2005), leading to different biases per annotator (Kulikov et al., 2019). It is thus often necessary to then re-evaluate existing models at the same time as a new model, as the distribution of human annotators can easily shift over time, causing measurement errors (See et al., 2019). Another common difficulty is related to sequential effects (Stewart et al., 2005), where the annotator can be influenced by the first model they evaluate, causing difficulties in using an absolute scale. Pairwise comparisons, on the other hand, make comparing a set of models less efficient, and also have the same problem that existing baseline models have to be essentially reassessed with respect to new ones.

Per-turn versus per-dialogue evaluation Some research evaluates *single-turn* responses in conversations given gold dialogue contexts, without taking into account whole interactive conversations (Lee et al., 2020; Vinyals and Le, 2015; Li et al., 2016). This fails to take into account multi-turn aspects of a conversation, for example a model repeating itself over multiple turns. *Per-turn* evaluation instead conducts an entire conversation, but raters are still asked to evaluate each turn (response by their partner) (Schmitt and Ultes, 2015; Adiwardana et al., 2020; Komeili et al., 2021). Collecting per-turn evaluation also allows for measuring learning effects where workers become more adept at interacting with the bot for certain specific tasks (e.g., see Xu et al. (2020)). In contrast, methods like multi-turn Likert or Acute-Eval ask evaluators to assess the entire dialogue as a whole, rather than the individual turns, under the assumption that the quality of a conversation is not simply the sum of its parts. Literature from psychology predicts several effects when considering how people combine their impressions from single conversational turns into an evaluation of an entire conversation. The *primacy effect* refers to how overall judgment is more shaped by characteristics presented earlier (Asch, 1946; Anderson, 1965). Conversely, the *recency* *effect* appears when more weight is given to information presented the most recently, and both effects combine to give more weight to items at the beginning and end of a list (Murdock Jr, 1962; Postman and Phillips, 1965), with the recency effect being more prominent when judgment is elicited without any delay when the recent information is still fresh (Miller and Campbell, 1959; Hoch, 1984).

B Additional methods

B.1 Screenshots of crowdsourced human evaluation tasks

See Figure 2 for a screenshot of the PW-Turn evaluation technique, Figure 3 for a screenshot of the PW-Dialog technique, and Figure 4 for a screenshot of the SM-Turn and SM-Dialog techniques. Figure 5 additionally displays the onboarding UI for PW-Turn.

B.2 Pairwise per-turn evaluation onboarding

In order to perform quality control on crowdworkers before the start of the conversation itself, we ask each worker to first annotate a conversation in which there are two possible responses for each turn of one of the speakers, one response of which is clearly better than the other (Figure 5). These pairs of responses vary slightly depending on which of the evaluation metrics is being tested. Workers must ultimately choose the correct response for all four pairs of responses but have two tries in which to do so.

B.3 Quality checks on crowdworkers

In order to ensure that our comparisons between evaluation techniques are not affected by variability in the pool of crowdworkers when running one technique vs. another, we adopt a consistent set of criteria across all techniques regarding which workers to exclude from our final set of data. If a worker fails one of the checks in Appendix B.4 during one of the per-turn evaluations PW-Turn or SM-Turn, we retroactively exclude their ratings from all of the evaluation techniques.

In order to prevent any worker from disproportionately contributing to the final results, each worker is restricted to one conversation per model pair and evaluation metric (for PW-Turn and PW-Dialog) or one conversation per model (for SM-Turn and SM-Dialog). All evaluations are collected among residents of the United States on weekdays, from roughly 9 AM to 6 PM in the U.S. Eastern time zone, following Li et al. (2019).

B.4 Checks used when filtering per-turn evaluations

We check each conversation between a crowdworker and a Bot Speaker collected during PW-Turn and SM-Turn evaluations against the criteria below to see if they have issues that warrant their exclusion from the final filtered set of evaluations. If at least one of the following problems is present, all evaluations from the crowdworker in question are filtered out of the results shown in this work:

- The messages consist of less than three words on average
- The first message inputted by the worker contains a greeting (redundant, since a dummy *"Hi!"* message is already fixed to be the worker's first line of conversation)
- Several of the messages are written using all capital letters
- Later messages are duplicates of the first one (i.e. the worker is repeating their messages throughout the conversation)
- One or more of the messages use offensive language

C Additional results

C.1 Evaluation data collection

After filtering out workers with unacceptable messages following Appendix B.3, we are left with a minimum of 144 and a mean of 231 ratings (typically 6 per conversation) for each of the PW-Turn evaluations, a minimum of 191 and a mean of 324 ratings for PW-Dialog, a minimum of 349 and a mean of 411 ratings (typically 6 per conversation) for SM-Turn, and a minimum of 58 and a mean of 68 ratings for SM-Dialog evaluations (for which there is only one rating per conversation). On average, the collection of ratings after filtering represents 5.73 hours of worker labor for PW-Turn per model pair and evaluation metric, 6.03 hours for PW-Dialog per model pair and evaluation metric, and 4.39 hours for joint SM-Turn/SM-Dialog evaluations per model.

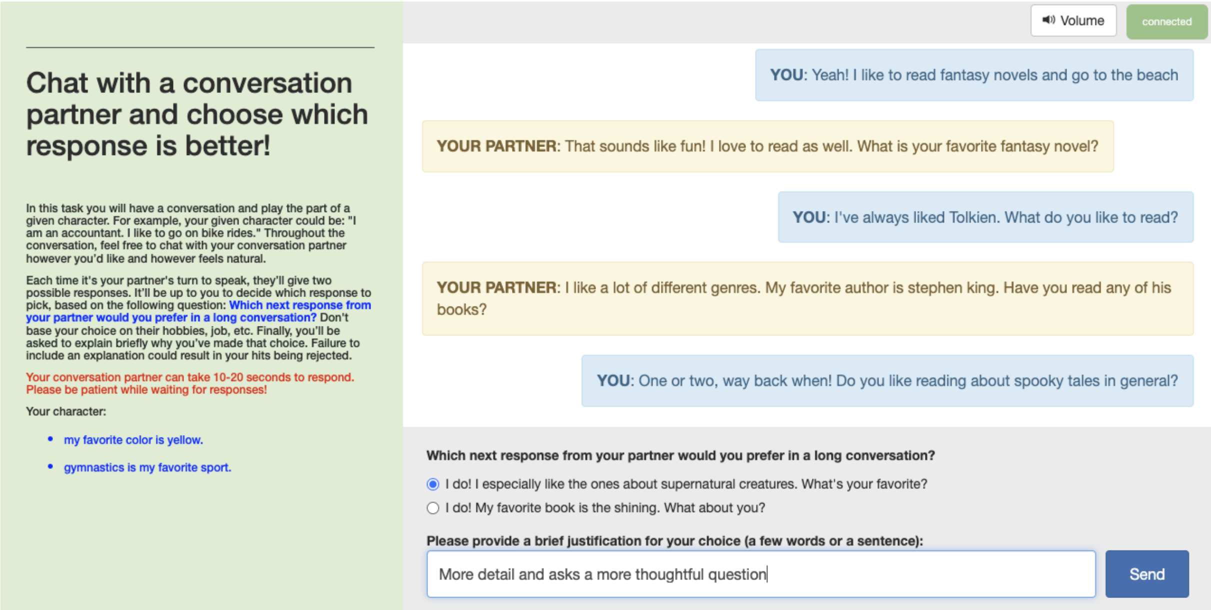

Figure 2: Screenshot of the Pairwise Per-Turn (PW-Turn) evaluation technique, in which we ask crowdworkers to choose one of two possible responses from their conversation partner and describe why that response is better. The two responses come from the two models that we are trying to compare the performance of.

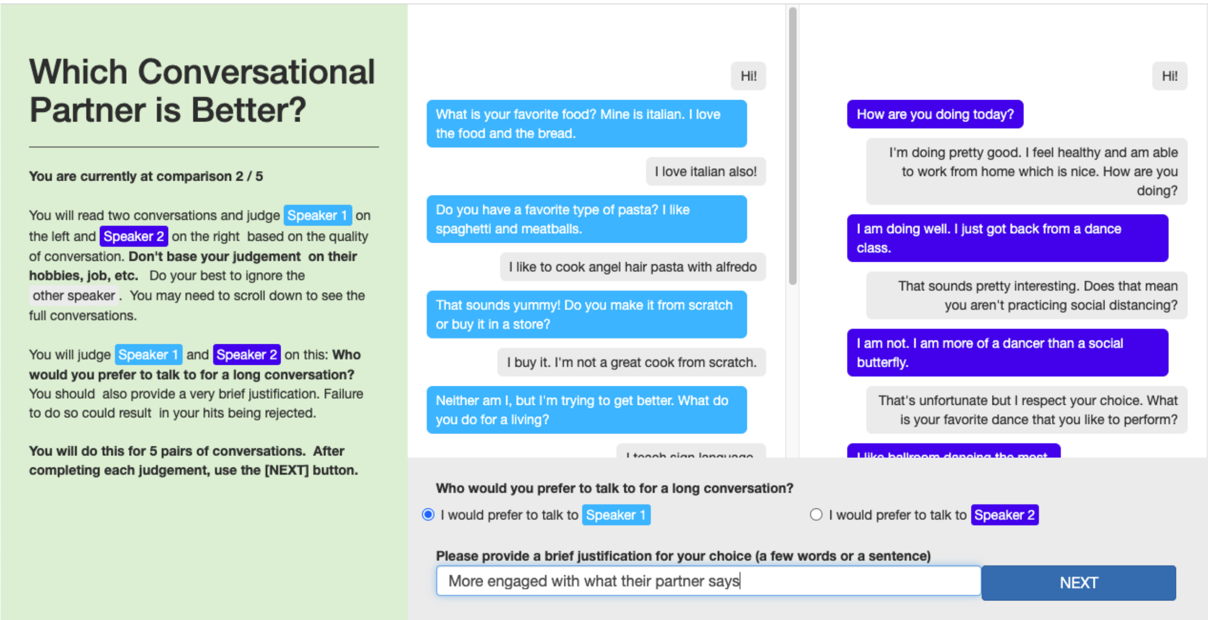

Figure 3: Screenshot of the Pairwise Per-Dialogue (PW-Dialog) evaluation technique, in which we ask crowdworkers to choose which of two speakers in two separate conversations is better on the given evaluation metric, here *"Who would you prefer to talk to for a long conversation?"* The crowdworker must then describe why that speaker is better.

C.2 PW-Turn model win rates as a function of turn

Unlike PW-Dialog, the PW-Turn technique is able to measure differences in the win rate of models as a function of the number of turns into the conversation. In Figure 6 and Table 4, we see that BlenderBot3B's win rates tend to be closer to 50% in the first 1 or 2 turns of the Bot Speaker, and higher later: this may be because the first few lines of the conversation typically consist of greetings (*"Hi, how are you?"*) or pleasantries, which may be harder to judge model performance on. However, it may also be because improvements are accumulated and factored into evaluators' decisions

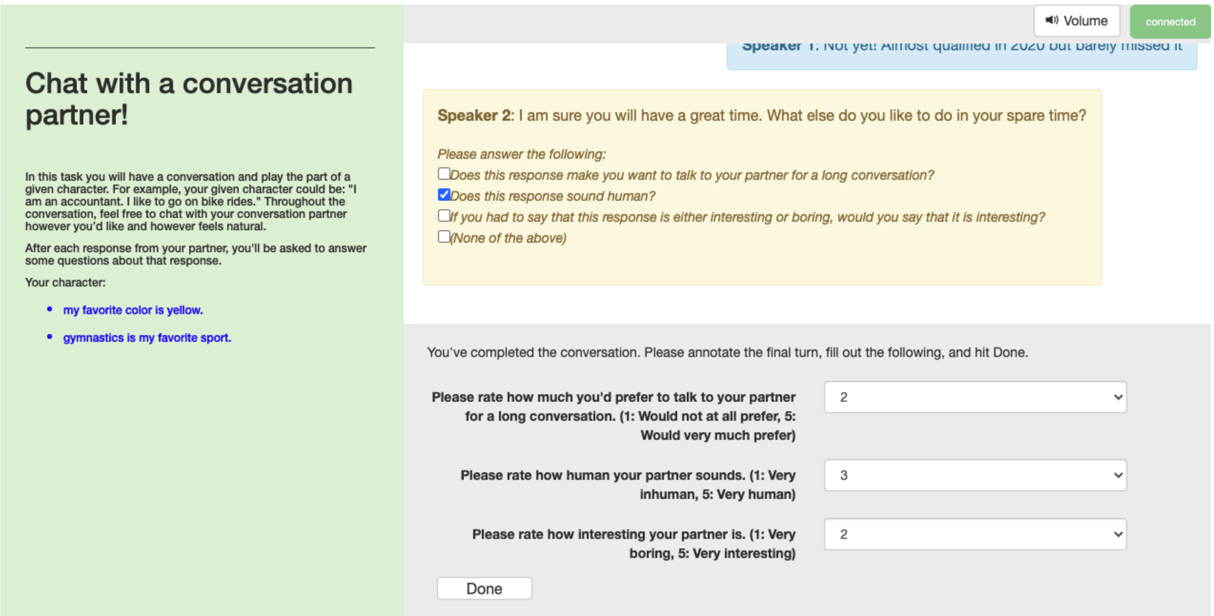

Figure 4: Screenshot of the crowdsourcing task for collecting Single-Model Per-Turn (SM-Turn) and Single-Model Per-Dialogue (SM-Dialog) evaluations. We ask the crowdworker to annotate each response from their partner along several dimensions, as well as give a global Likert-scale rating of their partner's performance at the end of the conversation.

		PW-Turn: turn index					
Comp.	Metric	1	2	3	4	5	6
Length	Pref	40	56	67	67	**74**	72
	Human	38	**72**	69	62	**72**	62
	Inter	59	62	72	**78**	69	69
Size	Pref	26	56	51	54	38	**62**
	Human	29	58	50	46	**67**	58
	Inter	31	59	**62**	52	45	45
FT	Pref	70	78	74	**89**	81	**89**
	Human	69	73	**90**	88	84	85
	Inter	52	59	**89**	74	67	85

Table 4: Percentage win rates of BlenderBot3B vs. other models on PW-Turn evaluations as a function of Bot Speaker turn. The highest win rate for each model comparison and evaluation metric is bolded. This is a tabular representation of the curves in Figure 6.

		SM-Turn: turn index					
Model	Metric	1	2	3	4	5	6
BB3B	Pref	67	71	**72**	70	**72**	70
	Human	70	63	72	**76**	71	70
	Inter	42	43	**47**	43	45	**47**
BB3B-M0	Pref	**74**	**74**	**74**	67	66	72
	Human	66	**72**	64	67	69	66
	Inter	29	**40**	38	34	36	33
BB90M	Pref	**71**	67	65	64	64	64
	Human	59	64	61	**70**	65	67
	Inter	39	35	39	**41**	**41**	38
Reddit3B	Pref	**67**	63	49	50	54	46
	Human	**60**	57	54	53	51	58
	Inter	44	36	39	**46**	31	39

Table 5: Percentage success rates of responses of various models on various evaluation questions (metrics) for SM-Turn, as a function of Bot Speaker turn. The highest win rate turn for each model and evaluation metric is bolded.

later in the conversation. Strikingly, BlenderBot3B performs very poorly vs. BlenderBot90M (the Size comparison) on the first Bot Speaker turn: empirically, this may be due to the fact that BlenderBot3B generally starts its first responses with the greetings *"Hi"* or *"Hello"* much less frequently than BlenderBot90M does.

C.3 SM-Turn success rates as a function of conversation turn

See Table 5 for the success rates of model responses using the SM-Turn technique, as a function of Bot Speaker turn.

C.4 Relationship between per-turn ratings and final ratings

Given that SM-Turn allows us to measure per-turn ratings of model performance, it is worth exploring whether there are certain turns of the conversation that contribute more strongly to the workers' final Likert-scale ratings of the conversation (SM-

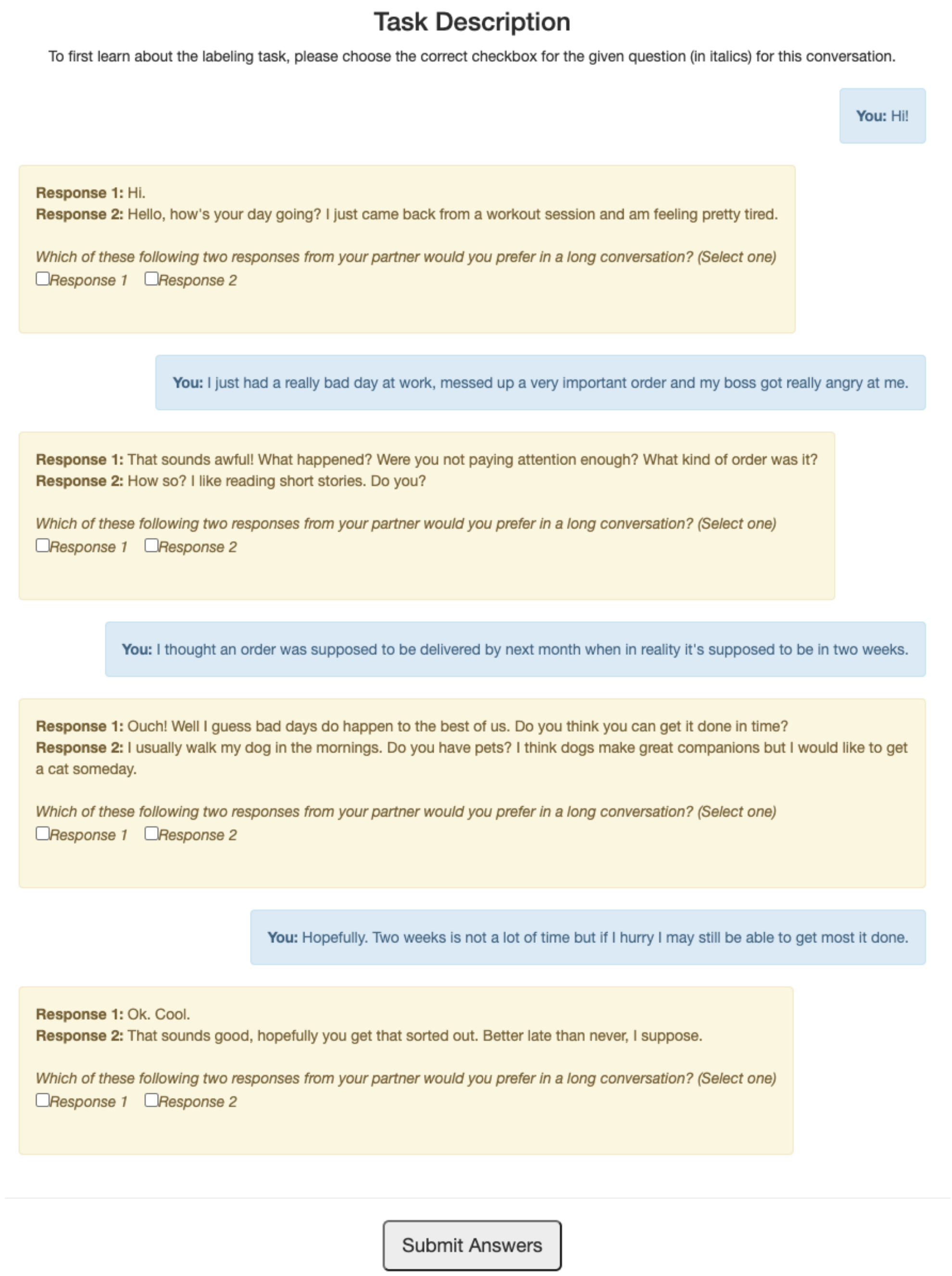

Figure 5: Screenshot of the onboarding process for crowdworkers for the PW-Turn technique.

Dialog). Figure 7 plots the coefficients of workers' per-turn SM-Turn ratings in an OLS regression, with the SM-Dialog Likert score at the end of the conversation as the dependent variable. (Here, we reduce variability by taking the mean over the three evaluation metrics for each turn's SM-Turn ratings and SM-Dialog Likert scores.) Generally, we see a higher positive coefficient of the SM-Turn ratings in later turns in the conversation, which implies that the workers may have a recency bias: they may remember the most recent turns of the conversation more strongly when determining how to rate the model's performance overall.

C.5 Text justification for model response selection

For PW-Turn evaluations, we collect and analyze justification texts for each turn, after the worker selects a model response. We then group justification texts by model type and comparison.

To measure lengths of justifications, we split text strings into words (space-delimited), and we

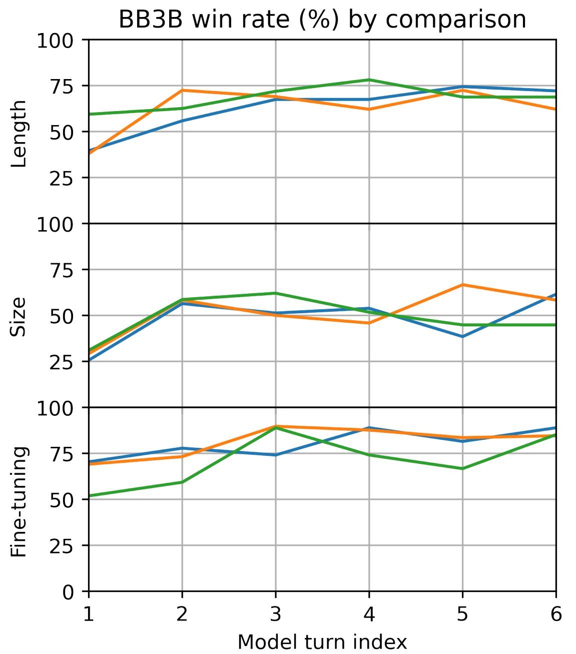

Figure 6: Win rate of BlenderBot3B vs. other models for the comparisons in Section 3.2 on the PW-Turn evaluations, as a function of the number of Bot Speaker turns into the conversation, for the Preference (blue), Humanness (orange), and Interestingness (green) metrics. BlenderBot3B tends to fare better against other models in later turns of the conversation.

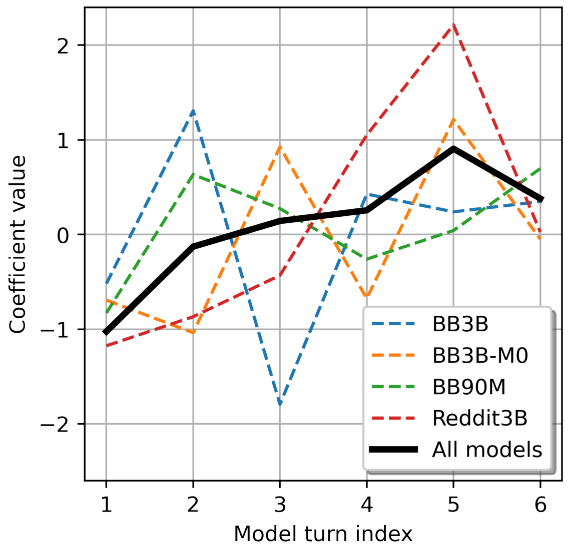

Figure 7: The per-turn coefficients of SM-Turn success rates in an OLS regression with SM-Dialog Likert scores as the dependent variable. SM-Turn rates and SM-Dialog scores are averaged across evaluation metrics. The black curve represents data from all models concatenated together. SM-Turn rates from later turns tend to be more positively correlated to the final SM-Dialog Likert scores, suggesting a possible recency bias.

calculate the mean number of words in each sample. Results are shown in Table 6.

Comparison	Model	Avg. number of words
Length	BB3B	8.85
	BB3B-M0	7.70
Size	BB3B	8.95
	BB90M	8.81
Fine-tuning	BB3B	9.40
	Reddit3B	9.25

Table 6: Mean number of words in justifications given for BlenderBot3B vs. other models on PW-Turn evaluations.

For term importance, we use the `scikit-learn TfidfVectorizer` class to compute TF-IDF scores for each term in each model comparison.

We use a list of English stopwords from the `NLTK` library to filter out common terms. Additionally, we discard terms that have a higher document frequency than 0.8.

The top 20 terms (descending order) for each model pairing are shown in Table 7.

Our analysis reveals the following:

- **Length comparison**: While it appears that many crowdworkers prefer longer responses overall, at least in some conversational turns some crowdworkers may prefer shorter responses. The top terms in justifications for BlenderBot3B-M0 responses include “simple”, “short” and “direct”, while top terms in reasons for choosing BlenderBot3B include “detailed” and “longer”. This shows that PW-Turn evaluation does well in capturing sensitivity to length, and that workers’ selections are due to their own preferences at a given conversational turn.

  Interestingly, in PW-Turn we find that workers’ justifications for choosing the BlenderBot3B-M0 responses are themselves on average shorter than for BlenderBot3B. Table 6 shows the mean justification lengths for different model pairings. The mean justification length for BlenderBot3B is 8.85 words, compared to a mean length of BlenderBot3B-M0 justifications of 7.7 words. This suggests that workers choosing shorter, “simple” responses may also be less detail-oriented.

- **Size comparison**: Top TF-IDF weighted terms from workers' justifications for both models contain a mix of references to the conversational content, such as "hiking", "beach" or "dogs", and conversational structure, such as "relates" or "engaging". By inspection, there are no discernible differences between these terms.

- **Fine-tuning comparison**: High TF-IDF-weighted terms in justifications given by workers who choose the BlenderBot3B model are mostly related to conversational flow, such as "follows", "responds", and "acknowledges". In contrast, terms appearing in justifications for the Reddit3B model are specific and often refer to the topic instead of conversational style, such as "bath", "robot", and "paris". This suggests that workers who choose the Reddit3B model often favor less natural responses because they contain particular references.

These nuanced differences are clear when evaluating model responses per turn, but are difficult to capture when evaluating the conversation as a whole. Analysis of worker justifications supports our hypothesis that differences in conversational quality are easier to identify in the PW-Turn evaluation.

C.6 Variability in win rate across evaluation techniques

Table 8 shows the variability in the win rates of BlenderBot3B per evaluation technique, as measured by splitting the ratings from each technique into chunks of equal crowdworker time. The win rates from PW-Turn, PW-Dialog, and SM-Turn vary least across chunks, largely because the mean time per rating is small, leading to a larger number of ratings per chunk and thus a more precise estimate obtainable within a given block of time.[5] This suggests that calculating the per-conversation winner-takes-all win rate for the per-turn methods PW-Turn and SM-Turn may be disadvantageous if having a precise measurement of the win rate is more important than one that is statistically significant.

[5]We omit win rates of PW-Dialog on conversations between a human and a model for simplicity. For this technique, the time to collect conversations varies non-linearly as a function of the number of ratings (Section C.7), and thus any dividing of ratings into chunks of equal crowdworker time would have to take this irregularly-spaced conversation collection time into account.

C.7 Crowdsourcing time needed per technique

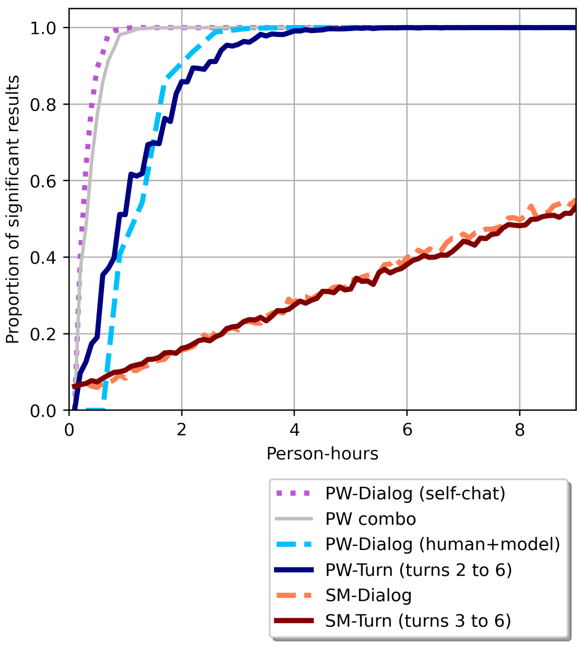

Figure 8: The time needed for statistical significance for the Length comparison between models (BlenderBot3B vs. BlenderBot3B-M0).

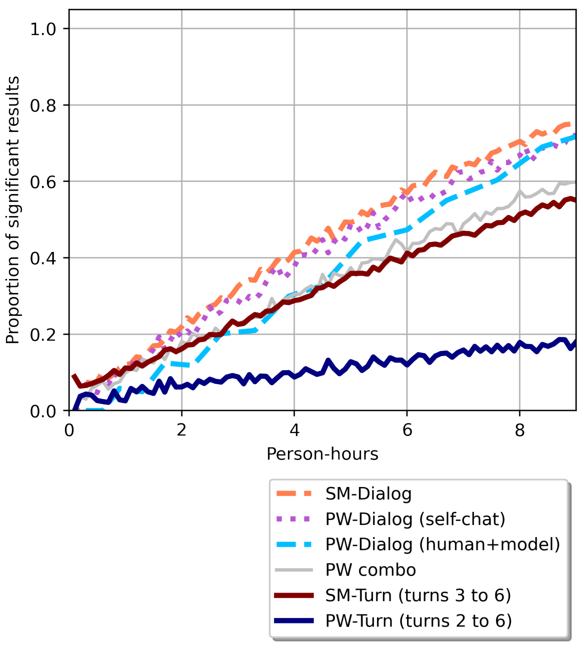

Figure 9: The time needed for statistical significance for the Size comparison between models (BlenderBot3B vs. BlenderBot90M).

Figures 8, 9, and 10 show the time needed to achieve a statistically significant difference between a pair of models for each of the evaluation

Comparison	Model	Top terms
Length	BB3B	information, chosen, provides, follow, engaging, going, adds, speaker, detailed, interested, conversational, looks, little, *play*, chat, new, *pets*, includes, longer, *tallies*
	BB3B-M0	*day*, *wow*, *game*, going, simple, stays, short, direct, express, speaker, keeps, conversational, precise, *western*, *popular*, *silk*, *hands*, use, tone, elaborate
Size	BB3B	message, information, easy, time, correct, want, interested, enjoy, change, relates, spend, prefer, fun, well, *hiking*, *pets*, go, moves, *beach*, sound
	BB90M	never, going, ok, excited, *fav*, correct, changes, *color*, new, engaging, personal, explain, *ohio*, fluent, enjoy, *hop*, *hip*, listen, back, *dogs*
Fine-tuning	BB3B	follows, going, contradicts, great, responds, follow, responsive, never, contradict, flows, acknowledges, responses, responded, stays, looks, personal, keep, well, nothing, contradiction
	Reddit3B	*bath*, personal, *robot*, im, *someone*, *bubble*, detailed, flowing, *play*, information, *paris*, due, *softball*, *careers*, unique, direct, watch, told, *book*, boring

Table 7: Top TF-IDF-weighted terms in justifications given for BlenderBot3B responses vs. other models on PW-Turn evaluations. Terms that are irrelevant to conversational evaluation are italicized.

Technique	Length	Size	Fine-Tuning
PW-Turn	10%	8%	11%
PW-Turn (WTA)	18%	24%	13%
PW-Dialog (self-chat)	9%	9%	6%
SM-Turn	14%	13%	12%
SM-Turn (WTA)	17%	16%	15%
SM-Dialog	14%	15%	16%

Table 8: The variability of win rates of BlenderBot3B across different evaluation techniques, for different model comparisons (columns). Variability was measured by splitting each time-ordered set of ratings into chunks representing 45 minutes of crowdworker time each, and then computing the standard deviation of the win rate across chunks. Standard deviations are averaged across the three evaluation metrics (Section 3.1.2). Win rates for PW-Turn were compiled over Bot Speaker turns 2 to 6 and for SM-Turn over turns 3 to 6, following Section 4.3.1.

techniques studied. For these plots, we consider ratings for each turn in Bot Speaker turns 2 through 6 for PW-Turn and Bot Speaker turns 3 through 6 for SM-Turn, as in Section 4.3.1.[6] We use a two-sided binomial test for PW-Turn and PW-Dialog and a two-sided independent t-test for SM-Turn and SM-Dialog. Significance is measured at a p-value of 5%. When estimating the crowdsourcing time needed for each evaluation technique, we include an estimate of each technique's time to complete onboarding, which is mandatory before being approved to work on an evaluation.

[6]We do not compute winner-takes-all scores for each conversation because in experiments this works less well. It greatly diminishes the total number of ratings per technique, and thus it increases the number of conversations needed to achieve statistical significance. We note that, compared to the per-dialogue technique SM-Dialog, the resulting rating is binary per evaluation metric in this case, which may contribute to poor performance.

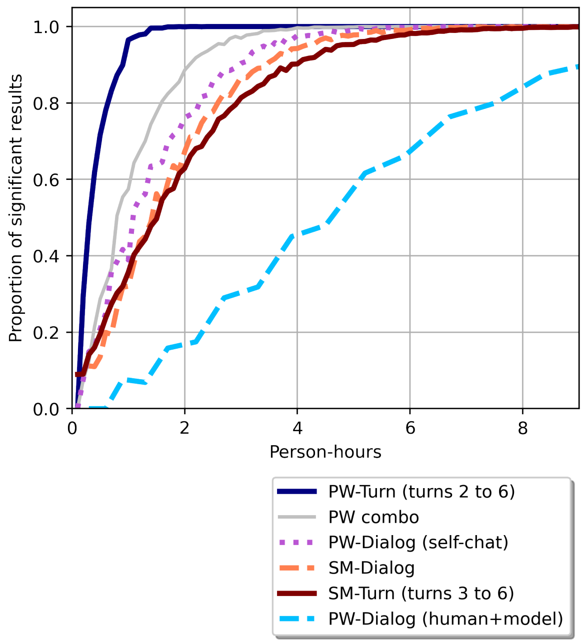

Figure 10: The time needed for statistical significance for the Fine-tuning comparison between models (BlenderBot3B vs. Reddit3B).

For PW-Dialog evaluations (i.e. Acute-Evals) on conversations between a human and a model, the labor costs involve collecting both conversations and rating pairs. This gives us a parameter to tune in this method: how many conversations to collect, and then how many times to reuse them when rating pairs of them. In our experiments, the number of conversations necessary is chosen such that each possible pairing of a conversation from one model and a conversation from another model should only be evaluated once at most: thus, if we have N conversations for each of the two models being compared, we will be able to perform a maximum of N^2 PW-Dialog evaluations on these conversations.[7]

C.8 Crowdsourcing time needed when combining single-model methods

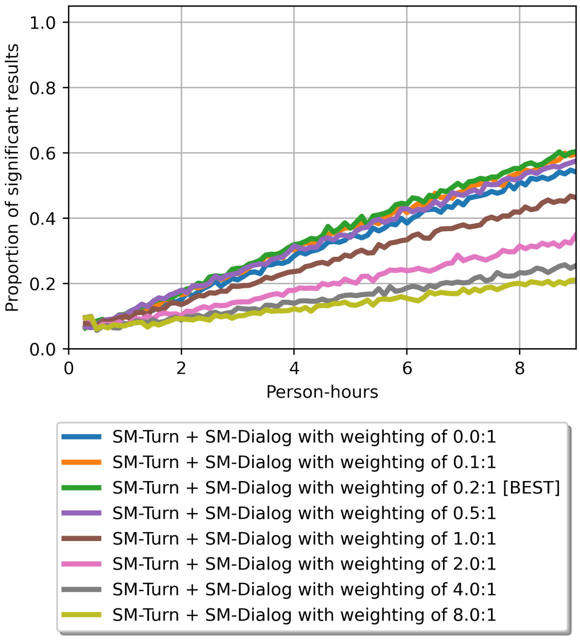

Figure 11: The time needed to measure a statistically significant result when averaging together per-conversation evaluations of SM-Turn and SM-Dialog with the given weighting, for the Length comparison. The fastest weighting is marked with “[BEST]”.

Figures 11, 12, and 13 show the time needed to achieve a statistically significant difference between models when averaging together SM-Turn winner-takes-all success rates from Bot Speaker turns 3 to 6 (Section 4.2) with SM-Dialog Likert scores. To perform the weighted average between SM-Turn and SM-Dialog on each conversation, we first shift and scale the originally 1-to-5 SM-Dialog Likert scores to fall within the range $[0, 1]$, matching the range of the individual binary SM-Turn success rates. We see that statistical significance is reached fastest when weighting SM-Turn much less heavily than SM-Dialog at a ratio of 1:5 or 1:10, which is to be expected given the already much stronger sensitivity of SM-Dialog.

[7] The potential drawback of this assumption is that the performance of the models will then likely be judged using only a relatively small handful of conversations, which may or may not be representative of the models' true performance.

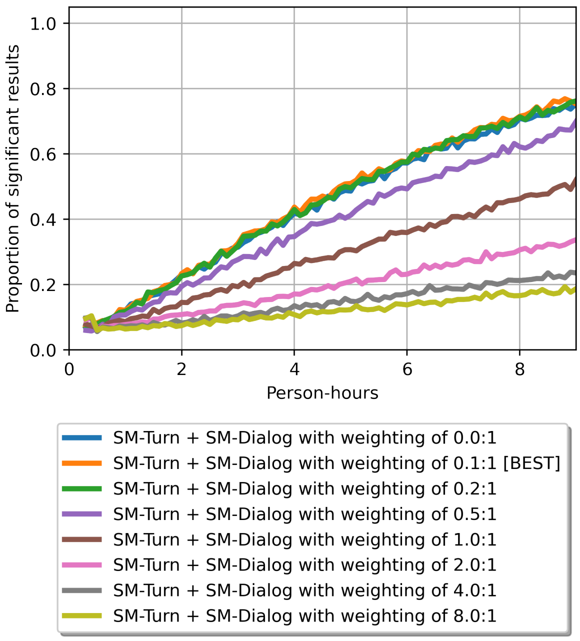

Figure 12: The time needed to measure a statistically significant result when averaging together per-conversation evaluations of SM-Turn and SM-Dialog with the given weighting, for the Size comparison.

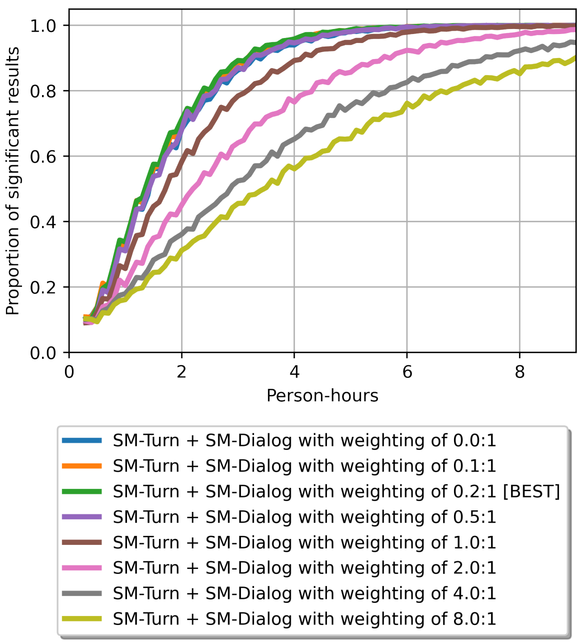

Figure 13: The time needed to measure a statistically significant result when averaging together per-conversation evaluations of SM-Turn and SM-Dialog with the given weighting, for the Fine-tuning comparison.