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

amazon | science

Introduction

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.

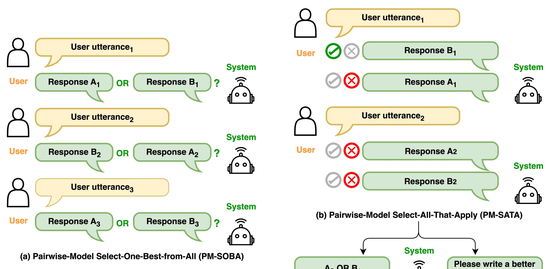

Figure 1. Illustration of Pairwise-Model evaluation.

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 those.

Models

We use GPT2-based models with a variety of sizes and fine-tuning data to test which evaluation techniques work best in different scenarios. During inference we use nucleus sampling to generate the response. The four models used here are:

- GPT2-XL/GPT2-M fine-tuned on Blended Skill Talk (BST) Dataset;
- GPT2-XL fine-tuned on Topical Chat (TCS) Dataset;
- GPT2-XL fine-tuned on Wizard-of-Wikipedia (WoW) Dataset.

Notes

To facilitate comparisons across different Pairwise-Model 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. We perform two kinds of comparisons in Pairwise-Model evaluations:

- 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.

Evaluation Mechanisms

We build a UI on Amazon Mechanic Turk (AMT) to compare responses from one or more response generators in a multi-turn interaction with a human worker. Workers are required to complete at least 10 turns in one dialog.

Pairwise-Model Evaluations (PM): At each turn, AMT workers are shown responses generated from two models, and then are asked to follow one of the settings below:

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.

Multi-Model Evaluations (MM): As a generalized form of PM evaluation, all three per-turn settings above are adopted in a 4-way comparison using all four GPT2 models.

Single-Model Evaluations (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.

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., $WR(A) = \frac{X_A}{N}$, where X_A is the number of times model A's response is selected among a total number of N turns.

Sample Size Estimation

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.

In this work, all sample sizes across different evaluations are estimated at a 95% confidence level with a 80% statistical power. 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.

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

Table 1. Estimations of the required number of turns for different setups. Each sample is equivalent to one turn.

Results

Data Cleaning: 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. After both dialog-level and turn-level filtering, we have 1,037 dialogs and 10,043 turns.

PM Evaluations: Comparing **sensitivity** between SOBA and SATA, 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. Comparing **consistency** between SOBA and SATA, we find moderate to high consistency between SOBA and SATA settings. Specifically, each setting can capture a significant win-rate difference (at least 10%) between GPT2XL-BST versus GPT2XL-TCS and GPT2M-BST, suggesting high statistical confidence through cross-validation.

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 2. 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.

MM Evaluations: 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. Comparing **consistency** between MM-SOBA and MM-SATA, despite smaller win-rate values, all three settings show highly consistent results that significantly reject the null hypothesis and suggest that at least one pair of models have statistically different win-rates. One **additional benefit** of MM-SATA is that we can further test all possible model pairs with existing data.

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 3. 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).

SM Evaluations: Due to limited sensitivity and consistency, SM evaluation fails to serve as a good baseline for measuring absolute model performance when presented alone.

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 works 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.

Conference Paper AAAI 2023
*Corresponding author: sijial@amazon.com