Towards Credible Human Evaluation of Open-Domain Dialog Systems Using Interactive Setup

Sijia Liu, Patrick Lange, Behnam Hedayatnia, Alexandros Papangelis,
Di Jin, Andrew Wirth, Yang Liu, Dilek Hakkani-Tur
Amazon Alexa AI
{sijial, patlange, behnam, papangea, djinamzn, wirandre, yangliud, hakkanit}@amazon.com

Abstract

Evaluating open-domain conversation models has been an open challenge due to the open-ended nature of conversations. In addition to static evaluations, recent work has started to explore a variety of per-turn and per-dialog interactive evaluation mechanisms and provide advice on the best setup. In this work, we adopt the interactive evaluation framework and further apply to multiple models with a focus on per-turn evaluation techniques. Apart from the widely used setting where participants select the best response among different candidates at each turn, one more novel per-turn evaluation setting is adopted, where participants can select all appropriate responses with different fallback strategies to continue the conversation when no response is selected. We evaluate these settings based on sensitivity and consistency using four GPT2-based models that differ in model sizes or fine-tuning data. To better generalize to any model groups with no prior assumptions on their rankings and control evaluation costs for all setups, we also propose a methodology to estimate the required sample size given a minimum performance gap of interest before running most experiments. Our comprehensive human evaluation results shed light on how to conduct credible human evaluations of open domain dialog systems using the interactive setup, and suggest additional future directions.

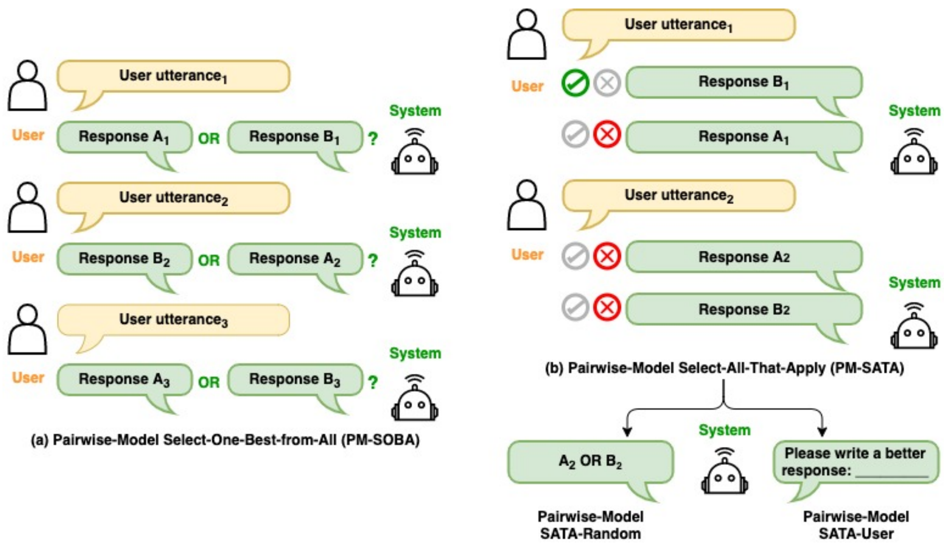
Introduction

Building open-domain chatbots that can converse with humans freely has been a challenging area for NLP. Unlike task-oriented conversations where task completion is critical, human users can talk about any topics during open-domain conversations. This open-ended nature thus has posed more difficulty on evaluating models in a credible and reproducible way. Plenty of work has focused on developing automatic evaluation metrics or performing static human evaluations to compare different conversational models. However, automatic metrics have been shown to have very weak correlations with human judgements on open-domain conversations (Liu et al. 2016; Lowe et al. 2018; Mehri and Eskenazi 2020); on the other hand, performing static human evaluations can be both time and cost intensive (Deriu et al. 2020). Additionally, most post-hoc static evaluations cannot reflect the quality of dialogs that are collected in a realistic interactive setup, and hence are not necessarily the most accurate assessment of a user's satisfaction with the conversation (Liu et al. 2016; Ghandeharioun et al. 2019). Therefore, apart from several public competitions such as Alexa Prize[1], ConvAI2 (Dinan et al. 2019) and DSTC9 Track 3 (Mehri et al. 2022) in English, as well as Dialog System Live Competition (Higashinaka et al. 2021) in Japanese, several studies have adopted interactive evaluations to either annotate per-turn model responses in a multi-turn dialog (Ghandeharioun et al. 2019; Adiwardana et al. 2020), or collect overall per-dialog ratings on a Likert scale to evaluate different aspects of conversation quality (Zhang et al. 2018; See et al. 2019; Dinan et al. 2018; Finch and Choi 2020; Ji et al. 2022). Smith et al. (2022) compared five different interactive setups including per-turn, per-dialog and self-play evaluations, and provided a comprehensive analysis for single-model and pairwise evaluations. However, questions still remain regarding which interactive evaluation setup is more appropriate for different evaluation scenarios (e.g., when more than two models need to be evaluated).

In this work, we apply the interactive evaluation framework to multiple models with a focus on *per-turn* evaluation techniques. We compare Multi-Model evaluation mechanisms with existing Single-Model and Pairwise-Model evaluations, and perform a thorough analysis across all three mechanisms. We adopt two *per-turn* evaluation setups: *Select One Best from All* (SOBA), where users choose the best response from a list of system response candidates and *Select All That Apply* (SATA) where users choose all the best responses from the list or choose none if they don't like of the responses. For the latter, users will either continue the conversation with a random system response (SATA-Random), or write a suggested system response (SATA-User), which will be used in the conversation.

In contrast to (Smith et al. 2022) where model rankings are pre-assumed before running evaluations, we instead compare the sensitivity and consistency of different evaluation setups without any assumptions on model rankings, and then use the results to understand model performance. To achieve this goal, we propose a methodology to determine the required sample size given a minimum performance gap of interest, which not only helps control evaluation costs

[1] https://www.amazon.science/alexa-prize/socialbot-grand-challenge



(a) Pairwise-Model Select-One-Best-from-All (PM-SOBA)

(b) Pairwise-Model Select-All-That-Apply (PM-SATA)

Figure 1: Illustration of Pairwise-Model evaluation. At each turn, the system provides two randomly ordered responses generated by models A and B, and the user does not know if two responses come from the same or different models. In PM-SOBA, users always select a *preferred* response. In PM-SATA, users can select 0 to 2 *appropriate* responses; when no response is selected, the system returns a *random* response if using SATA-Random, or asks the user to write a better *pseudo* system response if using SATA-User. Multi-Model evaluation uses the same setups, with more than two responses provided at each turn.

but also better generalizes to scenarios where model rankings remain unknown. It is worth noting that our focus is on comparing per-turn evaluation setups instead of investigating which models are better, and we leave per-dialog or self-chat evaluations to future work.

In this study, we have compared all three per-turn evaluation setups across Single-Model, Pairwise-Model, and Multi-Model evaluations, and find that:

- Pairwise-Model (PM) and Multi-Model (MM) evaluations generate highly consistent results for model rankings given a minimum performance gap of interest, while Single-Model (SM) has the least sensitivity to differentiate model performance;
- For a model pair of interest, Pairwise-Model Select-All-That-Apply (PM-SATA) works better at consistently measuring actual performance gap and providing explainability of model performance at higher costs, while Pairwise-Model Select-One-Best-from-All (PM-SOBA) is more sensitive and works better when models are clearly different;
- For more than two models, Multi-Model Select-All-That-Apply (MM-SATA) works better at performing a comprehensive test on both overall equality of all models and pairwise equality for all matched-pairs (which MM-SOBA cannot) with a reasonable sample size;
- Given a setup, the required sample size is a function of a minimum performance difference of research interest. Following the estimations, experiments show high consistency across Pairwise-Model (PM) and Multi-Model (MM) evaluations for most model pairs.

Related Work

Evaluating Open-Domain Conversations Using Interactive Setup. Recent work has adopted interactive settings to evaluate open-domain dialog models both in public competitions (Ram et al. 2018; Dinan et al. 2019; Mehri et al. 2022; Higashinaka et al. 2021) and development of several state-of-the-art models (Adiwardana et al. 2020; Roller et al. 2020). Mehri et al. (2022) recruited real users to evaluate a set of chatbots and provide turn-level ratings, and then asked Amazon Mechanical Turk (AMT) workers to assess each dialog across a set of metrics. However, users only interacted with one system for each dialog and thus only saw a single response at a time, while we allow workers to interact with more than one system to evaluate multiple response candidates simultaneously. Smith et al. (2022) compared per-turn, per-dialog and self-play setups for single-model and pairwise evaluations. However, it requires prior knowledge of model rankings to ideally choose the best evaluation setup, while researchers don't necessarily have that knowledge before running the evaluation. In contrast, we extend the evaluation framework to multiple models, and adopt two more novel per-turn setups, allowing users to select all appropriate responses, all without assumptions on model rankings.

Response Selection. In addition to the Select-One-Best-from-All (SOBA) setting used by (Smith et al. 2022), our work adopts a novel Select-All-That-Apply (SATA) setting, which is essentially a multi-turn response selection task for open-domain human-bot conversations. Most work in this field has either directly collected human-written positive and negative responses, or used adversarial methods to generate negative responses (Gao et al. 2020; Han et al. 2021; Deriu et al. 2022). In contrast, we allow workers to not only generate their own human-bot conversations but also annotate both positive and negative samples from a list of model-generated responses. To our knowledge, there is no published work that allows users to enter their own response during an interactive response selection procedure. By introducing a novel user input feature, the worker can steer the conversation in the direction they prefer when they don't like any model-generated responses, and make the dialog error-free and comprehensible.

Estimating Sample Size. Sample size estimation is a critical step before conducting any experiment to ensure adequate statistical power. To our knowledge, there has been little prior work that has discussed how to properly estimate sample size in the field of open-domain dialog evaluation. Smith et al. (2022) used a two-sided binomial test and two-sided independent t-test for pairwise evaluations and single-model evaluations respectively, and collected dialogs until statistically significant results are reached. However, this approach doesn't necessarily guarantee enough statistical power, as statistical power should be determined when designing an experiment rather than after observing experiment results (Hoenig and Heisey 2001). In this work, inspired by practices in clinical research (Friedman et al. 2015), we perform a variety of sample size estimations before running most experiments depending on the specific evaluation setting, and then examine sample independence using evaluation results. This methodology can not only generalize to any model pairs of interest without prior knowledge on rankings, but also promotes the repeatability of experiments by ensuring a good statistical power, which is particularly desirable for interactive evaluation work.

Methods

Models

In this work, we use GPT2-based (Radford et al. 2019) models with a variety of sizes and fine-tuning data to test which evaluation techniques work best in different scenarios. For each model, we fine-tune both the Language Modeling Head and Multiple Choice Head of GPT2 in a Transfer-Transfo fashion (Wolf et al. 2019). The Language Modeling Head takes in the dialog history and learns to predict the follow up response by minimizing the cross-entropy loss. The Multiple Choice Head is fine-tuned to select the ground-truth response amongst five candidates where four are randomly selected negative candidates. During inference we use nucleus sampling to generate the response. We leverage the HuggingFace's transformers library for all our models.[2] We use these four models with detailed descriptions in Appendix:

- GPT2-XL/GPT2-M fine-tuned on Blended Skill Talk (BST) Dataset (Smith et al. 2020);
- GPT2-XL fine-tuned on Topical Chat (TCS) Dataset (Gopalakrishnan et al. 2019);
- GPT2-XL fine-tuned on Wizard-of-Wikipedia (WoW) Dataset (Dinan et al. 2018).

Evaluation Mechanisms

Interactive Conversational Setup To enable interactive evaluation of human-bot conversations, we build an *Interactive Evaluation Service* to compare responses from one or more response generators in a multi-turn interaction with a user. Through a user interface on Amazon Mechanic Turk (AMT), workers can have live conversations with one or more models depending on the setup.

Pairwise-Model Evaluation (PM) As shown in Figure 1, at each turn, AMT workers are shown responses generated from two models, and then are asked to follow one of the settings below throughout the conversation:

1. Select One Best from All (SOBA): Workers always select a preferred response even when neither is good. The conversation continues using their selected response.
2. Select All That Apply with Random fallback (SATA-Random): Workers select 0 to 2 responses that they think are appropriate. When both or none are selected, the conversation continues with a randomly selected response.
3. Select All That Apply with User Input fallback (SATA-User): Same as SATA-Random, workers select 0 to 2 responses based on the appropriateness of responses. When both are selected, the conversation still continues with a randomly selected response. But when none is selected, the worker needs to write a better response that will be used to continue the conversation.

To facilitate comparisons across different PM per-turn settings, GPT2XL-BST is set as the baseline model and paired with the other models. However, we do not assume any model rankings and solely use evaluation results to understand the magnitude of model performance difference as well as attributes that may influence their performance. Two kinds of comparisons include:

- Size comparison: Comparing GPT2XL-BST versus GPT2M-BST, which are both fine-tuned on BST data but differ in model size.
- Fine-tuning dataset comparison: Comparing two model pairs with the same size but fine-tuned on different datasets: (1) GPT2XL-BST versus GPT2XL-TCS; (2) GPT2XL-BST versus GPT2XL-WoW.

It is worth mentioning that although we only use the general response appropriateness, these per-turn settings can be easily extended to more evaluation dimensions of research interest, such as persona, empathy, knowledgeable.

Multi-Model Evaluation (MM) As a generalized form of PM evaluation, all three per-turn settings used in PM are also adopted for Multi-Model evaluations. We perform a 4-way comparison using all four GPT2-based models.

Single-Model Evaluation (SM) Workers need to evaluate if the single provided response is appropriate. Either way, the conversation continues with that response. Users are not allowed to provide better system responses in this setup. This can serve as an independent performance baseline.

[2] https://github.com/huggingface/transformers

Metrics

We use Win-Rate to measure the relative model performance difference, which is the observed proportion of one model being selected among all the samples in an evaluation, i.e.,

$$WR(A) = \frac{X_A}{N} \quad (1)$$

where X_A is the number of times model A's response is selected among a total number of N turns. Hence, this metric is straight-forward and directly shows how often model A's response is appropriate.

Sample Size for Different Evaluations

An adequate sample size is critical to draw any statistically convincing conclusions with reasonable costs, which should also apply to the work of interactive evaluation on open-domain dialog systems. However, to our best knowledge, there is little discussion in this field on how to effectively estimate sample size before running experiments. Rather than continuing to collect more samples until a pre-assumed statistically significant result is reached, we propose a methodology to determine the required sample size before actually performing the experiment, bringing in two-fold benefits of ensuring a good statistical power and controlling evaluation costs. Although the quality of a dialog cannot be adequately captured by the sums of its turns, we assume that in a multi-turn interaction, first-party users have a full and unbiased knowledge of previous turns as context and thus can make a sensible judgement of good responses in the next turn. Therefore in this paper, following prior per-turn evaluation work (Adiwardana et al. 2020; Smith et al. 2022), each turn is considered as an individual and independent sample based on a partial dialog. Potential cascading effects between two consecutive turns and anchoring effects among multiple responses in one turn are later assessed in the results to examine sample independence on a per-turn level.

In this work, all sample sizes across different evaluations are estimated at a 95% confidence level with a 80% power (i.e., the probability of *correctly* rejecting the null hypothesis that two models perform equally well is 80% in the experiment). The motivation behind this is to make sure that all interactive evaluation setups are statistically comparable. The effect size is set to 0.1 for PM and MM evaluations, which we consider is the minimum meaningful difference in win-rates for any model pair in the experiments. In other words, we estimate the minimum number of turns required, for a win-rate difference of 0.1 between any model pairs to be statistically significant for 80% of the time if that model pair is truly different. Table 1 shows the estimated sample size for different settings. We only describe which tests are used below and leave more details to Appendix.

PM sample size Two-sided binomial test is used for the SOBA setting (Joseph L. Fleiss 2003; Friedman et al. 2015), and McNemar's test is used for the matched-pairs in SATA settings. We choose those tests mainly because of the analogy between human evaluations and certain clinical setups for which the tests are designed, e.g., in SATA setups, when a user evaluates two or more response candidates in one turn, the evaluations performed by the same user are not independent but actually correlated, suggesting that we need to treat this turn as a paired sample and use McNemar's test rather than two-sample t-test (McNemar 1947; Agresti 2007).

Mechanism	Setting	# Sample Size
Pairwise	SOBA	196
	SATA	667
Multiple (4-model)	SOBA SATA	430 445
Single	SATA	126

Table 1: Estimations of the required number of turns for different setups when the minimum win-rate difference is 10% at a 95% confidence level (two-tailed) with a 80% power. Each sample is equivalent to one turn.

Mechanism	# Dial.	# Turns	Avg. # Turns
Pairwise	668	6,357	9.5
Multiple	260	2,522	9.7
Single	109	1,164	10.7
Total	1,037	10,043	9.7

Table 2: Overall dialog statistics after data cleaning.

MM sample size Pearson's chi-squared test is used for the SOBA setting, and Cochran's Q test is used for the SATA settings to detect any pairwise differences in a multiple comparison (Cochran 1950; Joseph L. Fleiss 2003). The motivation is that MM-SOBA is a multinomial distribution where we view each model as a category and expect to see equal counts if all models have equal performance, so Pearson's chi-square test can be used in this case. Also, MM-SOBA is the only setting that requires a small pilot (e.g., 150-200 turns for four models) as prototype data to help estimate with enough power, because Pearson's chi-squared test is non-parametric and thus the distribution cannot be approximated without prior information. On the other hand, MM-SATA is an extension of PM-SATA and tests for multiple matched-pairs, so Cochran's Q test is more appropriate.

SM sample size Similar to PM, two-sided binomial test is used for SATA with a single response setting. The only difference is that an empirical win-rate of 0.8 rather than 0.5 is used as the null hypothesis. That is, we expect these models to perform well for at least 80% of the time.

Results

Data Cleaning

Several qualification checks are used for AMT workers including locations, at least 500 approved tasks, and an approval rate above 95%. For the same model pair/group, each worker is restricted to one conversation per setting. In each conversation, the worker needs to chat with the system for at least 10 turns to generate more diverse and in-depth interactions. In total, 640 paid workers have worked on our tasks with an average of 2.3 completed conversations per worker and a maximum of 18 conversations, ensuring diversity in the worker group, and 23 of them are rejected by initial worker-level quality screening. Moreover, all three per-turn evaluations are always launched simultaneously to reduce measurement errors caused by annotation quality shift over time. Although inter-annotator agreement doesn't apply to our interactive setups, we ask another small group of expert annotators to perform third-party annotation on a random sample of 60 conversations collected in Pairwise-Model Select-All-That-Apply (PM-SATA) setups, and find that the agreement between such third-party post evaluation and the conversation participants' own evaluation is moderate between 0.44-0.51. Details are included in Appendix.

We apply both dialog-level and turn-level filtering to all collected conversations. Those dialogs where workers have a fixed selection pattern in all turns (e.g., always selecting the first response) except in Single-Model evaluations, or where the average length of their utterances is less than 2 tokens, are filtered out entirely. All first turns are also removed from the results, as they usually consist of user greetings like "hi/hello" with limited response variation. After filtering 26% of dialogs and 32% of turns (including all first turns which contribute to about 10%), we still have 1,037 dialogs and 10,043 turns in total. Table 2 shows the dialog statistics for each setting. The unit cost for one dialog depends on the average annotation time, ranging from $0.6 to $1 per dialog as detailed in Appendix.

Pairwise-Model Evaluations

As shown in Table 3, we find that GPT2XL-BST performs significantly better than GPT2XL-TCS and GPT2M-BST in all three settings (i.e., SOBA, SATA-Random, SATA-User), with a noticeable win-rate advantage of 12%-24% and 12%-22% respectively, while performing more similarly to GPT2XL-WoW, with a smaller win-rate gap ranging from 6% to 24%. Here, given the observed results, we mainly focus on comparing the sensitivity and consistency of different per-turn settings, and leave more discussions on understanding model rankings as well as sample independence across all PM, MM and SM evaluations to the end of this section.

A. Comparing sensitivity between SOBA and SATA

For the same model pair, a setting with high sensitivity can test the existence of performance difference with fewer samples. We find that SOBA only requires 30% of the sample size needed for SATA, but still achieves significant results for GPT2XL-BST versus GPT2XL-TCS and GPT2M-BST. Win-rate differences between those two model pairs are also larger in SOBA than SATA. However, one possible drawback is that SOBA can introduce more false positives when two model responses are equally good or bad, as a user always needs to choose one to continue the conversation. This can lead to deviation or even exaggeration from true performance difference between a pair of models, especially when the pair of interest performs more similarly (e.g., GPT2XL-BST versus GPT2XL-WoW).

B. Comparing consistency between SOBA and SATA

Comparing all four model pairs, we find moderate to high consistency between SOBA and SATA settings. Specifically, each setting is capable of capturing a significant win-rate difference (at least 10%) between GPT2XL-BST versus GPT2XL-TCS and GPT2M-BST, suggesting high statistical confidence through cross-validation. For less distinguished pairs (e.g., GPT2XL-BST versus GPT2XL-WoW), the consistency decreases since not all settings show the same test results. However, such inconsistency doesn't necessarily mean that the model pair is equivalent. This may suggest that the model pair should be tested using a smaller win-rate difference (e.g., using 5% instead of 10%), or a higher statistical power should be used (e.g., using 90% power instead of 80%). The sample size then needs to be re-estimated based on this new combination of win-rate difference and statistical power as detailed in Appendix.

C. Additional benefits of ties

Our SATA settings make ties possible on a turn-level evaluation. One obvious benefit is that ties directly measure how frequently both model responses meet or fail to meet user's expectations. Apart from win-rates, more tie wins or fewer tie losses can also suggest a similarly good pair of models, e.g., GPT2XL-BST versus GPT2XL-WoW. Another benefit is that tie losses help capture failures on the system side, which can inform future model development. Moreover, the SATA-User setting allows users to steer the conversation towards a desired direction, even though such cases do not happen very often. More details are included in a separate section directly comparing SATA-Random and SATA-User.

Multi-Model Evaluations

As presented in Tables 4 and 5, we have three key findings.

A. Win-rate decreases with more models added

Compared with PM evaluations where win-rates are roughly centered at 50%, win-rates for each individual model across all three MM settings now shrink to about 19%-37% roughly centered at 25%. This suggests that workers are more selective when presented with more responses and hence are less likely to select all four responses. One possible explanation is that learning effects not only exist in multi-turn interactions (Xu et al. 2021) but also exists in multi-response selection, making workers more adept to form a fairer expectation of response quality.

B. Consistency between MM-SOBA and MM-SATA

Despite smaller win-rate values, we find that all three settings show highly consistent results that significantly reject the null hypothesis where all four models have the same win-rates. This suggests that at least one pair of models have statistically different win-rates. While MM-SOBA only tests for overall equality of all model win-rates, a further examination can be performed on MM-SATA's data using McNemar's test to efficiently identify which pair(s) are different.

C. Additional benefits of MM-SATA

For two MM-SATA settings, McNemar's test is used to detect any significant win-rate difference for each model pair (see Table 5). In both settings, two out of six pairs of models are tested different, i.e., GPT2XL-BST's win-rate is significantly higher than GPT2XL-WoW's and GPT2M-BST's win-rates. The results not only are consistent with our findings from PM evaluations, but also bring in additional benefits by comparing three more pairs that we have not examined in PM, e.g., GPT2XL-TCS versus GPT2XL-WoW. However, despite significance, we also observe shrinking win-rate gaps with more models added.

Model	Setting	# Dialogs	# Turns	Win-rate		Tie-win	Tie-loss
				Baseline	Model		
GPT2XL-TCS	SOBA	20	199	**62%***	38%	-	-
	SATA-Random	70	667	**58%***	48%	10%	5%
	SATA-User	66	667	**58%***	46%	7%	3%
GPT2XL-WoW	SOBA	21	198	**62%***	38%	-	-
	SATA-Random	70	671	**57%**	51%	14%	6%
	SATA-User	71	671	**54%**	48%	6%	4%
GPT2M-BST	SOBA	21	197	**61%***	39%	-	-
	SATA-Random	72	671	**57%***	45%	8%	6%
	SATA-User	71	668	**57%***	45%	11%	9%

Table 3: Pairwise-model evaluation: Win-rates of GPT2XL-BST (baseline) vs. other models, for all per-turn evaluation settings. Win-rates marked with asterisk (*) are statistically significant on a 95% level of confidence with a 80% statistical power.

Setting	# Dialogs	# Turns	Win-rate			
			GPT2XL-BST	GPT2XL-TCS	GPT2XL-WoW	GPT2M-BST
SOBA	46	436	**28%**	**19%**	27%	25%
SATA-Random	94	896	**37%**	**29%**	33%	30%
SATA-User	91	900	**32%**	**25%**	29%	27%

Table 4: Multi-model evaluation: 4-way comparison with all GPT2-based models. All three 4-way SOBA results are statistically significant using Pearson's Chi-squared test (one-tailed) or Cochran's Q test (one-tailed). These results suggest that there is at least one pair of model that is significantly different in terms of win-rates.

Single-Model Evaluations

Results for Single-Model (SM) evaluations are presented in Table 6. The model win-rate is between 85%-94% when only one single response is provided, and users cannot enter their own responses. 95% confidence intervals are also presented for comparison. Contradictory to our previous findings, GPT2M-BST turns out to be the top performing model with a noticeable advantage of 7%-11% over all other models. This is likely because of lack of appropriate dialog-level filtering for SM evaluations. When a worker always selects that single response provided in each turn, it is hard to separate the possibility of model actually performing well from annotation noise. Therefore, due to limited sensitivity and consistency, SM single-response evaluation fails to serve as a good baseline for measuring absolute model performance when no other models are compared together.

Discussion

In this work, we focus on comparing different per-turn evaluation settings for two or more models based on their sensitivity and consistency. We find that PM and MM evaluations generate very consistent results for model rankings with adequate sensitivity, while SM evaluation fails to serve as a baseline with the least sensitivity and consistency.

Explanability of model rankings We use four GPT2-based models that differ in sizes and fine-tuning data to compare different settings. With no assumptions on model rankings, we focus on using the evaluation results given a minimum performance difference of interest to understand why certain model pairs behave differently.

- Size comparison: GPT2XL-BST (with 1.5 billion parameters) performs better than GPT2M-BST (with 345 million parameters) in both PM and MM evaluations. It is well known that larger pre-trained models generally have better performance (Brown et al. 2020).
- Fine-tuning dataset comparison: GPT2XL-BST performs better than GPT2XL-TCS, while not behaving significantly different from GPT2XL-WoW. GPT2XL-TCS and GPT2XL-WoW are both fine-tuned on open-domain knowledge-grounded conversations, but since WoW was used in BST data collection and thus bears more similarity to BST than TCS.

Comparing two SATA settings As shown in Tables 3, 4 and 5, we see that SATA-Random and SATA-User generate highly consistent and equally sensitive results for all PM and MM evaluations. The sole difference between SATA-Random and SATA-User is their fallback strategy. Specifically, when all provided system responses fail to meet user's expectation, SATA-User asks user to write a better response for the system side to steer the conversation into a desired direction, while SATA-Random still continues the conversation by randomly selecting one response out of the list that are actually considered as negative cases by users. We manually checked a sample of 33 user-written responses, and found that about 73% of those responses were at least as good as system responses and about 55% of responses were better than system responses. However, with the SATA-User setting, we also notice that different from organic users recruited through advertising (Mehri et al. 2022), paid workers sometimes can produce false positives when trying to avoid spending more time writing a response.

Examining turn-level sample independence When turn-level samples are collected in a multi-turn interaction, it is necessary to examine potential cascading effects where one selected model may gain advantage over other unselected model(s) in the next turn. For PM and MM evaluations, we

	SATA-Random			SATA-User		
Comparison	$\|p_i - p_j\|$	Q/χ^2	p**-value**	$\|p_i - p_j\|$	Q/χ^2	p**-value**
All 4-Way	-	14.77	0.002	-	9.52	0.023
GPT2XL-BST vs. GPT2XL-TCS	**0.079***	11.69	0.000	**0.069***	8.42	0.003
GPT2XL-BST vs. GPT2XL-WoW	0.041	3.06	0.080	0.029	1.42	0.233
GPT2XL-BST vs. GPT2M-BST	**0.072***	9.45	0.002	**0.050***	4.37	0.036
GPT2XL-TCS vs. GPT2XL-WoW	0.037	2.86	0.090	0.040	3.07	0.079
GPT2XL-TCS vs. GPT2M-BST	0.006	0.09	0.763	0.019	0.69	0.402
GPT2XL-WoW vs. GPT2M-BST	0.031	1.80	0.178	0.021	0.81	0.366

Table 5: Multi-Model pairwise comparison: a comprehensive examination of all pairs of models using McNemar's test. $\|p_i - p_j\|$ is the absolute difference between the pair of model win-rates. There are 2 pairs marked with asterisk (*) that are statistically significant with a p-value ≤ 0.05 in both SATA-Random and SATA-User.

Model	# Dial./# Turns	Win-rate	95% CI
GPT2XL-BST	12/128	87%	[0.81, 0.93]
GPT2XL-TCS	11/126	84%	[0.78, 0.91]
GPT2XL-WoW	12/126	83%	[0.76, 0.89]
GPT2M-BST	13/127	94%	[0.91, 0.98]

Table 6: Single-model single-response evaluation: Win-rates of all models. workers can select the provided response or none if they think the response is not appropriate.

calculate the probability of each model getting re-selected in two consecutive turns. Compared with Table 3 and 4, we do not see any consistently enlarged win-rate gaps among different models in either PM or MM evaluations. This may suggest that workers do not favor certain types of model responses in two consecutive turns, or that models are capable of adapting to the previous turn so that earlier advantages do not accumulate. Either way, since we do not observe significant cascading effects, we conclude that these turn-level samples can be empirically considered as independent samples in a multi-turn dialog. One interesting finding is that although not often seen, ties do show some cascading effects, suggesting that ties are likely clustered within dialogs created by a subset of workers.

Examining anchoring effects We also examine the existence of anchoring effects in PM and MM evaluations where workers may be systematically biased towards the first response they see in a list of candidates even though no absolute ratings are required. Although candidates are shown in a random order, We still find that for all turns collected in PM evaluations, the first shown response has on average a selected rate of 55%, higher than the other response with an average selected rate of 49%. For MM evaluations where four response candidates are shown together, the first two responses also have a higher average selected rate of 34%-35%, while the last two responses only have an average rate of 28%. These suggest that anchoring effects can contribute to about 6% of win-rate difference in both PM and MM evaluations, and hence a larger minimum win-rate gap of interest should be used to test model rankings with more confidence.

Repeatability of experiments As a key contribution of this work, we provide a methodology to promote repeatability of human evaluations by ensuring a good statistical power with sample size estimation. On a model-pair level, we test their performance difference in three settings (i.e., SOBA, SATA-Random, SATA-User) of the same statistical power (i.e., 80%), which can be viewed as repetitive measures of the same model pair/group. Those highly consistent results show satisfying repeatability. On a setting level, we also test each setting more rigorously by repeating the exact same experiment three times for GPT2XL-BST versus GPT2XL-TCS as one of the distinguished model pairs that we have observed. All three repetitions generate significant results at a 95% confidence level, suggesting high repeatability for all three per-turn settings.

Evaluation costs When choosing the best evaluation setup, there is a trade-off between efficacy and costs. In our example of four models, we need to perform 3-6 PM evaluations to exhaust all possible pairings, while only one MM evaluation is required to test all pairs altogether if using SATA. Translated to turn-level sample size, the number of turns required in MM-SATA will be between the best and worst scenarios of PM-SOBA but smaller than PM-SATA.

Conclusion and Future Work

In this work, we extend the interactive evaluation settings to multiple models with a focus on per-turn evaluation techniques, and show that two novel Select-All-That-Apply settings work well with additional benefits from allowing ties and user-written responses. Besides, we propose a methodology to estimate required sample size given a minimum performance gap, which promotes repeatability, helps control costs, and does not require prior knowledge on rankings and hence will work for any pair of models. A thorough analysis comparing Single-Model, Pairwise-Model, and Multi-Model evaluations is also conducted based on sensitivity and consistency of different settings to help choose the best evaluation setup for more research scenarios.

While our work has taken a step forward towards credible human evaluations for open-domain dialog systems, it is worth noting that per-turn evaluations alone cannot adequately evaluate the whole conversation, where per-dialog or self-play evaluations or a mix of different techniques should be further investigated in future work.

Acknowledgements

We would like to thank Karthik Gopalakrishnan for his initial passion and ideas with this work, and Seokhwan Kim and Shikib Mehri for their inspiring feedback. We would also like to especially thank Prof. Joel Greenhouse from Department of Statistics & Data Science, Carnegie Mellon University, for his feedback and suggestions over the statistical methodology covered in this project.

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; and Le, Q. V. 2020. Towards a Human-like Open-Domain Chatbot. arXiv:2001.09977.

Agresti, A. 2007. *An Introduction to Categorical Data Analysis, Second Edition*. New York: Wiley. ISBN 978-0-471-22618-5.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cochran, W. G. 1950. The Comparison of Percentages in Matched Samples. *Biometrika*, 37(3/4): 256–266.

Deriu, J.; Tuggener, D.; von Däniken, P.; Campos, J. A.; Rodrigo, Á.; Belkacem, T.; Etxabe, A. S.; Agirre, E.; and Cieliebak, M. 2020. Spot the Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems. In *EMNLP*.

Deriu, J.; Tuggener, D.; von Däniken, P.; and Cieliebak, M. 2022. Probing the Robustness of Trained Metrics for Conversational Dialogue Systems. arXiv:2202.13887.

Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A. H.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A. D.; Serban, I.; Lowe, R.; Prabhumoye, S.; Black, A. W.; Rudnicky, A. I.; Williams, J.; Pineau, J.; Burtsev, M. S.; and Weston, J. 2019. The Second Conversational Intelligence Challenge (ConvAI2). *ArXiv*, abs/1902.00098.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Finch, S. E.; and Choi, J. D. 2020. Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols. arXiv:2006.06110.

Friedman, L.; Furberg, C.; DeMets, D.; Reboussin, D.; and Granger, C. 2015. *Chapter 8. Sample Size*. Springer.

Gao, X.; Zhang, Y.; Galley, M.; Brockett, C.; and Dolan, B. 2020. Dialogue Response Ranking Training with Large-Scale Human Feedback Data. arXiv:2009.06978.

Ghandeharioun, A.; Shen, J. H.; Jaques, N.; Ferguson, C.; Jones, N.; Lapedriza, A.; and Picard, R. 2019. Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems. arXiv:1906.09308.

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; Hakkani-Tür, D.; and AI, A. A. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *INTERSPEECH*, 1891–1895.

Han, K.; Lee, S.; Lee, W.; Lee, J.; and hun Lee, D. 2021. An Evaluation Dataset and Strategy for Building Robust Multi-turn Response Selection Model. arXiv:2109.04834.

Higashinaka, R.; Funakoshi, K.; Inaba, M.; Tsunomori, Y.; Takahashi, T.; and Akama, R. 2021. *Dialogue System Live Competition: Identifying Problems with Dialogue Systems Through Live Event*, 185–199. Lecture Notes in Electrical Engineering. Germany: Springer Science and Business Media Deutschland GmbH. Publisher Copyright: © 2020, The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd.

Hoenig, J. M.; and Heisey, D. M. 2001. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, 55(1): 19–24.

Ji, T.; Graham, Y.; Jones, G. J. F.; Lyu, C.; and Liu, Q. 2022. Achieving Reliable Human Assessment of Open-Domain Dialogue Systems. arXiv:2203.05899.

Joseph L. Fleiss, M. C. P., Bruce Levin. 2003. *Statistical Methods for Rates and Proportions, Third Edition*. Wiley. ISBN 9780471526292.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 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*, 2122–2132. Austin, Texas: Association for Computational Linguistics.

Lowe, R.; Noseworthy, M.; Serban, I. V.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2018. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. arXiv:1708.07149.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2): 153–157.

Mehri, S.; and Eskenazi, M. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 681–707. Online: Association for Computational Linguistics.

Mehri, S.; Feng, Y.; Gordon, C.; Alavi, S. H.; Traum, D.; and Eskenazi, M. 2022. Interactive Evaluation of Dialog Track at DSTC9. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5731–5738. Marseille, France: European Language Resources Association.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Ram, A.; Prasad, R.; Khatri, C.; Venkatesh, A.; Gabriel, R.; Liu, Q.; Nunn, J.; Hedayatnia, B.; Cheng, M.; Nagar, A.; King, E.; Bland, K.; Wartick, A.; Pan, Y.; Song, H.; Jayadevan, S.; Hwang, G.; and Pettigrue, A. 2018. Conversational AI: The Science Behind the Alexa Prize. arXiv:1801.03604.

Roller, S.; Boureau, Y.-L.; Weston, J.; Bordes, A.; Dinan, E.; Fan, A.; Gunning, D.; Ju, D.; Li, M.; Poff, S.; Ringshia, P.; Shuster, K.; Smith, E. M.; Szlam, A.; Urbanek, J.;

and Williamson, M. 2020. Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions. arXiv:2006.12442.

See, A.; Roller, S.; Kiela, D.; and Weston, J. 2019. What makes a good conversation? How controllable attributes affect human judgments. *Proceedings of the 2019 Conference of the North*.

Smith, E. M.; Hsu, O.; Qian, R.; Roller, S.; Boureau, Y.-L.; and Weston, J. 2022. Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents. arXiv:2201.04723.

Smith, E. M.; Williamson, M.; Shuster, K.; Weston, J.; and Boureau, Y.-L. 2020. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2021–2030.

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Xu, J.; Ju, D.; Li, M.; Boureau, Y.-L.; Weston, J.; and Dinan, E. 2021. Recipes for Safety in Open-domain Chatbots. arXiv:2010.07079.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 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)*, 2204–2213. Melbourne, Australia: Association for Computational Linguistics.