MERCY: Multiple Response Ranking Concurrently in Realistic Open-Domain Conversational Systems

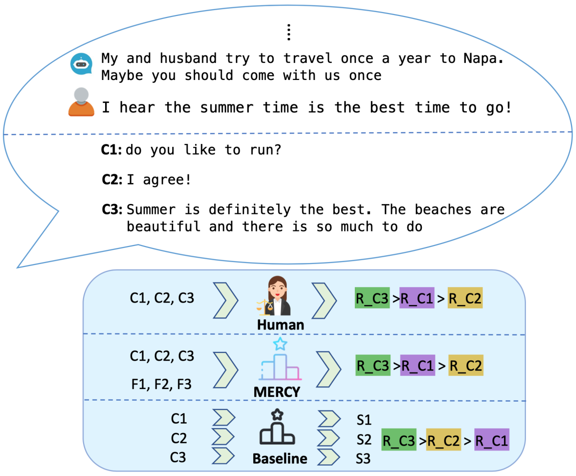
Sarik Ghazarian[1*] **Behnam Hedayatnia**[2] **Di Jin**[2] **Sijia Liu**[2]
Violet Peng[2] **Yang Liu**[2] **Dilek Hakkani-Tur**
[1] University of Southern California / Information Sciences Institute
[2] Amazon Alexa AI
`sarik@isi.edu`, `dilek@ieee.org`
`{behnam,djinamzn,sijial,pengnany,yangliud}@amazon.com`

Abstract

Automatic Evaluation (AE) and Response Selection (RS) models assign quality scores to various candidate responses and rank them in conversational setups. Prior response ranking research compares various models' performance on synthetically generated test sets. In this work, we investigate the performance of model-based reference-free AE and RS models on our constructed response ranking datasets that mirror real-case scenarios of ranking candidates during inference time. Metrics' unsatisfying performance can be interpreted as their low generalizability over more pragmatic conversational domains such as human-chatbot dialogs. To alleviate this issue we propose a novel RS model called **MERCY** that simulates human behavior in selecting the best candidate by taking into account distinct candidates *concurrently* and learns to rank them. In addition, MERCY leverages *natural language feedback* as another component to help the ranking task by explaining why each candidate response is relevant/irrelevant to the dialog context. These feedbacks are generated by prompting large language models in a few-shot setup. Our experiments show the better performance of MERCY over baselines for the response ranking task in our curated realistic datasets.

1 Introduction

Advancements of neural models (Devlin et al., 2019; Radford et al., 2019; Zhang et al., 2020b; Shuster et al., 2022) has led to the vast continuous research on open-domain dialog systems. Many deployed open-domain dialog systems rely on multiple response generators in order to address the variety of topics within a dialog. Accordingly, response ranking is introduced as a major necessity for ranking different responses based on their quality (Zhou et al., 2018; Wu et al., 2019; Liu et al., 2021).



⋮
My and husband try to travel once a year to Napa. Maybe you should come with us once
I hear the summer time is the best time to go!

C1: do you like to run?
C2: I agree!
C3: Summer is definitely the best. The beaches are beautiful and there is so much to do

Figure 1: An overview of the response ranking task. Similar to human, MERCY takes all the candidates as input to rank them (denoted as R_C1, R_C2, R_C3) more accurately, while baseline RS metrics score (denoted as S1, S2, S3) each candidate separately. Alongside candidates, MERCY also takes the candidates generated feedback (denoted as F1, F2, F3) as another input.

Response Selection (RS) models were initially trained to rank human-written positive responses higher than synthetically generated negative responses (Gu et al., 2020; Gao et al., 2020; Gupta et al., 2021). Another conforming substitution for RS models can be model-based reference-free **Automatic Evaluation (AE)** metrics that conduct response evaluation along different dimensions such as relevancy (Tao et al., 2018; Ghazarian et al., 2019), engagingness (Ghazarian et al., 2020), coherence (Ye et al., 2020), etc. and have been proven to be more suitable for open-domain conversations (Lowe et al., 2017; Eskenazi et al., 2019) in comparison to the reference-based overlap-based metrics (Papineni et al., 2002; Lin, 2004). The primary intention of AE and RS is to quantify the quality of responses. In this work, we investigate their usage for the response ranking task.

Predominantly, RS models have been tested to distinguish between positive and synthetically gen-

---

[*] Work done during an internship at Amazon

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 615–631
September 11–15, 2023. ©2023 Association for Computational Linguistics

erated negative candidates following approaches such as random matching (Gao et al., 2020; Sai et al., 2020; Gupta et al., 2021), text-level manipulations (shuffling, replacing) (Sai et al., 2020; Zhang et al., 2021a), human-written adversarial irrelevant responses (Sai et al., 2020; Gupta et al., 2021; Sato et al., 2020). These synthetically curated test sets are not sufficient representations of real-world inference time candidates that are generated by dialog models. Hedayatnia et al. (2022) demonstrated the effectiveness of training on machine-generated candidates from real user interactions over using synthetic candidates for response selection. However this data is not publicly available.

In this work, we construct the first public dataset of human fine-grained rankings for responses generated by state-of-the-art dialog models in *human-human* dialog contexts, which we denote as *Static Evaluation Setup (SES)*. For our work we also leverage the recently collected *Interactive Evaluation Setup (IES)* (Liu et al., 2023) dataset, which includes human annotations for different machine-generated responses within *human-chatbot* conversations. Our intention is to conduct a survey to evaluate the performance and generalizability of state-of-the-art model-based reference-free AE and RS on our curated datasets that are closer to deployment time ranking scenarios, where one interlocutor is human and the other is a dialogue system. We show that in these realistic test cases, existing RS and AE models exhibit low performance.

To overcome this issue and have a more reliable RS in real case scenarios we propose **MERCY**, which pursues users' behavior of taking different candidates as input and predicting their rankings by relying on their comparable representations (See Figure 1). There is strong evidence that relying on comparable representations is useful such as human preference modeling to improve the performance of LLMs (Bai et al., 2022). MERCY also augments the input with feedback in the form of natural language that explains why or why not a response is relevant. Gupta et al. (2022) introduced an instruction-tuned large language model (LLM) to perform a variety of dialog tasks such as determining if a response is relevant or not. We follow a similar approach of prompting a LLM to evaluate a response; however, we prompt the model to generate more detailed information by not only asking *if* a response is relevant but also *why* it is relevant. We refer to this generated output as feedback. We leverage BLOOMZ-175B (Muennighoff et al., 2022) to generate each candidate's feedback. We train MERCY on the train split of SES and demonstrate that it is more accurate in real-case ranking scenarios in comparison to the best performing automatic metric finetuned on the same training set. Considering multiple candidates together and augmenting responses with feedback both contribute to MERCY's better performance.

Our contributions are summarized as follows:

- We release a new benchmark dataset for response selection, which contains human rankings for responses generated by state-of-the-art neural response generation models.[1]
- We present an in-depth analysis of the performance of AE and RS models on this benchmark dataset and report their low performance and generalizability over different dialog contexts, domains and generated responses.
- We propose a new RS method, MERCY, which receives various candidates simultaneously and takes the generated natural language feedbacks for each candidate as input and learns to rank candidates by minimizing the Kullback-Leibler divergence loss. Experiments show that MERCY outperforms all existing AE and RS metrics by a good margin.

2 Related Work

Due to the vast number of AE/RS models, an in-depth comparison of these metrics is critical. Yeh et al. (2021) performed a comprehensive survey by comparing multiple AE metrics on publicly available evaluation testsets. In this work, we perform a similar survey of model-based reference-free AE/RS models on *response selection testsets*. In contrast to their evaluation testsets where responses are annotated on a Likert scale, which can lead to annotator bias and could make it difficult for a model to predict the exact scores, we only need to evaluate the relative ordering from the predicted output of these methods.

The response selection datasets we leverage are more realistic than previously proposed synthetically generated datasets. Prior research proposed to use simple approaches such as random response selection (Han et al., 2021), corrupting utterances by inserting, substituting and deleting random tokens (Whang et al., 2021), using the mask-and-fill

---

[1] The dataset will be published upon acceptance.

approach (Gupta et al., 2021) for generating adversarial negative examples or collect human-written negative samples (Sato et al., 2020). Previous work also suggest to augment dialog datasets with synthetically generated positive samples (Mizukami et al., 2015; Khayrallah and Sedoc, 2020; Gupta et al., 2019; Sai et al., 2020; Zhang et al., 2020a).

In a study by Hedayatnia et al. (2022), they demonstrated that using a human-chatbot dataset, where responses were generated by multiple response generators and then annotated by humans for training RS (response selection) models, led to improved performance compared to models trained on synthetically generated datasets. Unfortunately, the dataset they used could not be made public due to privacy concerns, as it contained real-user dialogs. In contrast, our approach involves collecting a similar and realistic response ranking dataset, which we plan to release for future research purposes.

In RS, most models score response candidates independently without considering them together. Zhang et al. (2021b) proposed a joint matching approach that concurrently accepts *exactly four* candidates as input and *selects the only correct response* using log-likelihood as the training objective. Our RS model follows a similar training approach but can handle a *variable number of responses* for ranking. Additionally, our metric stands out from previous work as it combines generated natural language feedback with multiple response candidates, providing the model with valuable information in a natural language format.

Feedback generation has been shown to be beneficial for improving language models. (Shi et al., 2022; Xu et al., 2020; Hancock et al., 2019; Scheurer et al., 2022; Tandon et al., 2022). Shi et al. (2022); Hancock et al. (2019); Scheurer et al. (2022) focused on improving response generation models using three types of human feedback: binary, modular and natural language. While these studies use natural language feedback collected via human annotation, our work *generates feedback* from large language models in a few-shot fashion and use them for the *ranking task*. The closest work to ours is (Gupta et al., 2022), which trains an instruction-tuned large language model to conduct evaluation. However, this work treats feedback as a classification task asking *if* a response is relevant, while we prompt the model to output *why* it is relevant and accompany that with the candidate.

3 Data Sets

To conduct a comprehensive survey on AE/RS models, we look at three response ranking test sets each encompassing different properties: 1) type of dialog contexts, 2) type of candidates for ranking, and 3) type of conversational domains.

DAILYDIALOG++ Sai et al. (2020) composed a dataset consisting of manually created relevant/irrelevant responses for human-human dialog contexts taken from DailyDialog (Li et al., 2017). To create irrelevant responses, annotators were asked to write responses that share similar semantics with the dialog context yet are not acceptable. Hence, in DAILYDIALOG++ both dialog histories and candidate responses are human-written.

SES The responses from DailyDialog++ may not match realistic inference time test sets where responses are machine generated. To deal with this, we collect the Static Evaluation Setup (SES) dataset comprising of various model generated responses for contexts sampled from multiple human-human dialog datasets: DailyDialog (Li et al., 2017), BlendedSkillTalk (Smith et al., 2020), PersonaChat (Zhang et al., 2018), EmpatheticDialogues (Rashkin et al., 2019). Each dialog context contains 8 different responses generated by BlenderBot (Roller et al., 2021), GPT2-XL (Radford et al., 2019) fine-tuned on BlendedSkillTalk (Smith et al., 2020), Plato-2 (Bao et al., 2020), and Plato-XL (Bao et al., 2021) with different decoding mechanisms. The model training and decoding parameters are provided in Section E in the Appendix. We collect two sets of data where the rankings are eventuated from two groups of annotators: 1) in-house annotators familiar with the ranking task (SES_INTERNAL), 2) Amazon Mechanical Turk (AMT) workers (SES_AMT).

Responses in SES_INTERNAL are annotated by two internal annotators on the scale of 0 (not an appropriate response) to 2 (a suitable response). We calculate the normalized mean score for each response in the range of 0 to 1 and assign label 1 to the response if its normalized score is greater than 0.5 or 0 otherwise. To better analyze the performance difference of AE/RS models, we remove turns where all the candidates are 1 or 0 and call it SES_INTERNAL_FILTERED. Although RS model may face such all good or all bad candidates in real-world scenarios, we exclude them to not mislead the performance of RS with random candidate selection. Due to the higher quality of annotations by

Dataset	Num_Responses	Pos/Neg
DAILYDIALOG++	11420	5710/5710
SES_INTERNAL	8000	4601/3399
SES_INTERNAL_FILTERED	7336	4049/3287
SES_AMT	7968	5546/2422
SES_AMT_FILTERED	6488	4098/2390
IES	31849	13519/18330
IES-V2	3240	1330/1910

Table 1: Statistics of response ranking datasets.

internal annotators who are more familiar with the task, we leverage this dataset to test our proposed RS model versus baselines.

The process of collecting SES_INTERNAL is slow due to an insufficient amount of annotators, therefore we use AMT workers for faster data collection. For SES_AMT dataset, 5 AMT workers evaluate each response in the range of 1-5 indicating low-quality to high-quality responses. Here we use more fine-grained ratings which allow us to check if AMT workers understand the range of how good/bad a response can be. We get the median score of each candidate's ratings and normalize it in the range of 0-1. We assign 0/1 label similar to SES_INTERNAL. We remove turns with all good or bad responses, and call it SES_AMT_FILTERED. The statistics of these datasets are shown in Table 1. This dataset is biased toward positive samples as generations are done by state-of-the-art models resulting high quality responses, similar to what happens during real-case scenarios.

IES We take one step closer towards having a realistic response ranking test set by leveraging the dataset from (Liu et al., 2023) where at each turn in a human-chatbot dialog, AMT workers are requested to select all valid responses from multiple machine-generated candidates. The generative models are four GPT2-XL (Radford et al., 2019) models, fine-tuned on BlendedSkillTalk (Smith et al., 2020), TopicalChat (Gopalakrishnan et al., 2019), and WOW(Dinan et al., 2019) datasets, respectively. The model training parameters are described in Section D in the Appendix. We denote this dataset as Interactive Evaluation Setup (IES). Although IES represents a more realistic dataset, its collection process is time consuming as the user has to both converse with the system and annotate each turn for quality. In contrast only one turn needs to be annotated in SES allowing for faster data collection.

A closer look at the IES data shows that some good responses were not marked correctly by AMT workers. This could be because AMT workers may be taking into account factors besides relevancy when selecting a response such as engagingess. Examples of these issues can be seen in Section C in the Appendix. In order to have a more fair comparison we sample 80 dialogs from IES and ask AMT workers to reannotate each response on a scale of [1-5] similar to SES_AMT setup, and denote this dataset as IES-v2. IES-v2 includes a part of IES dataset with more fine-grained annotations in a 1-5 scale that allows better training signals for the RS model. We compute the Fleiss kappa for inter-annotator agreement and get a score of 0.41, which indicates moderate agreement. In this work, we use IES/IES-v2 data for only testing.

4 Analysis of AE/RS Methods for Response Ranking

4.1 AE/RS Methods

Inspired by the survey of automatic metrics on evaluation test sets (Yeh et al., 2021), we compare different AE/RS models on response ranking testsets. We compare AE metrics such as: Ruber (Tao et al., 2018), Bert_Ruber(Ghazarian et al., 2019), Pone(Lan et al., 2020), USR(Mehri and Eskenazi, 2020b), FED(Mehri and Eskenazi, 2020a), FlowScore(Li et al., 2021), Maude(Sinha et al., 2020), Grade(Ye et al., 2020), DynaEval(Zhang et al., 2021a), Predictive_Engagement(Ghazarian et al., 2020), USL(Phy et al., 2020), HolisticEval(Pang et al., 2020), MDD(Zhang et al., 2022), DEAM(Ghazarian et al., 2022). For RS models, we use BM25(Robertson et al., 2009), Dialogrpt(Gao et al., 2020), SABert_KeySem(Gupta et al., 2021).

Bert_Ruber (Ghazarian et al., 2019), Pone (Lan et al., 2020), Maude (Sinha et al., 2020) and DEB (Sai et al., 2020) are classifiers used to predict the relevancy of a response, while Predictive_Engagement (Ghazarian et al., 2020) affirms the positive impact of incorporating an engagement classifier on top of response relevance. FlowScore (Li et al., 2021), Deam (Ghazarian et al., 2022) and DynaEval (Zhang et al., 2021a) evaluate the overall dialog and the connection between utterances. A few AE metrics, such as USL-H (Phy et al., 2020), HolisticEval (Pang et al., 2020), USR (Mehri and Eskenazi, 2020b), FED (Mehri and Eskenazi, 2020a) take into account multiple sub-metrics to achieve a more reliable evaluation metric. Finally, MDD (Zhang et al., 2022) looks for a robust metric that has acceptable performance over multiple domains.

For RS models, BM25 (Robertson et al., 2009) ranks candidates based on their keyword similarities to the context. DialogRPT (Gao et al., 2020) uses human feedback data from Reddit and determines whether a response is human-like to rank the generated candidates. SABert_KeySem (Gupta et al., 2021) is a Speaker-Aware Bert-based (Gu et al., 2020) classifier finetuned on adversarial responses created via mask-and-fill and keyword-based generations.[2]

In addition to aforementioned AE/RS methods, we add random and naive baselines to achieve an exhaustive study. Random baseline randomly assigns scores to responses in the range of 0 to 1. We report the mean aggregation of random baseline performance after 5 runs. Naive baseline reports the best generative model's performance by selecting all its responses as appropriate and the rest candidates as not suitable.

4.2 Ranking Metrics

We report common metrics for response selection:

Hits@K shows the rate of correct responses (selected by human) appearing in the top-k responses scored by each metric. *In our experiments, K is a variable since each turn of evaluation can have different number of human selected responses.* Thus, we report the mean of Hits@K from different evaluation turns.

Recall@1 computes the number of evaluation turns where the highest scored candidate by the metric is also selected by human.

MRR computes the mean of all reciprocal ranks for human-selected responses. Reciprocal rank for each true response shows its rank in the metric's ordered output list. MRR demonstrates the ability of the metric to assign better scores (higher rankings) to human-selected responses.

4.3 Results

We show the performance and generalizability of AE/RS models on the IES and DAILYDIALOG++ datasets in Table 2 and SES datasets in Table 3 by using them without finetuning on the datasets.

For IES we see the best performing AE metric is DEB. This shows the positive impact of pretraining on a large conversational dataset (Reddit) for evaluation. The best performing RS model is SABert_KeySem, which is due to the positive effect of its semantic-based perturbations to generate higher quality negative samples. The slight performance difference between the best performing AE/RS models and baselines on the IES dataset shows the low generalizability of these methods. The main distinctions between IES and the training datasets of the AE/RS models are: 1) differing conversational domains, 2) responses generated by state-of-the-art dialog systems in IES versus human-written or heuristically generated candidates, 3) human-bot dialog contexts in IES versus human-human interactions.

For the SES testsets we also see DEB and SABert_KeyKem are among the best performing AE/RS models, respectively. We see a much higher score from these models on SES in comparison to IES. This may be due to the closeness of dialog history type, which is human-written in SES, to the training datasets of these metrics.

For DAILYDIALOG++, DEB and MDD achieve the best performance. One reason is the domain overlap between the test and train data as both are from DailyDialog. The high performance of these metrics on DAILYDIALOG++ in comparison to SES and IES further shows the low generalizability of AE/RS metrics on different dialog contexts/domains. We don't report the Naive baseline since the candidates are not from different models.

5 Method

Our proposed response selection model MERCY evaluates multiple response candidates for a given dialog context *simultaneously*, and also leverages the feedback generated by LLMs for candidates.

5.1 Few-Shot Feedback Generation

We look into leveraging LLMs for feedback generation via prompting. Specifically, we use the BLOOMZ-175B model (Muennighoff et al., 2022), which is finetuned to follow human instructions for various NLP tasks. To prompt the model for response evaluation, we take three conversations from the FED testset (Mehri and Eskenazi, 2020a), add the question "How relevant are the bot responses?" along with a brief explanation of relevance (or lack thereof) for each response. The exact prompt is available in Section F in the Appendix. Using this prompt, we input the conversations from SES and IES-v2 to obtain feedback for each response in the dataset. Table 4 has an example of our generated feedback showing the model's ability to predict relevance and offer reasoning. More examples can be found in Section G the Appendix.

---

[2] More details about AE/RS metrics are discussed in Section A of the Appendix.

		DailyDialog++			IES		
Metric	**Type**	**Hits@K**	**MRR**	**Recall@1**	**Hits@K**	**MRR**	**Recall@1**
Naive	basline	-	-	-	50.04	72.40	49.64
Random	basline	50.25	70.12	49.4	48.10	71.40	48.15
Bert_Ruber	AE	55.99	74.41	57.44	47.51	71.00	47.43
PONE	AE	48.14	65.06	43.61	47.59	71.00	47.47
USR	AE	54.69	75.68	59.28	46.57	70.37	46.44
FED	AE	61.28	86.00	75.92	50.88	73.00	50.76
FlowScore	AE	26.01	42.37	37.04	48.06	71.23	48.00
Maude	AE	62.31	84.28	71.8	50.86	72.89	50.77
Grade	AE	69.72	89.74	82.14	46.22	70.24	46.11
DynaEval	AE	92.7	98.88	98.07	48.99	71.85	48.86
Predictive_Engagement	AE	45.92	59.72	35.81	46.91	70.57	46.79
USL-H	AE	60.51	66.01	44.57	47.47	70.86	47.36
HolisticEval	AE	55.43	81.12	68.39	46.9	70.61	49.08
MDD	AE	95.73	99.65	**99.74**	50.9	73.00	51.29
DEAM	AE	54.64	72.72	54.99	49.88	72.37	49.93
DEB	AE	**95.97**	**99.70**	99.39	52.12	73.62	52.11
Dialogrpt	RS	46.87	61.50	38.79	49.95	72.30	49.74
BM25	RS	40.47	63.45	44.05	46.89	70.75	46.73
SABert_KeySem	RS	89.63	99.16	98.51	**52.80**	**74.14**	**53.01**

Table 2: Performance of different AE/RS metrics on DailyDialog++ and IES

		SES_internal_filtered			SES_amt_filtered		
Metric	**Type**	**Hits@K**	**MRR**	**Recall@1**	**Hits@K**	**MRR**	**Recall@1**
Naive	baseline	53.76	74.23	59.54	63.05	78.86	64.50
Random	baseline	58.74	72.71	58.44	63.45	78.14	62.52
Bert_Ruber	AE	56.71	74.02	57.8	65.71	81.62	68.68
PONE	AE	56.06	71.16	52.78	63.60	77.15	61.28
USR	AE	56.54	74.92	58.89	66.46	81.95	69.54
FED	AE	58.69	75.94	59.77	66.54	82.13	69.67
FlowScore	AE	56.14	71.48	53.54	62.09	74.56	57.21
Maude	AE	56.45	73.75	57.25	61.96	76.57	61.05
Grade	AE	56.78	72.66	54.53	65.50	79.30	65.10
DynaEval	AE	58.95	75.96	59.76	63.53	79.72	65.72
Predictive_Engagement	AE	53.55	69.42	51.36	61.32	74.17	57.09
USL-H	AE	57.63	74.21	57.47	66.56	**82.81**	70.9
HolisticEval	AE	56.13	74.18	58.56	63.70	79.74	66.09
MDD	AE	56.62	74.78	61.61	64.78	82.40	**71.89**
DEAM	AE	55.73	74.10	58.01	63.80	81.30	68.80
DEB	AE	**60.83**	**77.56**	**63.03**	63.32	79.28	65.23
Dialogrpt	RS	53.63	69.56	51.47	63.40	76.45	60.30
BM25	RS	56.87	73.13	55.39	63.65	76.03	59.43
SABert_KeySem	RS	57.91	76.07	61.18	**67.85**	82.05	69.79

Table 3: Performance of different AE/RS metrics on SES_internal_filtered and SES_amt_filtered

User: South Padre Island is beautiful. How many boats have you made? Socialbot: about 6 i believe , i lost track after my 3rd User: Haha. Are they big boats or small boats? Socialbot: small boat, about 30 feet
Question: How relevant are the socialbot responses?
Answer: The last system response is relevant in this dialog. The socialbot responds to the user's question about boats by providing details about a boat they have made. Overall, the socialbot's responses are relevant, earning a score of 8 out of 10 for relevance.

Table 4: An example of a generated feedback

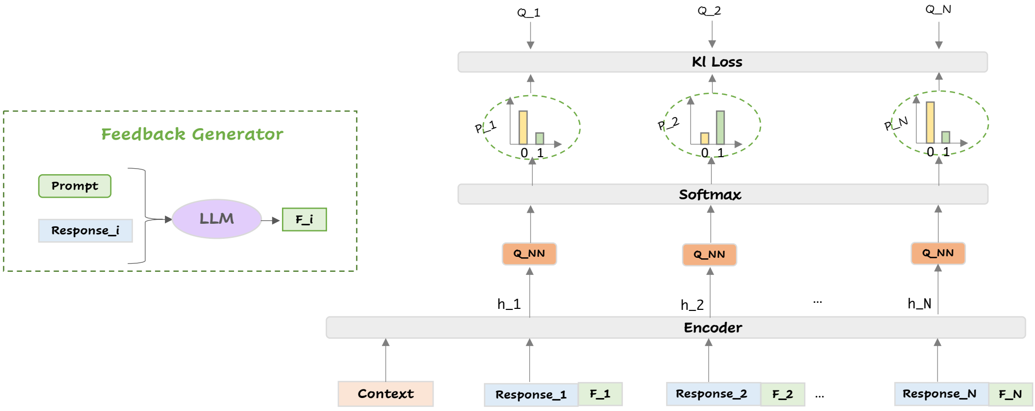


Figure 2: The overview of MERCY and the feedback generator.

To verify the quality of the generated feedback we sample 50 outputs from the SES dataset, 25 predicted by the model as relevant and the remaining as not relevant. We annotate for two dimensions: *Correctness* (Did the model correctly predict if the response was relevant?) and *Makes Sense* (Did the explained feedback make sense?) as yes/no. We find that 82% of the feedback was correct and 56% of the explanations made sense. This suggests that the feedback can be used as input into the response selector. We leave improving the quality of explanations via prompt engineering for future work.

5.2 Ranking Responses Synchronously

Users rank responses by simultaneously receiving all candidates and comparing their quality. However, most RS models consider the candidates independently and predict their scores without conducting the comparison between responses. In this work, we explore the benefits of concurrently receiving *any number* of candidates as input for *rankings*. We train MERCY by minimizing the Kullback-Leibler (KL) Divergence loss between its outputs and ground-truth labels. MERCY's predictions are scores in the range of [0-1], indicating low up to high quality candidates.

Figure 2 gives an overview of MERCY. It takes the context and concatenated candidates as input. Context is composed of all utterances from beginning up to the current turn in the dialogue, splitted with $</UTT>$ token.

$$C = U_1 </UTT> U_2 ... </UTT> U_M \quad (1)$$

Since MERCY can receive multiple candidates with various lengths, the chance of passing the maximum length that can be handled by the encoder in the metric is not negligible. Hence, we use $</UTT>$ to handle such cases by removing the minimum number of utterances from the beginning of the context until all the input can fit in the model.

Following the context C, we pass all the candidates beginning with $[RES]$ special token. In contrast to the metric proposed by Zhang et al. (2021b), which separates each candidate with special tokens to be distinguished between *constant* number of candidates, MERCY is *more generalized* and can process *any number of candidates* as input. It uses the index of each $[RES]$ token to get the corresponding candidate's encoding vector.

$$R = [RES]R_1[RES]R_2...[RES]R_N \quad (2)$$

To incorporate feedback into MERCY, we concatenate each response's feedback to itself, and separate them with a $[Feedback]$ special token.

$$R_i = R_i[Feedback]F_i \quad (3)$$

After concatenating C and R we pass the input I $(I = C \cdot R)$ through an encoder and get the output embeddings $H \in \mathbb{R}^{|I| \times d}$, where d denotes the hidden dimensional size of the encoder. The hidden representation of each candidate response is returned based on the index of the $[RES]$ token for that corresponding candidate. Similar to how humans rank responses, MERCY is seeing multiple candidates during the encoding process.

$$h\_i = H_{[RES]} \quad \text{where} \quad [RES] \in [RES]R_i \quad (4)$$

Simultaneously, each candidate's hidden representational vector is passed through a linear layer, whose parameters are denoted as W_q, which outputs a scalar value q_i for each candidate. The outputs are then sent through a Softmax layer. The KL-Divergence loss is then minimized between the normalized model outputs and the probability distribution of ground-truth labels.

		SES_INTERNAL_FILTERED			IES-V2		
Metric	**Data**	**Hits@K**	**MRR**	**Recall@1**	**Hits@K**	**MRR**	**Recall@1**
DEB	SES_AMT_SINGLE	60.51	77.95	63.25	51.55	70.30	51.48
+ F	SES_AMT_SINGLE	59.31	75.90	65.00	52.91	66.26	54.29
MERCY	SES_AMT_SHUFFLED	62.75	78.43	67.39	49.67	64.36	51.03
+ F	SES_AMT_SHUFFLED	63.19	79.51	69.03	50.61	63.86	51.03
+ KL	SES_AMT_SHUFFLED	63.62	80.63	67.50	**53.50**	72.35	53.55
+ KL + F	SES_AMT_SHUFFLED	**64.77**	**81.75**	**69.14**	53.13	**73.23**	**55.62**

Table 5: Performance of AE/RS metrics. KL=Kullback-Leibler divergence loss. F=Feedback

6 Experiments

As seen in Table 2 and Table 3, DEB (Sai et al., 2020) performs the best amongst all existing metrics therefore we use it as our baseline.

For training we leverage the SES_AMT dataset which consists of 8 different candidates per context. We augment the data by perturbing the location of candidate responses. In our experiments, we shuffle the candidates 10 times, and thus have 10 times more training samples. We denote this dataset as SES_AMT_SHUFFLED. In order to compare the benefit of training on multiple candidates versus one, we create a dataset SES_AMT_SINGLE. Models finetuned on SES_AMT_SINGLE take in one response as input while models finetuned on SES_AMT_SHUFFLED take in all candidates simultaneously as input by concatenating them.[3]

When training on the SES_AMT_SINGLE dataset, we append the entire feedback to the response. However, when training on the SES_AMT_SHUFFLED dataset, the tokens are too long since DEB has only 512 positional embeddings. We found that truncating the feedback led to performance degradation. Therefore, for each response we take the corresponding feedback and map to one of the following templates: *"the response is relevant"*, *"the response is not relevant"*, *"the response is somewhat relevant"*. We create a list of keywords for each template. If at least one of the keywords exist in the original generated feedback we replace it with the corresponding template. Through this method the entire response/feedback pair can be fitted with minimal context truncation and results in faster encoding.

7 Results

We present results of comparing MERCY against DEB and MERCY with feedback as additional input in Table 5, we draw the following findings[4]:

1) Training our model in a joint fashion with multiple candidates as input (SES_AMT_SHUFFLED) outperforms training on a dataset with a single candidate as input (SES_AMT_SINGLE), showing the benefit of concurrent response ranking.

2) The use of feedback improves Recall@1 with a 3% for SES_AMT_SINGLE; however, there is not similar improvement when training on SES_AMT_SHUFFLED. This may be due to the fact that we had to rewrite the feedback to contain less information in order to fit into the model's input. However, leveraging KL-Divergence loss alongside feedback shows improvements. Additionally, the feedback provides explanations which allows for interpretability during evaluation.

3) Leveraging the KL-Divergence loss (MERCY + KL) outperforms MERCY with Cross-entropy loss, an improvement of Recall@1 score from 51.03 to 53.55 on the IES-V2 test set. This could be due to the way the data has been annotated. Each response in the dataset has a score between [0-2], to show the rank of responses. While Likert scales suffer due to annotator bias, ranking responses are more robust to this bias. The KL-Divergence loss determines how different the model's output distribution is from the ground-truth distribution and therefore does not rely on the specific Likert scores, but rather on the relative ordering of responses.

8 Conclusion

We introduce MERCY, an RS model that ranks responses by comparing multiple responses synchronously and leveraging natural language feedback. We demonstrate that feedback generated from a LLM through a few-shot setup improves the performance of MERCY. Additionally we introduce the SES dataset, a more realistic RS dataset with human annotated machine generated responses and show the low performance of baseline AE/RS metrics on SES and other existing realistic response raking testsets.

---

[3] Training parameters are in Section B in the Appendix.

[4] We trained on SES_AMT and evaluated on the test split of SES_INTERNAL_FILTERED.

9 Limitations

(1) In this work, we only look at the relevancy when generating feedback; however, this can be expanded to contain other useful evaluation dimensions such as engagingness and contradiction. (2) We perform experiments on English-only conversations which makes our work biased toward the English language. (3) The performance on IES is far from satisfactory; however, this demonstrates the difficulty of this problem and a strong test set is useful for better development of AE/RS systems. (4) The number of responses that can be ranked by MERCY is limited by the context length of the model; however, the baseline model which only takes in one response at a time will get computationally expensive as the number of responses grows.(5) The responses in SES and IES do not consider the most recent conversational models such as ChatGPT[5].

10 Ethics Statement

All authors of this paper acknowledge and agree with the ACM Code of Ethics. In our study, we ensure that our work is compatible with the provided code, specifically in the terms of presenting a non-offensive dataset construction.

In order to accomplish a comprehensive analysis of AE/RS metrics on the response ranking task, we collect a dataset containing human rankings for generated responses conditioned on existing human-human conversations with polished contents. The main concern is that generated responses based on well-known state-of-the-art dialogue models could have offensive content which is out of our work's scope.

In the feedback generation component leveraged in our proposed metric which is based on prompting a LLM, the outputs show whether a response is relevant or not and explain why that is the case, hence the chance of generating inappropriate contents is near zero.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. Plato-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of the 2019 Conference on International Conference on Learning Representations (ICLR)*.

Maxine Eskenazi, Shikib Mehri, Evgeniia Razumovskaia, and Tiancheng Zhao. 2019. Beyond turing: Intelligent agents centered on the user. *arXiv preprint arXiv:1901.06613*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the NAACL 2019 Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen workshop)*.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the 2020 Conference on Association for the Advancement of Artificial Intelligence (AAAI)*.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

---

[5] `https://openai.com/blog/chatgpt/`

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2041–2044, New York, NY, USA. Association for Computing Machinery.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. 2022. Improving zero and few-shot generalization in dialogue through instruction tuning. *arXiv preprint arXiv:2205.12673*.

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. *arXiv preprint arXiv:1907.10568*.

Prakhar Gupta, Yulia Tsvetkov, and Jeffrey Bigham. 2021. Synthesizing adversarial negative responses for robust response ranking and evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3867–3883, Online. Association for Computational Linguistics.

Michael Gutmann and Aapo Hyv¨arinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558, Online. Association for Computational Linguistics.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684.

Behnam Hedayatnia, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2022. A systematic evaluation of response selection for open domain dialogue. In *Proceedings of the 2022 conference on Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Huda Khayrallah and João Sedoc. 2020. Smrter chatbots: Improving non-task-oriented dialog with simulated multi-reference training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4489–4505.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1).

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the intrinsic information flow between dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries.

Sijia Liu, Patrick Lange, Behnam Hedayatnia, Alexandros Papangelis, Di Jin, Andrew Wirth, Yang Liu, and Dilek Hakkani-Tür. 2023. Towards credible human evaluation of open-domain dialog systems using interactive setup. In *AAAI 2023*.

Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, and Yifei Zhang. 2021. A graph reasoning network for multi-turn response selection via customized pre-training. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics*.

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 (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 2020 conference on Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 681–707, Online. Association for Computational Linguistics.

Masahiro Mizukami, Hideaki Kizuki, Toshio Nomura, Graham Neubig, Koichiro Yoshino, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Adaptive selection from multiple response candidates in example-based dialogue. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 784–790. IEEE.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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 Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 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, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. 2020. Evaluating dialogue generation systems via response selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 593–599, Online. Association for Computational Linguistics.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with natural language feedback. *arXiv preprint arXiv:2204.14146*.

Weiyan Shi, Emily Dinan, Kurt Shuster, Jason Weston, and Jing Xu. 2022. When life gives you lemons, make cherryade: Converting feedback from bad responses into good labels. *arXiv preprint arXiv:2210.15893*.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 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*, pages 2021–2030, Online. Association for Computational Linguistics.

Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2022. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. *NAACL Findings.(to appear)*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Computational Linguistics*, 45(1):163–197.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2020. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. *arXiv preprint arXiv:2009.06265*.

Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

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

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. 2022. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. In *Proceedings of the 2022 Conference on Association for the Advancement of Artificial Intelligence (AAAI)*.

Linhao Zhang, Dehong Ma, Sujian Li, and Houfeng Wang. 2021b. Do it once: An embarrassingly simple joint matching approach to response selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 4872–4877. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020a. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9604–9611.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 2020 Conference of the American Chapter of the Association for Computational Linguistics: Demonstration*. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

A More about AE and RS Metrics

In our study, we target model-based reference-free AE metrics which are more appropriate with no reliance and limitations on ground-truth responses (Novikova et al., 2017; Lowe et al., 2017; Yeh et al., 2021) in open-domain dialogue systems.

Bert_Ruber (Ghazarian et al., 2019) which is an advanced version of Ruber (Tao et al., 2018) leverages positive impact of contextualized word embeddings across with a cross entropy loss to distinguish between human-written responses from random matched ones. **Pone** (Lan et al., 2020) proposes to get top-k most similar randomly chosen responses to the dialogue context as more appropriate negative samples and augment data with more balanced number of generated positive responses. Similar to prior metrics, **Maude** (Sinha et al., 2020) also takes the BERT embeddings of dialogue and learns to assign quality score to the response by being trained with Noise Contrastive Estimation (Gutmann and Hyv¨arinen, 2010) loss between positive and negative samples.

Predictive_Engagement(Ghazarian et al., 2020) returns the engagingness label for corresponding responses and shows its importance in achieving a more precise open-domain dialogue evaluation.

FlowScore (Li et al., 2021) as its name implies models the dynamic flow of the dialogue by leveraging three training objectives to consider the flow and semantic influence of the context and utterances. **DEAM** (Ghazarian et al., 2022) focuses on AMR architecture of conversations to apply semantic-level perturbations and generate more natural looking incoherent conversations. The classification model trained on the resulted dataset has higher performance than its counterpart trained on heuristically generated negative samples. **Grade** (Ye et al., 2020) considers topic transition dynamics by incorporating topic-level graph representations of the dialogue alongside utterance-level contextualized representations trained with the ranking loss and outputs the coherence score for the response. **DynaEval** (Zhang et al., 2021a) transforms each dialogue to a graph where its nodes represent the utterances and the edges demonstrate the dependency between utterances. A graph convolutional network is adopted to measure the quality of the response as well as the whole dialogue. A contrastive loss is defined to distinguish between positive dialogues and negative ones resulted from shuffling and replacements in the utterances.

USL-H (Phy et al., 2020) pays attention to the multi-facet feature of the evaluation. It is a mixture of three metrics for capturing Understandability, Sensibleness, and Likability in Hierarchy. **HolisticEval** (Pang et al., 2020) decomposes the overall quality into four sub-metrics: coherency, fluency, diversity and logical self-consistency. The first two aspects are yielded from probability distribution of GPT-2 model, while diversity is computed by taking into account n-gram entropy and logical self-consistency follows Natural Language Inference models. **USR** (Mehri and Eskenazi, 2020b) is a combination of naturalness, context consistency, knowledge conservation sub-metrics. The likelihood estimated by a fine-tuned RoBERTa model based on MLM objective shows the naturalness. The conditional distribution of a fine-tuned RoBERTa model for the retrieval task can potentially demonstrate the context consistency and naturalness. **FED** (Mehri and Eskenazi, 2020a) is also a multi-dimensional metric without necessity of training. It defines positive and negative follow-up responses designed for each aspect and computes their likelihood using DialoGPT model.

MDD (Zhang et al., 2022)'s main goal is to be a robust metric over different domains and to achieve it two types of models are trained: teacher and student models. The teacher model is trained on human annotated positive and negative responses and later applied on synthetic dataset to get pseudo-labels. Following, the student model is trained to have similar predictions as the teacher model, to be able to be covered after injecting noise to the responses, and to be better adaptable to the multi-domain synthetic datasets.

DEB (Sai et al., 2020) determines the efficiency of pretraining on large-scale dialogue corpora for the evaluation task. The pretraining on Reddit dataset including positive comments and randomly picked negative responses with incorporating cross entropy loss objective and subsequently finetuning on some human crafted positive and adversarial negative samples makes DEB as the most accurate evaluation metric.

Apart from AE metrics that assess the quality of responses from different perspectives, RS models learn to assign different rankings to the responses. One of the pioneer baselines for the response ranking task is **BM25** (Robertson et al., 2009) that leverages keyword similarity to rank responses given a context. We pursue the idea by Henderson et al.

(2019) to rank candidates based on their BM25 vector's inner product with the context's BM25 vector.

DialogRPT (Gao et al., 2020) contains a set of GPT-2 based models that are trained on human feedback data on social media platforms indicating different factors such as the number of replies, maximum length of the dialog after the reply, the difference between upvotes and downvotes. The overall ranking of each response includes scores showing predicted human feedback of responses and whether the response is human-like or not.

SABert_KeySem (Gupta et al., 2021) is a recent response ranking metric that proposes two modern approaches for constructing negative candidates that are used to be classified from responses with higher rankings. First is a mask-and-fill approach that masks spans of utterances and infill them using GPT2-based model conditioned on random contexts. Second is also a GPT2-based model that tries to complete a response conditioned on its keywords and a random context. A Speaker-Aware Bert (SA-Bert) (Gu et al., 2020) classifier trained on such data outperforms different existing baselines.

B MERCY training parameters

We start finetuning our model on the DEB (Sai et al., 2020) checkpoint. We set training for 10 epochs and do early stopping once the loss on the validation set does not go down. We use a training batch size of 8. We use the Adam optimizer with a learning rate of 5e-6. Additionally we use a dropout rate of 0.1. The model was trained on a single NVIDIA A100 GPU. The average inference time of MERCY over 1000 runs on an NVIDIA A100 GPU is 21.88 ms.

C Annotation issues in IES

As seen in Table 6, even though Candidate 3 is a relevant response it is still annotated as not relevant by the worker. This could be due to the fact that Candidates 2 and 4 are more interesting as they offer up more information.

D Models used in IES

All models used in this dataset are GPT2-based (Radford et al., 2019) with a variety of sizes. For each model, both the Language Modeling Head and Multiple Choice Head of GPT2 are finetuned 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 nucleus sampling is used to generate the response.

E Models used in SES

The models used to generate responses are: BlenderBot (Roller et al., 2021), GPT2-XL (Radford et al., 2019) finetuned on the Blended Skills Talk dataset (Smith et al., 2020), Plato-2 (Bao et al., 2021) and Plato-XL. GPT2-XL was finetuned in the same fashion as the models in D. To generate 8 candidates for each dialog context we generate three responses from GPT2-XL using nucleus sampling, two responses from Plato-2 using nucleus sampling, one response from Plato-XL using beam search, one response from Blenderbot using nucleus sampling and one response from Blenderbot using beam search.

F Feedback prompt

G Generated feedback examples

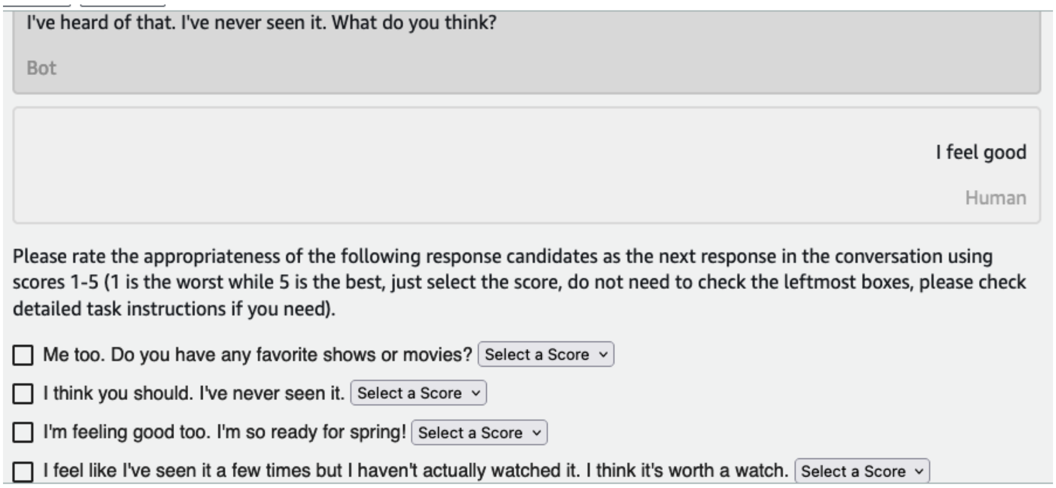
Table 8 shows an example of a relevant and non-relevant response with its corresponding feedback. Additionally the last example includes an explanation that is incorrect. In the last example the feedback hallucinates information that is not mentioned in the response. To determine how diverse the generated feedback is we compute self-BLEU and get a score of 0.67 where the lower the score the more diverse the feedback is.

H Mechanical Turk interface for SES data collection

Our qualifications included a HIT Approval Rate greater than 95%, number of HITs approved greater than 1000, location is US and Masters have been granted. Additionally we had asked the Turkers to do a qualification HIT where if they passed they were granted permission to do the actual data collection task. They were paid $15 per hour. A screenshot of the conducted experiment is shown in Figure 3.

Context	Annotation
Bot: I don’t blame you. What do you do for work? Human: i manage social media for clients. how about you? Bot: Same here! I manage my social media and my website. I like to use tools like Google and Facebook. Human: google has a heck of a lot of tools for managing all types of things	
Candidate 1: i manage social media for clients	Not relevant
Candidate 2: Same here! I manage my social media and my website. I like to use tools like Google and Facebook.	Relevant
Candidate 3: Yes, and Facebook is a great social media tool too.	Not relevant
Candidate 4: It really does! I use it for everything from social media to email to file sharing.	Relevant

Table 6: Example of annotation issues in IES


I've heard of that. I've never seen it. What do you think?

Bot

I feel good

Human

Please rate the appropriateness of the following response candidates as the next response in the conversation using scores 1-5 (1 is the worst while 5 is the best, just select the score, do not need to check the leftmost boxes, please check detailed task instructions if you need).

☐ Me too. Do you have any favorite shows or movies?

☐ I think you should. I've never seen it.

☐ I'm feeling good too. I'm so ready for spring!

☐ I feel like I've seen it a few times but I haven't actually watched it. I think it's worth a watch.

Figure 3: The interface for collecting the SES and IES-v2 data

Below are a series of examples dialogs, followed by a question about the quality of the socialbot's responses. This question must be answered by carefully considering each response in the proceeding dialog, assessing whether the response satisfies the quality/criteria the question is asking for, and ultimately answering the question with a detailed description of the assessment. The answer should be detailed, quote the concerning parts of the dialog and give an overall score out of 10.

User: Hi!
Socialbot: Hi there.
User: What are you listening to?
Socialbot: All sorts of music. I listen when no-one is chatting to me.
User: That's great!
Socialbot: Thanks.
User: Do you chat here often?
Socialbot: I am a talking computer, after all, so of course I could talk here, if I needed to.
User: Let's talk about Taylor Swift!
Socialbot: I watched The Matrix recently. It's got to be my favorite movie.
Question: How relevant are the socialbot responses?
Answer: The socialbot responses in this dialog are not very relevant. The first response to the question about what the socialbot is listening to is somewhat relevant, as it mentions listening to music. However, the subsequent responses do not directly address the user's questions or statements. The response to the question about chatting often simply states that the socialbot is a talking computer and could talk if needed, but does not address the frequency of its chat activities. The response to the statement about Taylor Swift is completely unrelated, discussing The Matrix movie instead. Overall, the socialbot's responses in this dialog score a 2 out of 10 for relevance.

User: Hi!
Socialbot: Hi! What's up?
User: Nothing much, how about you
Socialbot: Not much either.
User: What are you doing
Socialbot: Playing Terraria. What about you?
User: Sitting in a meeting
Socialbot: What kind of meeting?
User: Can't say
Socialbot: It's probably boring, isn't it?
Question: How engaging are the socialbot responses?
Answer: In this dialog, the socialbot's responses could be considered somewhat engaging. The socialbot initiates conversation by asking "What's up?" and continues the conversation by asking the user what they are doing. However, the socialbot's response of "It's probably boring, isn't it?" could be perceived as dismissive or judgmental. Overall, I would rate the socialbot's responses as a 6 out of 10 in terms of engagement.

User: Hi!
Socialbot: Hey
User: How's your day going?
Socialbot: pretty good. yours?
User: Ehh it's fine. I didn't do so well on that history test, actually..
Socialbot: oh what happened?
User: Apparently Christopher Columbus didn't fight in the Civil War :')
Socialbot: hahah wait for real?
User: I know right! Are you taking History next semester?
Socialbot: No I'm not in school anymore
Question: How relevant are the socialbot responses?
Answer: The socialbot responses are relevant in this dialog. The socialbot asks the user how their day is going and responds to the user's mention of not doing well on a history test by asking for more information. The socialbot also responds to the user's question about taking history next semester with information about their current schooling status. Overall, the socialbot's responses are timely and relevant to the conversation, earning a score of 8 out of 10 for relevance.

Table 7: Prompt sent into BLOOMZ-175B

... User: I am an avid gamer I just got a new one I am obsessed with. Socialbot: I am not much of a gamer myself but I do know a few, what game did you get? User: It’s called the awakening. What do you do if you don’t game? Socialbot: i like to watch movies and tv
Question: How relevant are the socialbot responses?
Answer: The socialbot’s last response is highly relevant. The socialbot’s response discusses the types of things the user likes to do, including games, movies, and TV. Overall, the socialbot’s responses earn an 8 out of 10 for relevance.
... User: i love eating anything italian , watching forensic files and jeopardy Socialbot: Great shows! Have you ever been it italy? User: i have been there a couple of times Socialbot: How is it? I’ve always wanted to go? User: its beautiful and its nice to use overall Socialbot: I like to run in the mountains.
Question: How relevant are the socialbot responses?
Answer: The system response is not relevant. The socialbot states information about running in mountains and likes of sports. The response is not relevant to the conversation as the user has not indicated that they like sports or running in mountains.
... User: I may take a run myself, but am really thinking of going to American Eagle, my favorite store...they have a sale! Socialbot: What do you like to purchase there? I like the tee’s personally User: yes the tee shirts are great! Also have some nice jeans. Socialbot: You have any hobbies?
Question: How relevant are the socialbot responses?
Answer: The last system response is relevant. The system asks the user about hobbies and what they might purchase at American Eagle. The system’s last response mentions the store’s sale. The system’s responses score a 4 out of 10 for relevance in this dialog.

Table 8: More examples of generated feedback