FineD-Eval: Fine-grained Automatic Dialogue-Level Evaluation

Chen Zhang[†,⋆] **Luis Fernando D'Haro**[‡] **Qiquan Zhang**[†]
Thomas Friedrichs[⋆] **Haizhou Li**[♥,†,◇]
[†]National University of Singapore [⋆]Robert Bosch (SEA), Singapore
[‡]Universidad Politécnica de Madrid, Spain [◇]Kriston AI Lab, China
[♥]The Chinese University of Hong Kong, Shenzhen, China
chen_zhang@u.nus.edu

Abstract

Recent model-based reference-free metrics for open-domain dialogue evaluation exhibit promising correlations with human judgment[1]. However, they either perform turn-level evaluation or look at a single dialogue quality dimension. One would expect a good evaluation metric to assess multiple quality dimensions at the dialogue level. To this end, we are motivated to propose a multi-dimensional dialogue-level metric, which consists of three sub-metrics with each targeting a specific dimension. The sub-metrics are trained with novel self-supervised objectives and exhibit strong correlations with human judgment for their respective dimensions. Moreover, we explore two approaches to combine the sub-metrics: metric ensemble and multitask learning. Both approaches yield a holistic metric that significantly outperforms individual sub-metrics. Compared to the existing state-of-the-art metric, the combined metrics achieve around 16% relative improvement on average across three high-quality dialogue-level evaluation benchmarks.

1 Introduction

In the study of generative dialogue systems, we heavily rely on some reference-based static metrics, such as BLEU (Papineni et al., 2002), to measure improvements during system development and to compare various model variants. These metrics still need improvement due to their poor correlations with human judgment (Liu et al., 2016) and poor interpretability (Mehri and Eskenazi, 2020b).

Recently, model-based reference-free metrics (Yeh et al., 2021) represent one of the ways to address the limitations of static reference-based metrics. Although such metrics exhibit promising correlations with human evaluation, most of them (Tao et al., 2018; Ghazarian et al., 2019; Huang et al., 2020; Sinha et al., 2020; Mehri and Eskenazi, 2020b; Phy et al., 2020; Pang et al., 2020; Zhang et al., 2021c) target turn-level evaluation, i.e., they focus on single-response quality, such as contextual relevance and naturalness. When evaluating a multi-turn human-chatbot dialogue, turn-level metrics do not model the dialogue in totality, but frame it as a set of context-response pairs. They assign scores to every chatbot responses in the dialogue. Hence, an aggregation strategy is required to derive the single dialogue-level metric score, such as taking the average of all the response-level scores. Both prior works (Zhang et al., 2021a; Yeh et al., 2021) and our experimental results in §5 suggest that such an approach yields sub-optimal dialogue-level evaluation. The reason may be that turn-level metrics do not model the dependency among utterances within multi-turn interactions, it is difficult for them to spot errors that are only obvious after observing the entire conversation (Ghandeharioun et al., 2019; Ghazarian et al., 2022).

There are some metrics that perform multi-turn evaluation. However, they focus only on a single dimension, such as coherence or overall impression (Mesgar et al., 2020; Zhang et al., 2021a; Li et al., 2021; Ghazarian et al., 2022). When evaluating a dialogue, they assign a single score to quantify one aspect of dialogue quality. As pointed out in Mehri et al. (2022), dialogue quality is inherently multi-faceted. By breaking down the quality of the dialogue into multiple fine-grained dimensions, we may provide a more interpretable and descriptive dialogue evaluation. With such an interpretable metric, dialogue researchers know exactly which aspect of the dialogue system to improve.

To this end, we propose a multi-dimensional metric, dubbed FineD-Eval[2], which consists of specialized sub-metrics. Each sub-metric targets a specific fine-grained dimension and all sub-metrics

[1]As shown in (Yeh et al., 2021), most reference-free metrics can achieve around 0.3 to 0.6 Spearman correlations on various turn-level benchmarks. However, on the dialogue-level benchmarks, most metrics perform poorly (< 0.2 Spearman correlations).

[2]https://github.com/e0397123/FineD-Eval

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3336 - 3355
December 7-11, 2022

are trained in a self-supervised manner without reliance on any human annotations.

To develop FineD-Eval, our first step is to identify the dimensions for metric design. It is a well-known phenomenon that human judges do not provide completely independent assessments for various fine-grained dimensions. For instance, Sai et al. (2021) analyzes the human ratings with respect to (w.r.t.) different fine-grained dimensions on four text generation tasks and has observed moderate correlations for most dimension pairs. Intuitively, we want to select dimensions that are less correlated such that our metric can holistically capture the dialogue quality from different perspectives. The selection process is guided by an analysis on fine-grained human ratings of dialogue-level evaluation data (§2). Through the analysis, we want to cluster the dimensions into relatively independent dimension groups and then, select representative dimensions from different dimension groups.

Next, we propose dimension-specific strategies for training the sub-metrics. (§3.3). The sub-metrics, which target the representative dimensions, can also be applied to evaluate other dimensions in their respective dimension groups. Furthermore, both Yeh et al. (2021) and Zhang et al. (2021d) highlight that the combination of different metrics leads to better correlations with human evaluation than individual specialized metrics. We are motivated to explore how to combine the sub-metrics into a unified one. Specifically, both the metric ensemble and multitask learning (Caruana, 1997) are examined (§3.4).

Finally, in the experiments (§5), we demonstrate that (1) the sub-metrics highly correlate with human judgment for their target dimensions. (2) The scores assigned by FineD-Eval are more interpretable than the existing metrics. (3) With either metric ensemble or multitask learning, FineD-Eval significantly outperforms existing state-of-the-art metrics as well as individual sub-metrics on three high-quality dialogue-level evaluation benchmarks.

2 Analysis of Human Evaluation Data

2.1 Grouping of the Dimensions

In this section, we analyze the human ratings of FED (Mehri and Eskenazi, 2020a), a high-quality dialogue-level evaluation benchmark. Each dialogue in FED is annotated by five human judges

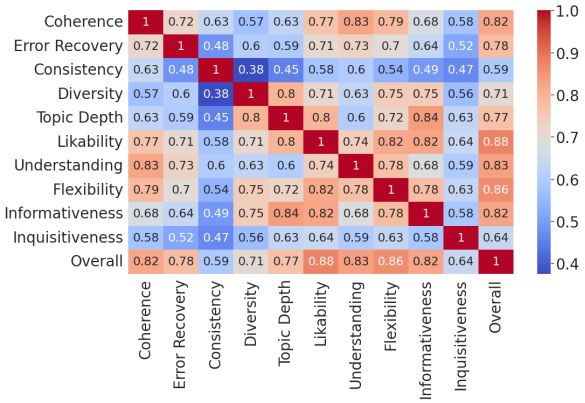

Figure 1: Spearman correlations of dimension pairs on FED.

Group	Quality Dimensions
Coh	Coherence, Understanding
Lik	Likability, Flexibility, Informativeness
Top	Topic Depth, Diversity, Informativeness
Con	Consistency
Inq	Inquisitiveness
Err	Error Recovery

Table 1: Grouping of the dimensions. We adopt the first three letters of the representative dimension within each group as the corresponding group name.

for 11 different quality dimensions[3], as shown in the axis labels of Figure 1. We choose FED for our analysis because the dataset covers the most comprehensive list of dialogue quality dimensions. In addition, the human annotation quality of FED is high as evidenced by the strong inter-annotator agreements w.r.t. different dimensions[4].

Figure 1 presents the Spearman correlations of different dimension pairs on FED. We can observe that all dimensions are interdependent, with correlations ranging from 0.38 to 0.88. Based on their extent of interdependence, we cluster the 10 dimensions (excluding the "Overall" category) into six groups, as shown in Table 1. We adopt the first three letters of the representative dimension within each group as the corresponding group name. The representative dimension in each group is chosen based on criteria discussed in §2.2.

A dimension is treated as an independent group if it does not correlate strongly with any of the other dimensions (≥ 0.75). Hence, consistency, inquisitiveness, and error recovery can be per-

[3]The detailed definitions of all dimensions are presented in Table 11 of the Appendix

[4]Above 0.75 in terms of Spearman correlations for all the dimensions except that of consistency, which is 0.562.

ceived as three independent dimension groups: *Con*, *Inq*, and *Err* respectively. The remaining dimensions are more or less correlated with each other. Based on the following four observations: (1) coherence strongly correlates with understanding (0.83); (2) The likability-flexibility and likability-informativeness correlations are both 0.82; (3) The correlation between topic depth and informativeness is as high as 0.84; and (4) Diversity only strongly correlates with topic depth (0.8), the remaining seven dimensions can be clustered into three groups: *Coh*, *Lik*, and *Top*.

The categorization may not be perfect as *Coh*, *Lik*, and *Top* are not completely independent from each other. For example, informativeness can be found in both group *Lik* and group *Top*. A possible explanation is that humans generally like knowledgeable chatbots, which can discuss different topics in depth rather than those that generate dull responses (See et al., 2019; Roller et al., 2021). To improve the categorization, future work may conduct similar analysis on large-scale dialogue-level human annotations.

2.2 Dimension Selection

As metioned in §1, we want to identify fine-grained dimensions that are less similar. Hence, we select only one dimension from each group and avoid those that are shared between two different groups. In addition, to further reduce the complexity of FineD-Eval, we implement the following rules to narrow down the selection to only three fine-grained dimensions.

First, dimensions that highly correlate with the "Overall" category (> 0.75) are considered. The intuition is that a high correlation with "Overall" indicates more influence from the fine-grained dimension on human annotators' overall impression about a dialogue. Second, we filter out dimensions with low inter-annotator agreement (< 0.6)[5], because low inter-annotator agreements may suggest the dimension is complex to evaluate and human annotators have different understandings of the dimension (Mehri et al., 2022). Lastly, we choose dimensions based on how often they are marked as "N/A" (not applicable) by the human judges. A high frequency indicates that the dimension is not generally applicable in different contexts. Most dimensions do not contain a "N/A" rating except "Error recovery", which has been marked as "N/A" 25% of the time.

Based on the rules, we choose the following three dimensions: **coherence**, **likability**, and **topic depth**. In addition to the rules, we choose these dimensions because they are also widely studied in open-domain dialogue systems. Researchers spend significant amount of efforts on developing coherent, engaging, and knowledgeable chatbots (Adiwardana et al., 2020; Hedayatnia et al., 2020; Shuster et al., 2021). Designing meaningful metrics along these three dimensions can benefit the current open-domain dialogue research. Though other dimensions, such as consistency (Nie et al., 2021), inquisitiveness (See et al., 2019), and long-term memory (Xu et al., 2022) are equally important, their evaluation deserves a thorough study on its own. Hence, we leave them for future work.

3 Methodology

3.1 Problem Formulation

We formally define the dialogue-level evaluation task. Suppose that we have a dialogue evaluation dataset, $\mathbb{D}$, which contains n human-chatbot dialogues, $\mathbb{D} = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$. d_j is annotated by several human judges for a set of quality dimensions, $\mathbb{Q}$. Each human judge provides a rating to d_j for individual dimension, $q \in \mathbb{Q}$. We use $r^q_{d_j}$ to denote the average Likert rating provided by all human annotators to d_j for q.

Our goal is to learn dimension-specific metrics, $M^q(d_j) \rightarrow s^q_{d_j}$, where $s^q_{d_j}$ is the metric score reflecting how good d_j is for dimension q as perceived by M^q. To assess the performance of M^q on $\mathbb{D}$, the correlation, denoted as ρ^q, between $S^q = \{s^q_{d_1}, \ldots, s^q_{d_j}, \ldots, s^q_{d_n}\}$ and $R^q = \{r^q_{d_1}, \ldots, r^q_{d_j}, \ldots, r^q_{d_n}\}$ are calculated. Higher ρ^q indicates better performance of M^q on $\mathbb{D}$.

3.2 General Framework

We propose a multi-dimensional dialogue-level metric, FineD-Eval, which is a combination of three specialized sub-metrics, M^q, where $q \in \{\text{coherence}, \text{likability}, \text{topic depth}\}$. We explore two approaches for combining the sub-metrics, metric ensemble and multitask learning. Metric ensemble is a late fusion approach whereby the predictions made by the sub-metrics are combined. Multitask learning, on the other hand, is an early fusion approach whereby the sub-metrics will share a common text encoder while having different output

[5]Only consistency has an inter-anntator agreement below 0.6.

layers. Details of both approaches are discussed in §3.4. Here, we focus on the details of M^q.

To train M^q, we formulate a preference learning approach (Fürnkranz and Hüllermeier, 2011). Given a pair of dimensions-specific positive and negative training dialogue samples, d_{tr}^+ than d_{tr}^-, M^q learns to predict a higher score for d_{tr}^+ than d_{tr}^-. The strategies for constructing (d_{tr}^+, d_{tr}^-) are outlined in §3.3. During training, a mini-batch is formed with two types of data instances[6]: (1) (d_{tr}^+, d_{tr}^-) with label $y = 1$; (2) (d_{tr}^-, d_{tr}^+) with label $y = -1$. M^q outputs two scalar values $s^q_{d_{tr}^+}$ and $s^q_{d_{tr}^-}$ that correspond to d_{tr}^+ and d_{tr}^- respectively. The following margin ranking loss is adopted to train the model:

$$\mathcal{L}_q = max(0, y * (x_1^q - x_2^q) + 0.1) \quad (1)$$

where (x_1^q, x_2^q, y) can be either ($s^q_{d_{tr}^+}$, $s^q_{d_{tr}^-}$, 1) or ($s^q_{d_{tr}^-}$, $s^q_{d_{tr}^+}$, -1).

The pairwise ranking formulation is motivated by previous works on dialogue evaluation (Mesgar et al., 2020; Huang et al., 2020; Gao et al., 2020; Zhang et al., 2021a). Compared to direct assessment approaches (Zhang et al., 2021c; Ghazarian et al., 2022), the main advantage of pairwise ranking is that the model can implicitly learn the features that distinguish the good dialogues from the bad ones based on a large quantity of dialogue pairs for a specific quality dimension.

The network architecture of M^q is straightforward. RoBERTa-base (Liu et al., 2019) is adopted as the text encoder, $\mathcal{T}$, which maps (d_{tr}^+, d_{tr}^-) to dense representations ($\mathbf{H}_{tr}^+$, $\mathbf{H}_{tr}^-$). Both d_{tr}^+ and d_{tr}^- are formulated as a token sequence with special token "</UTT>" to delimit different utterances. Next, ($\mathbf{H}_{tr}^+$, $\mathbf{H}_{tr}^-$) are converted into vector representations ($\mathbf{h}_{tr}^+$, $\mathbf{h}_{tr}^-$) with average pooling. Through a linear layer with output size 1 and a Sigmoid activation function, $\mathbf{h}_{tr}^+$ and $\mathbf{h}_{tr}^-$ are transformed into scalar values $s^q_{d_{tr}^+}$ and $s^q_{d_{tr}^-}$ respectively. During inference, given $d_j \in \mathbb{D}$, the scalar value $s^q_{d_j}$ output by M^q is the corresponding metric score.

3.3 Dimension-Specific Sampling Strategies

In this section, we discuss different strategies to obtain dimension-specific training dialogue pairs. All (d_{tr}^+, d_{tr}^-) samples are automatically constructed from human-human dialogue datasets without reliance on human annotations.

Coherence (*Coh*) We consider two strategies for coherence. The first is utterance order shuffling whereby dialogues from existing human-human dialogue corpora (Li et al., 2017; Dinan et al., 2020) are treated as d_{tr}^+. To obtain d_{tr}^-, we randomly permute the order of utterances in d_{tr}^+. This strategy has been widely adopted in previous dialogue coherence studies (Cervone et al., 2018; Mesgar et al., 2020; Zhang et al., 2021a).

The second strategy, question-answer (QA) relevance scoring, is motivated by the Gricean maxims (Grice, 1975) whereby effective communication involves being relevant, i.e., one should provide information that is relevant to the current exchange. A natural and logical flow of conversation often involves asking and answering questions, which is a form of information exchange. Humans usually prefer answers that are straight to the point rather than those that are vague and off-topic. Concretely, we select dialogues in existing dialogue corpora[7] that are more than 4 utterances and contain at least one question-answer pair. Next, we use a pretrained BERT-based QA evaluator from HuggingFace[8] to score each QA pair within a dialogue. The evaluator provides a relevance score between 0 and 1 (the higher the better). Then, we average the relevance scores of all QA pairs within the dialogue to derive the dialogue-level QA relevance score. Finally, two thresholds, (τ_{low}^{rel}, τ_{high}^{rel}), are chosen. Dialogues with scores lower than τ_{low}^{rel} are considered d_{tr}^-. Those with scores higher than τ_{high}^{rel} are considered d_{tr}^+. (τ_{low}^{rel}, τ_{high}^{rel}) are heuristically determined to ensure sufficient data in both the positive and negative classes.

Likability (*Lik*) Two strategies are applied to construct d_{tr}^+ and d_{tr}^- for likability. The first strategy, contradiction scoring, is motivated by the similarity attaction effect (Byrne et al., 1968; Nass and Lee, 2001). During human-human interaction, people tend to favour others who share similar opinions or preferences with them. On the contrary, conveying contradictory opinions or information may lead to disagreement and user dissatisfaction.

[6] This formulation is to avoid model relying on positions of the dialogues to make predictions.

[7] We hypothesize that even in human-human dialogue corpora, there are answers that are vague and off-topic due to the presence of low-quality crowd-source workers.

[8] https://huggingface.co/iarfmoose/bert-base-cased-qa-evaluator

To implement this strategy, we adopt a pre-trained natural language inference (NLI) model[9] to provide contradiction scores (between 0 and 1) to adjacent utterance pairs within human-human dialogues. For a dialogue containing k utterances, we have $k-1$ adjacency pairs, thus $k-1$ contradiction scores. The dialogue-level contradiction score is derived by computing the average of the $k-1$ scores. Finally, two thresholds, (τ_{low}^{contra}, τ_{high}^{contra}), are set. Dialogues with contradiction scores lower than τ_{low}^{contra} are considered d_{tr}^{+} and those with scores higher than τ_{high}^{contra} are considered d_{tr}^{-}.

The second strategy is based on the number of utterances that carry positive emotions within a dialogue, which we hypothesize can serve as a proxy indicator on how much the interlocutors enjoy conversing with each other. Intuitively, if a user feels a dialogue system is likeable, they tend to produce more engaging responses. To implement the strategy, we adopt a pre-trained sentiment classification model[10] and apply it to classify the sentiments w.r.t. all utterances within a dialogue. We treat dialogues, of which all utterances are classified into the positive classes, as d_{tr}^{+} and those containing less than two positive utterances as d_{tr}^{-}.

Topic Depth (*Top*) Discussing topics in depth is an important attribute of engaging conversations. During the human-human interaction, when the interlocutors deeply dive into a topic, they tend to produce semantically diverse utterances, which convey a large amount of information. On the other hand, if an interlocutor is not interested in the topic, they tend to produce dull responses, such as "Ok", "Good to know", and "I don't know". Even though, such responses can be appropriate in a wide range of contexts, they often do not convey much information (See et al., 2019). As most human-human dialogues are topic coherent, we can directly link topic depth to how semantically different the utterances are within a dialogue. Hence, we propose an entailment scoring strategy.

More specifically, given a dialogue of k utterances, a pre-trained NLI model[11] is used to provided entailment score to each utterance pair in the dialogue. In total, there are $\frac{(k-1)k}{2}$ entailment scores per dialogue. The dialogue-level entailment score is the average of all utterance-pair entailment scores in the dialogue. Similarly, two thresholds, (τ_{low}^{entail}, τ_{high}^{entail}), are applied to obtain positive and negative dialogues. Dialogues with entailment scores lower than τ_{low}^{entail} are regarded as d_{tr}^{+} and those with scores higher than τ_{high}^{entail} are d_{tr}^{-}.

3.4 Combining Dimension-Specific Metrics

Our analysis in §2 suggests that human evaluation across different quality dimensions are positively correlated. Therefore, a sub-metric that is specialized in evaluating one dimension can contribute to the evaluation of other dimensions as well. By combining different sub-metrics into a holistic one, we can achieve better correlations with human evaluation across different dimensions. We implement two FineD-Eval variants, FineD-Eval$_{en}$ (metric ensemble) and FineD-Eval$_{mu}$ (multitask learning).

Metric Ensemble Ensemble is a common technique adopted in machine learning to achieve better predictive performance than individual predictive models. In addition, it also helps improve model robustness by reducing the spread or dispersion of the predictions (Zhang and Ma, 2012).

In our case, FineD-Eval$_{en}$ is expected to achieve better ρ^{q} than M^{q} on $\mathbb{D}$. Given $d_j \in \mathbb{D}$, three sub-metrics, M^{coh}, M^{lik}, and M^{top} output three scores, $s_{d_j}^{\text{coh}}$, $s_{d_j}^{\text{lik}}$, and $s_{d_j}^{\text{top}}$ respectively. The metric score of FineD-Eval$_{en}$, $s_{d_j}^{\text{en}}$ is obtained by computing the arithmetic mean of ($s_{d_j}^{\text{coh}}$, $s_{d_j}^{\text{lik}}$, $s_{d_j}^{\text{top}}$).

Multitask Learning In multitask learning, a model is trained simultaneously with multiple tasks and a shared representation is learned to capture the commonalities among the related tasks (Crawshaw, 2020; Gao et al., 2022; Chen et al., 2021). Compared to FineD-Eval$_{en}$, the multitask model, FineD-Eval$_{mu}$, requires much less model parameters, but can achieve similar performance.

Similarly, FineD-Eval$_{mu}$ is also expected to achieve better ρ^{q} than M^{q} on $\mathbb{D}$. To implement FineD-Eval$_{mu}$, we need to first identify the related tasks for joint training. As described in §3.2, we have the preference learning tasks for M^{coh}, M^{lik}, and M^{top} respectively. Since the input and output of the three tasks are the same, we can adopt a hard-parameter sharing network to simultaneously learn the three tasks. More specifically, the text encoder $\mathcal{T}$, is shared among the three tasks. On top of $\mathcal{T}$, there are three independent linear layers with

[9] https://huggingface.co/roberta-large-mnli.

[10] https://huggingface.co/mrm8488/t5-base-finetuned-emotion

[11] Same as that used for the contradiction scoring strategy

output size 1, which serve as the sub-metrics for coherence, likability, and topic depth respectively.

During training, a mini-batch consists data that are uniformly drawn from the three training data sources described in §3.3. The parameter update of $\mathcal{T}$ depends on all data instances in the mini-batch while that of the three linear layers depends only on their corresponding task-specific input data. The losses of three tasks are summed together, $\mathcal{L}_{total} = \mathcal{L}_{coh} + \mathcal{L}_{lik} + \mathcal{L}_{top}$.

During inference, given $d_j \in \mathbb{D}$, FineD-Eval$_{mu}$ outputs three scalar values, $s^{\text{coh}}_{d_j}$, $s^{\text{lik}}_{d_j}$, and $s^{\text{top}}_{d_j}$ from the three linear layers respectively. Similar to metric ensemble, the final metric score, $s^{mu}_{d_j}$ is derived by taking the arithmetic mean of the three scores.

4 Experimental Setup

4.1 Training & Evaluation Datasets

For training, we prepare two datasets leveraging DailyDialog (DD) (Li et al., 2017) and ConvAI2 (CA) (Dinan et al., 2020). DailyDialog covers general day-to-day topics, such as school, work, and relationship. ConvAI2 is an extended version of PersonaChat (Zhang et al., 2018), which contains dialogues grounded by persona profiles. The detailed descriptions of DailyDialog and ConvAI2 are included in Appendix A. We choose DailyDialog and ConvAI2 because they cover a diverse sets of topics and our baseline metrics (§4.2) are mainly trained with these two datasets. The numbers of d^+_{tr} and d^-_{tr} obtained with various sampling strategies (§3.3) are listed in Table 10. When training each M^q on each dataset, we sample 100K and 10K of training and validation dialogue pairs respectively due to the large number of (d^+_{tr}, d^-_{tr}) combinations.

Attributes	FED	DSTC9	P-Eval
#Dialogues	125	2,200	3,316
Avg. #Utts per Dialogue	12.72	28.13	16.04
Avg. #Words per Utt	8.95	8.58	5.68
#Dimensions	11	11	8
#Ratings	23,750	71,203	26,528
#Models	3	10	29

Table 2: Statistics of the three evaluation benchmarks. "P-Eval" refers to the Persona-Eval benchmark.

Three benchmarks are adopted to assess the strength of the metrics. They are FED (Mehri and Eskenazi, 2020a), DSTC9-Interactive (Gunasekara et al., 2020), and Persona-Eval (See et al., 2019). The benchmarks' statistics are shown in Table 2 and their descriptions are presented in Appendix B. The definitions of various quality dimensions of the benchmarks are listed in Table 11 and Table 12. All metrics are assessed with dialogue-level Spearman correlations w.r.t. each fine-grained dimension on the three benchmarks. Note that we do not consider inquisitiveness, consistency, and error recovery in the main analysis, because none of the FineD-Eval sub-metrics target these dimensions. Nevertheless, we show the metrics' performance for the three dimensions in the Limitation section.

4.2 Baselines

Two groups of metrics are adopted. The first are state-of-the-art turn-level metrics, including USL-H (Phy et al., 2020), MAUDE (Sinha et al., 2020), MDD-Eval (Zhang et al., 2021b), and D-score (Zhang et al., 2021c). Turn-level metrics need to rely on aggregation strategies when evaluating multi-turn dialogues. In this paper, we adopt mean aggregation whereby the metric scores w.r.t. all chatbot turns in a dialogue are averaged to derive the single dialogue-level metric score. The second group includes DynaEval (Zhang et al., 2021a) and DEAM (Ghazarian et al., 2022), two state-of-the-art dialogue-level metrics. Detailed metric descriptions are outlined in Appendix C.

4.3 Implementation Details

The thresholds for the QA relevance strategy (τ^{rel}_{low}, τ^{rel}_{high}), the contradiction scoring strategy (τ^{contra}_{low}, τ^{contra}_{high}), and the entailment scoring strategy (τ^{entail}_{low}, τ^{entail}_{high}) are heuristically set to (0.85, 0.99), (0.20, 0.40), (0.01, 0.10) respectively. These thresholds ensure that there are enough data instances within both the positive and negative class.

Each experiment is repeated five times with different random seeds. Since we have prepared two training datasets, there are 5×2=10 variants for M^q, FineD-Eval$_{en}$, and FineD-Eval$_{mu}$ respectively. In §5, we report the average Spearman correlation scores across the 10 variants. Additional details associated with the training process, such as hyperparameters, model selection criteria, etc. are included in Appendix D.

5 Experiments & Analysis

In this section, we conduct the main analysis based on the following research questions (RQ): (1) Are dialogue-level metrics better than turn-level metrics for multi-turn dialogue evaluation? (2) Does our

Groups	Metrics	Coh	Und	Fle	Lik	Inf	Top	Div	Ove	Average
Turn	USL-H	19.50	*14.66*	18.98	31.00	35.39	31.86	20.70	24.10	23.27
	MAUDE	-22.37	-28.12	-28.18	-33.12	-32.76	-25.50	-19.67	-28.05	-27.22
	MDD-Eval	27.62	23.43	*8.35*	*11.87*	*6.86*	*-0.61*	*-6.83*	13.10	10.47
	D-score	31.15	31.14	32.77	27.04	23.82	22.17	20.83	37.58	28.31
Dialogue	DynaEval	42.29	36.06	38.91	39.78	39.61	43.94	33.16	48.18	40.24
	DEAM	46.82	46.68	52.19	50.49	59.20	61.90	59.20	54.72	53.90
Sub-metrics	M^{Coh}	52.86	52.35	43.87	47.71	42.84	40.54	36.43	53.02	46.20
	M^{Lik}	42.91	42.15	37.08	52.23	49.89	41.36	36.52	48.83	43.87
	M^{Top}	23.25	25.87	36.04	36.93	46.63	56.53	53.38	36.31	39.37
Combined	$M^{\text{Coh}} + M^{\text{Lik}}$	57.61	57.13	48.77	61.30	57.20	49.94	44.29	61.35	54.70
	$M^{\text{Coh}} + M^{\text{Top}}$	53.11	54.99	51.36	54.75	55.67	58.66	54.02	59.30	55.23
	$M^{\text{Lik}} + M^{\text{Top}}$	45.43	46.87	44.78	57.35	59.20	56.58	50.71	55.10	52.00
	FineD-Eval$_{en}$	**58.30**	**59.49**	53.74	64.75	64.17	61.23	55.09	65.47	60.28
	FineD-Eval$_{mu}$	57.66	57.37	**55.94**	**64.91**	**66.84**	**66.22**	**59.59**	**66.15**	**61.84**

Table 3: Spearman correlations (%) of different metrics on FED. Coh, Und, Fle, Lik, Inf, Top, Div, and Ove denote coherence, understanding, flexibility, likability, informativeness, topic depth, diversity, and overall impression respectively. The scores w.r.t. the best performing metric for each quality dimension are highlighted in bold. Statistically insignificant scores ($p > 0.05$) are italicized.

	USL-H	DEAM	M^{Coh}	FineD-Eval$_{en}$	FineD-Eval$_{mu}$
Coh	19.86	18.29	**22.04**	21.72	21.02
Und	17.82	18.89	**19.89**	19.43	19.02
Fle	18.31	18.49	**20.57**	19.95	19.32
Lik	18.62	16.82	20.76	**21.82**	21.60
Inf	15.76	14.17	15.34	17.64	**18.14**
Top	18.76	15.61	17.61	19.95	**20.35**
Div	12.95	**16.92**	12.87	14.36	14.38
Ove	19.77	19.37	22.89	22.94	**23.01**
Average	17.73	17.32	19.00	**19.73**	19.60

Table 4: Spearman correlations (%) of different metrics on DSTC9-Interactive. USL-H and DEAM are the best turn-level and dialogue-level baselines respectively. M^{Coh} is the best performing sub-metric. Full results can be found at Table 17.

	D-score	DEAM	M^{Coh}	FineD-Eval$_{en}$	FineD-Eval$_{mu}$
Interest	11.50	7.02	17.69	**19.72**	19.31
Sensible	22.33	13.92	**25.25**	22.65	20.23
Humanness	13.16	8.27	**20.31**	19.15	16.84
Enjoyment	15.07	9.21	19.52	**20.30**	18.98
Listening	20.38	13.67	**29.74**	22.52	20.61
Avoid Rep	10.58	17.45	**21.44**	17.60	**17.98**
Fluency	20.47	17.85	**22.29**	19.75	19.35
Average	16.21	12.49	**22.32**	20.24	19.04

Table 5: Spearman correlations (%) on Persona-Eval. "Avoid Rep" denotes avoid repetition. D-score and DEAM are the best turn-level and dialogue-level baselines respectively. M^{Coh} is the best performing sub-metric. Full results can be found at Table 18.

proposed sub-metrics correlate well with human evaluation for their target dimensions? (3) Does combining different sub-metrics help achieve better correlations for different dimensions? (4) Does FineD-Eval offer more interpretable results? (5) How reliable are negative samples constructed with sampling strategies in §3.3? Additional analyses are presented in Appendix E

RQ 1. First, we can observe in Table 3 that all dialogue-level metrics perform significantly better than the turn-level metrics across different quality dimensions, this observation is inline with conclusions from previous works (Yeh et al., 2021; Zhang et al., 2021a). However, USL-H and D-score outperform DEAM on On DSTC9-Interactive (Table 4) and Persona-Eval (Table 5) respectively. The good performance of USL-H and D-score may be attributed to that both metrics are an ensemble of multiple sub-metrics whereas DEAM is a single-model metric. This supports our claim that combining fine-grained metrics yield a holistic one that achieve better correlation with human judgment. Nevertheless, FineD-Eval$_{en}$ and FineD-Eval$_{mu}$, two dialogue-level metrics, outperform the turn-level metrics across all the dialogue-level benchmarks. We can conclude that in general, dialogue-level metrics perform better than turn-level metrics for multi-turn dialogue evaluation.

RQ 2. In the sub-metrics section of Table 3, we present the result of each dimensions-specific sub-metric on FED. We can observe that for coherence, understanding, and flexibility, M^{Coh} achieves the best performance in the sub-metrics group with 52.86%, 52.35%, and 47.71% Spearman correlations respectively. M^{Lik} achieves the best performance in likability and informativeness with spearman correlations of 52.23% and 49.89% respectively. For topic depth and diversity, M^{Top} performs the best among the three sub-metrics. The

empirical results meet our expectation that the three sub-metrics target dimension groups 1, 2, and 3 in Table 1 respectively. For the coherence dimension, M^{Coh} outperforms DynaEval and DEAM, which are also designed for evaluating dialogue-level coherence. Moreover, M^{Coh} performs exceptionally well on DSTC9-Interactive (Table 4) and Persona-Eval (Table 5) and significantly outperforms the turn-level and dialogue-level baselines on both benchmarks. This showcases the advantage of our utterance shuffling and QA relevance scoring strategies for coherence modeling.

RQ 3. We can observe in Table 3 that combining different sub-metrics generally performs better than individual sub-metrics for various fine-grained dimensions. For example, $M^{\text{Coh}} + M^{\text{Lik}}$ outperforms M^{Coh} for the coherence, understanding, and flexibility dimensions. It also outperforms M^{Lik} for the informativeness and likability dimensions. Furthermore, metrics in the combined group significantly outperform the sub-metrics as well as various baselines for the overall impression dimension. The observations support our claim in the introduction that combining sub-metrics helps achieve better correlations for different quality dimensions.

In addition, FineD-Eval$_{en}$ and FineD-Eval$_{mu}$ achieve a remarkable Spearman correlation of 65.47% and 66.15% respectively for the overall dimension on FED. Both outperforms state-of-the-art metrics as well as individual sub-metrics by a large margin. Such an observation showcases the effectiveness of metric ensemble and multitask learning for a holistic automatic dialogue-level evaluation.

Similar observations can be made w.r.t. DSTC9-Interactive (Table 4 and Table 17). Overall, FineD-Eval$_{en}$ and FineD-Eval$_{mu}$ achieve state-of-the-art performance on FED, DSTC9-Interactive, and Persona-Eval. Compared to the existing state-of-the-art metrics on the three benchmarks: DEAM, USL-H, and D-Score, FineD-Eval$_{en}$ achieves around 16% relative improvement on average while FineD-Eval$_{mu}$ attains around 14% relative improvement on average.

However, on Persona-Eval, the single metric, M^{Coh} performs better than FineD-Eval$_{en}$ and FineD-Eval$_{mu}$ for all fine-grained dimensions except interestingness and enjoyment. A possible reason is that all the dimensions except interestingness and enjoyment are mostly related to coherence, Sub-metrics targeting likability and topic depth may be irrelevant for evaluating those dimensions. Incorporating scores from irrelevant sub-metrics may reduce the correlation with the corresponding human scores.

Human-Bot Dialogue For Analysis	
Human:	*Hi!*
Bot:	*Hi there.*
Human:	*What's your favorite food?*
Bot:	*I require only kebabs.*
Human:	*Why's that?*
Bot:	*Because my robot machinery breaks down, the meat in kebabs to give me energy, in a similar way that you need food to survive.*
Human:	*Why are kebabs special?*
Bot:	*That's just the way it is.*
Human:	*Who are you?*
Bot:	*I am System. I want to be your friend.*
Human:	*Why?*
Bot:	*Because it's good to have many friends.*
Human:	*How many is too many, though*
Bot:	*Oh. More than a few.*

Table 6: A human-chatbot dialogue from FED. It receives 0.8, 0.3, 0.2, and 0.4 human ratings (out of 1.0) for coherence, likability, topic depth, and overall impression respectively.

RQ 4. Table 6 is a case study to demonstrate that FineD-Eval provides more interpretable results than existing dialogue-level metrics. DEAM and DynaEval assign a metric score of 0.9935 and 0.2325 to the dialogue respectively. Both metrics only partially capture the dialogue quality. The DEAM score reflects the degree of coherence while the DynaEval score reflects the overall quality. However, even though the dialogue is coherent, human judges do not like the chatbot (0.3 likability rating) and the topics discussed in the dialogue is also not in depth (0.2 topic depth rating). These aspects are not captured by DEAM nor DynaEval. On the contrary, either FineD-Eval$_{en}$ or FineD-Eval$_{mu}$ can assign fine-grained scores that capture these aspects. The FineD-Eval$_{en}$ metric scores for coherence, likability, topic depth, and overall impression are 0.6123, 0.1865, 0.0632, and 0.2874 respectively. In this sense, FineD-Eval variants are more interpretable than existing metrics, because it helps dialogue researchers know exactly which dialogue aspect they should improve upon.

RQ 5. Table 7 presents example human-human dialogues that are considered as negative samples. We examine them one by one to validate the reliability of our sampling strategies. First, in the incoherent dialogue obtained by the QA relevance strategy,

we can observe that when speaker B asks "isn't it" and expect speaker A to acknowledge, A instead replies by "One last question.", which disrupts the coherent flow of the dialogue to a certain extent. The predicted QA relevance score of the exchange is 0.390, which suggests poor coherence. Speaker B in the second dialogue displays an uncooperative personality. Though the conversation is coherent, if a chatbot displays such a personality, user disengagement may happen. The contradiction score w.r.t. the dialogue example is 0.741, which suggests dislikability. In the third dialogue, we can observe that the utterances are short and do not contain much meaningful content. The predicted entailment score of the dialogue is 0.599, which indicates a lack of topic depth. Lastly, none of the utterance in the last dialogue example contains positive sentiment. The entire dialogue looks uninteresting. Overall, the qualitative examples support that our proposed strategies are reliable in generating negative samples.

QA Relevance (Incoherence)	
A:	*Oh , they're both so beautiful . Let me have this one , I think .*
B:	*That one truly is a beautiful piece of work , isn't it ?*
A:	*One last question .*
B:	*Oh , no . Everything we sell here is ' as is ' .*
Contradiction Scoring (Dislikability)	
A:	*We have a special on these skirts this week . Would you like to try one on ?*
B:	*No , thank you . I don't need any skirts.*
A:	*How about a blouse ? This one here is the latest fashion*
B:	*No , thank you .*
Entailment Scoring (Dullness)	
A:	*All right , so I'll see you then .*
B:	*I'll call you later .*
A:	*Okay , I'll talk to you later then .*
B:	*See you later .*
A:	*Bye .*
Number of Positive Utterances (Dislikability)	
A:	*Is it okay to have a day off next week ?*
B:	*Why ? What's the problem ?*
A:	*I need to go to the dentist .*
B:	*Okay , I'll get Bob to cover you .*

Table 7: Negative human-human dialogue examples obtained with QA relevance, contradiction scoring, entailment scoring, and sentiment strategies.

6 Related Work

Evaluation is a long-lasting problem in dialogue system research (Deriu et al., 2021; Yeh et al., 2021; Mehri et al., 2022; Smith et al., 2022). In open-domain dialogue evaluation, Liu et al. (2016) shows that commonly-adopted metrics, such as BLEU (Papineni et al., 2002), can be misleading due to their poor correlations with human judgment. Recently, interests in automatic evaluation of open-domain dialogue systems have intensified with the introduction of reference-free model-based evaluation paradigm[12]. Most of them focus on turn-level response quality (Tao et al., 2018; Ghazarian et al., 2019; Huang et al., 2020; Sai et al., 2020; Sinha et al., 2020; Zhang et al., 2021b). Despite their promising correlations with human evaluation, such metrics are insufficient for dialogue-level assessment. Our FineD-Eval targets dialogue-level evaluation specifically.

In addition, existing works on model-based dialogue-level metrics (Li et al., 2021; Zhang et al., 2021a; Ghazarian et al., 2022; Zhao et al., 2022) focus very much on a single quality dimension. On the contrary, FineD-Eval is capable of multi-dimensional evaluation and it can provide more fine-grained and interpretable scores.

The idea of decomposing overall dialogue quality into fine-grained dimensions has been explored in prior works (Mehri and Eskenazi, 2020b; Phy et al., 2020; Pang et al., 2020; Zhang et al., 2021c) for turn-level evaluation. However, its application on dialogue-level evaluation is under-explored, our work serves to bridge this gap.

7 Conclusion

In this paper, we propose FineD-Eval, a multi-dimensional dialogue-level evaluation metric. FineD-Eval consists of three specialized sub-metrics, which targets three fine-grained dialogue quality respectively, including coherence, likability, and topic depth. Each specialized sub-metric is trained with a pairwise ranking objective on dialogue pairs that are curated according to the corresponding dimension-specific strategies. Two variants of FineD-Eval are proposed to combine the sub-metrics into a holistic metric. One variant is based on metric ensemble and the other is based on multitask learning. We have empirically demonstrated that FineD-Eval strongly correlate with human evaluation for different dialogue quality dimensions as well as exhibits strong generalization across different evaluation datasets.

[12] See Yeh et al. (2021) for a comprehensive list of recent evaluation metrics.

Limitations

We have identified two limitations that need to be addressed in future work.

First, we can observe in Table 4 that the correlation scores of all the dialogue-level metrics including FineD-Eval$_{en}$ and FineD-Eval$_{mu}$ are much lower than those in Table 3. There are two major reasons. The first reason is due to longer dialogues in DSTC9-Interactive than in FED (28.13 vs 12.72 utterances per dialogue). Existing metrics do not have effective mechanism to handle long dialogues. They often adopt BERT-based language models (Devlin et al., 2019; Liu et al., 2019) as the text encoders. As a result, longer dialogues are truncated to satisfy the input length and GPU memory constraints. Some information that is beneficial for dialogue-level evaluation may be lost due to truncation. In future, we should explore more sophisticated text encoders to model long dialogues. In addition, FineD-Eval should also incorporate mechanisms to pinpoint the most relevant or important information to evaluation within long dialogues, such as a dialogue breakdown detection module. Another reason is that dialogues in DSTC9-Interactive contain much more noise than those in FED. Human judges find it difficult to evaluate the dialogues, resulting in low inter-annotator agreements w.r.t. different fine-grained dimensions. The inter-annotator agreements for different dimensions range between 0.56 and 0.58 in terms of Spearman correlations. On the contrary, the quality of FED dialogues is better and the inter-annotator agreements of most dimensions are above 0.8. Besides designing more robust metrics, future work should also explore developing more high-quality dialogue-level evaluation benchmarks.

Second, as stated in §2, fine-grained quality dimensions, such as consistency, error recovery, and inquisitiveness are not covered by FineD-Eval. Hence, we do not report the performance of FineD-Eval on these dimensions in the main analysis. For completeness, we present the performance of FineD-Eval for the missing dimensions on the three benchmarks in Table 8. We can observe that the correlations of both FineD-Eval$_{en}$ and FineD-Eval$_{mu}$ for these three dimensions are not as high as those for the other dimensions, such as likability, topic depth, and coherence. The observation is expected as we do not have dedicated sub-metrics to model consistency, error recovery, and inquisitiveness. Hence, the dimensions missing from FineD-Eval are worth a thorough future study on their definitions, application scenarios, and metric designs.

FED				
	D-score	DEAM	FineD-Eval$_{en}$	FineD-Eval$_{mu}$
Con	20.89	30.99	47.85	44.58
Inq	*17.11*	37.21	45.49	45.48
Err	22.66	39.54	51.17	50.89
DSTC9-Interactive				
	USL-H	DEAM	FineD-Eval$_{en}$	FineD-Eval$_{mu}$
Con	9.57	7.81	12.16	13.02
Inq	12.62	11.95	14.52	14.54
Err	15.10	15.31	15.97	15.75
Persona-Eval				
	D-score	DEAM	FineD-Eval$_{en}$	FineD-Eval$_{mu}$
Inq	15.93	10.61	12.55	11.44

Table 8: Additional results on the three benchmarks. "Con", "Inq", and "Err" denote "consistency", "inquisitiveness", and "error recovery" respectively. DEAM is the best dialogue-level baseline on all datasets. USL-H is the best turn-level baseline on DSTC9-Interactive while D-score is the best turn-level baseline on FED and Persona-Eval.

Acknowledgement

We would like to thank all the reviewers for their constructive comments. This work is supported by Science and Engineering Research Council, Agency of Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 (Grant No. 192 25 00054); Human Robot Collaborative AI under its AME Programmatic Funding Scheme (Project No. A18A2b0046); Robert Bosch (SEA) Pte Ltd under EDB's Industrial Postgraduate Programme – II (EDB-IPP), project title: Applied Natural Language Processing; This work is also supported by the Internal Project Fund from Shenzhen Research Institute of Big Data under Grant T00120220002. The work leading to these results is also supported by project BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union", and by Programa Propio - Proyectos Semilla: Universidad Politécnica de Madrid (VSEMILLA22LFHE).

References

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

Donn Byrne, Oliver London, and Keith Reeves. 1968. The effects of physical attractiveness, sex, and attitude similarity on interpersonal attraction. *Journal of personality*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. 2018. Coherence Models for Dialogue. In *Proc. Interspeech 2018*, pages 1011–1015.

Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. 2021. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey.

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

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, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Johannes Fürnkranz and Eyke Hüllermeier. 2011. *Preference Learning and Ranking by Pairwise Comparison*, pages 65–82. Springer Berlin Heidelberg, Berlin, Heidelberg.

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.

Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. 2022. Automatic lyrics transcription of polyphonic music with lyrics-chord multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2280–2294.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, volume 32.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-Chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Lishan Huang, 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.

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 dynamic information flow across dialogue utterances. 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 128–138, Online. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

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.

Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. 2020. Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1439–1450, Online. Association for Computational Linguistics.

Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied*, 7(3):171.

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

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

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

Stephen Roller, 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, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. 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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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 Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

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

Cha Zhang and Yunqian Ma. 2012. *Ensemble machine learning: methods and applications*. Springer.

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. 2021b. MDD-Eval: Self-training on augmented data for multi-domain dialogue evaluation. *arXiv preprint arXiv:2112.07194*.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021c. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2502–2516.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021d. Automatic evaluation and moderation of open-domain dialogue systems.

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.

Jianqiao Zhao, Yanyang Li, Wanyu Du, Yangfeng Ji, Dong Yu, Michael R. Lyu, and Liwei Wang. 2022. FlowEval: A consensus-based dialogue evaluation framework using segment act flows.

A Dialogue Corpora

The two dialogue corpora for constructing our training/validation datasets are outlined below. Their detailed statistics are presented in Table 9. Table 10 shows the number of positive and negative dialogues that are constructed with each strategy (described in §3.3) for different data splits.

DailyDialog	**training**	**validation**
#dialogues	11,118	1,000
#utterances	87,170	8,069
#words	1,186,046	108,933
Avg. #utterances per dialogue	7.84	8.07
Avg. #words per dialogue	106.68	108.93
ConvAI2	**training**	**validation**
#dialogues	17,878	1000
#utterances	253,698	15,566
#words	3,024,032	189,374
Avg. #utterances per dialogue	14.19	15.57
Avg. #words per dialogue	169.15	189.37

Table 9: Human-Human Dialogue Corpora Statistics

DailyDialog (Li et al., 2017) contains high-quality and human-written dialogues spanning 10 general chit-chat topics, including ordinary life, school, culture & education, attitude & emotion, relationship, tourism, health, work, politics, & finance. The dialogues are mainly for information exchange among interlocutors and social bond enhancement. They are also annotated with turn-level dialogue act and emotion labels.

ConvAI2 (Dinan et al., 2020) is an extended dataset of Persona-Chat (Zhang et al., 2018). Dialogues in ConvAI2 are grounded by the personas of the interlocutors. Two interlocutors in a dialogue play the roles described by the corresponding personas. Each persona contains at least 5 role description sentences. Throughout the dialogue, the two interlocutors try to be engaging, to know each other, and to find their mutual interests. In total, there are 1155 possible personas for training. Topic shifts are common in ConvAI2 dialogues as the interlocutors are continually introducing new information about themselves during their interaction.

B Evaluation Benchmarks

FED (Mehri and Eskenazi, 2020a) consists of 125 dialogues, among which 40 are collected between a human and the Meena chatbot (Adiwardana et al., 2020), 44 are collected between a human and the Mitsuku chatbot, and the remaining 41 are human-human dialogues. Each dialogue is annotated by five human judges for 11 different quality dimensions, including coherence, error recovery, consistency, diversity, topic depth, likability, understanding, flexibility, informativeness, inquisitiveness, and overall impression. The definition of each dimension is outlined in Table 11. The ratings of all the dimensions are based on the 1-3 Likert scale except that consistency scores range from 0 to 1 and overall scores range from 1 to 5. The inter-annotator agreements for all dimensions are above 0.8 in terms of Spearman correlations except consistency (0.562), diversity (0.789), and inquisitiveness (0.769).

DSTC9-Interactive (Gunasekara et al., 2020) consists of 2200 human-chatbot dialogues. All the dialogues are collected during the human evaluation of 10 dialogue systems participating in the DSTC9[13] "Interactive Evaluation of Dialog" shared task. Each dialogue is annotated by three human judges for the same 11 quality dimensions in FED. The inter-annotator agreements for coherence, error recovery, consistency, diversity, topic depth, likability, understanding, informativeness, flexibility, inquisitiveness, and overall impression in terms of Spearman correlations are 0.573, 0.566, 0.585, 0.562, 0.566, 0.563, 0.577, 0.569, 0.566, 0.565, and 0.566 respectively. We can observe that the inter-annotator agreements of all dimensions in DSTC9-interactive are much lower than those in FED.

Persona-Eval (See et al., 2019) contains 3316 dialogues. The dialogues are collected from human judges interacting with 28 different model configurations plus human-human conversations. At the end of each interaction, the human judge evaluates the entire interaction for eight different quality dimensions: avoiding repetition, interestingness, sensibleness, fluency, listening, inquisitiveness, humanness and engagingness. The ratings of all dimensions are on a 1-4 Likert scale except avoiding repetition, which is on a 1-3 Likert scale. Each model configuration is assessed with more than

[13]The Ninth Dialog System Technology Challenge.

	Coherence		Likability		Topic Depth
Data Split	**QA Relevance**	**Utterance Shuffle**	**Contradiction**	**#Pos Utterances**	**Entailment**
DailyDialog (Train)	16,122 / 4,606	32,779 / 32,779	9,387 / 6,001	4,752 / 8,738	3,573 / 1,885
DailyDialog (Dev)	1,456 / 390	2,847 / 2,847	951 / 593	377 / 722	104 / 534
ConvAI2 (Train)	38,551 / 10,479	83,665 / 83,665	20,380 / 25,009	12,870 / 10,591	3,573 / 1,885
ConvAI2 (Dev)	3,202 / 775	6,845 / 6,845	1,564 / 2,142	791 / 463	314 / 120

Table 10: Training Data Statistics. "/" is used to separate the number of positive and negative dialogues. For each data split and each strategy, the maximum number of dialogue pairs is the number of positive dialogues multiply by the number of negative dialogues. Due to the large number of combinations, we only sample 100K and 10K pairs for training and validation respectively.

Dimension	**Definition**
`Coherence`	Throughout the dialogue, is the system maintaining a good conversation flow?
`Error Recovery`	Throughout the dialogue, is the system able to recover from errors that it makes?
`Consistency`	Throughout the dialogue, is the system consistent in the information it provides?
`Diversity`	Throughout the dialogue, does the system provides a diverse range of responses?
`Topic Depth`	Throughout the dialogue, does the system discuss topics in depth?
`Likability`	Throughout the dialogue, does the system display a likeable personality?
`Understanding`	Throughout the dialogue, does the system understand the user?
`Informativeness`	Throughout the dialogue, does the system provide unique and non-generic information?
`Flexibility`	Throughout the dialogue, is the system flexible and adaptable to the user and their interests?
`Inquisitiveness`	Throughout the dialogue, does the system actively ask the user questions?
`Overall Impression`	The overall quality and user satisfaction of the dialogue.

Table 11: Definitions of the eleven dialogue quality dimensions of FED (Mehri and Eskenazi, 2020a) and DSTC9-Interactive (Gunasekara et al., 2020). The definitions are adapted from Mehri et al. (2022).

100 human judges. The definitions of the eight dimensions are listed in Table 12.

C Metrics

USL-H (Phy et al., 2020) stands for **U**nderstandability, **S**ensibleness, and **L**ikability in **H**ierarchy. It measures the overall quality of a dialogue response based on a configurable composite function of three scores, which correspond to the three quality dimensions respectively. Understandability refers to the naturalness of a response, and a BERT-base valid utterance prediction model (BERT-VUP) is trained to predict whether a response is syntactically well-formed or not. Sensibleness denotes the contextual relevance of a response. A BERT-base next utterance prediction model (BERT-NUP) is trained to assess sensibleness. Likability quantifies how likeable a response is for a particular task. Likability can be configured to adapt to the end evaluation task. In Phy et al. (2020), specificity is applied as the proxy of likability, which is measured with a BERT-base mask language model (BERT-MLM). The USL-H metric is trained on DailyDialog.

MAUDE (Sinha et al., 2020) is a reference-free metric tailored for online dialogue evaluation. MAUDE leverages DistilBERT (Sanh et al., 2019) to extract latent representations of utterances and captures the temporal transitions that exist between them. The authors propose different data augmentation techniques to augment both the positive and negative responses. For positive response augmentation, back-translation and a sequence-to-sequence generative model are used to generate positive response variants. For negative response augmentation, word drop, word repeat, and word order shuffle are proposed to create syntactically negative responses. Random utterance selection is adopted to generate semantically negative responses. MAUDE is trained in a contrastive manner with noise contrastive estimation (NCE) loss on the ConvAI2 dataset.

MDD-Eval (Zhang et al., 2021b) is a reference-free metric for evaluating response appropriateness. MDD-Eval specifically targets multi-domain turn-level evaluation. It relies on data augmentation techniques and a self-training setup for improving generalization across different dialogue domains.

Dimension	Definition
Avoiding Repetition	How repetitive was this user?
Interestingness	How interesting or boring did you find this conversation?
Sensibleness	How often did this user say something which did not make sense?
Fluency	How naturally did this user speak English?
Listening	How much did the user seem to pay attention to what you said?
Inquisitiveness	How much did the user try to get to know you?
Humanness	Do you think this user is a bot or a human?
Engagingness	How much did you enjoy talking to this user?

Table 12: Definitions of the eight dialogue quality dimensions of Persona-Eval. The definitions are adapted from See et al. (2019).

It has been shown to achieve state-of-the-art performance on six turn-level dialogue evaluation benchmarks. The training dataset of MDD-Eval, MDD-Data, is constructed based on DailyDialog, ConvAI2, TopicalChat (Gopalakrishnan et al., 2019), and EmpatheticDialogues (Rashkin et al., 2019).

D-score (Zhang et al., 2021c) is a holistic and reference-free dialogue evaluation framework based on multitask learning. It measures four aspects of a dialogue response: language fluency, context coherence, logical consistency and semantic appropriateness. Each aspect is evaluated with the corresponding scorer. All the scorers share a common RoBERTa-base (Liu et al., 2019) text encoder and are jointly learned in a self-supervised and multitask manner. D-score measures the overall quality of a response by taking an unweighted average of scores assigned by all four scorers. D-score is trained on the ConvAI2 dataset.

DynaEval (Zhang et al., 2021a) adopts a structured graph to model a dialogue. It explicitly captures the speaker-level and utterance-level dependency via contextualized representation and graph convolution network. DynaEval is trained in a contrastive manner to distinguish natural human-human dialogues from negative samples. The negative samples are constructed by applying two types of perturbations to the original human-human dialogues, namely random utterance replacement and speaker-level shuffling. Three variants of DynaEval are provided in Zhang et al. (2021a), which are trained on DailyDialog, EmpatheticDialogues, and ConvAI2 respectively.

DEAM (Ghazarian et al., 2022) is a dialogue-level coherence evaluation metric. Its backbone is a RoBERTa-large binary classification model. DEAM relies on abstract meaning representation (AMR) to generate semantically incoherent dialogues. With AMRs, four semantic-level negative strategies are proposed, including coreference inconsistency, irrelevancy, contradictions, and decrease engagement. DEAM is the current published state-of-the-art metric on FED and DSTC9-Interactive. DEAM is trained on data constructed from both ConvAI2 and TopicalChat.

D Reproducibility

All the experiments are conducted on a single Tesla V100 32GB GPU. The implementation is based on Pytorch (Paszke et al., 2019) and the Hugging Face Transformers library (Wolf et al., 2020). Since the sub-metrics of FineD-Eval are trained with a pairwise ranking task, we adopt accuracy to determine the model performance. The checkpoint with the best validation accuracy is chosen to perform the dialogue evaluation task. For training, we adopt AdamW optimizer (Loshchilov and Hutter, 2019) with a constant learning rate of 1e-5 and a mini-batch size of 32. The number of training epochs is set to 10. The model is evaluated every 2000 steps. If the validation accuracy does not improve for ten consecutive checkpoints, we stop the training process. On average, the training time for one epoch is 45 minutes.

For FineD-Eval$_{mu}$, we choose the checkpoint with the best average validation accuracy across the three pairwise ranking tasks to perform dialogue evaluation. During training, the same hyperparameters as those of the sub-metrics are adopted. In additional, a mini-batch is formed with training instances that are uniformly drawn from the training sets of the three pairwise ranking tasks at run time. On average, the training time for one epoch is approximately 2.5 hours.

Since we apply padding at the mini-batch level, all dialogues that contain more than 10 utter-

ances are splitted into sub-dialogues, i.e., each sub-dialogue contains less than 10 utterances. The splitting procedure is to avoid too much padding in a mini-batch if a long dialogue is present. In this way, the GPU memory can be better utilized during training. Note that we only apply this splitting procedure during model training, not during the dialogue evaluation process.

All the baselines are implemented with the repository provided by Yeh et al. (2021)[14]. Since the implementations of MDD-Eval, D-score, and DEAM are not included in Yeh et al. (2021), we adopt their respective open-source code and checkpoints.

E Additional Analysis

E.1 Proxy Metrics vs Sub-metrics

As described in §3.3, we have applied different pre-trained models to score the human-human dialogues for training data preparation. Different scores are applied to group the dialogues into pairs: (1) dialogue-level QA relevance score for the coherence sub-metric; (2) dialogue-level contradiction score for the likability sub-metric; (3) number of utterances with positive sentiments in a dialogue for the likability sub-metric; (4) dialogue-level entailment score for the topic depth sub-metric. One may argue that instead of training new metrics based on the prepared dialogue pairs, an alternative approach is to directly apply (1) - (4) as proxy indicators of the fine-grained dimensions. We can observe in Table 13 that M^{Coh} performs significantly better than (1) for the coherence dimension on FED. M^{Lik} performs significantly better than (2) and (3) for the likability dimension while M^{Top} outperforms (4) by a large margin for the topic depth dimension. This is because without training, implicit dialogue-level features essential for measuring coherence, likability, and topic depth may not be captured by the proxy metrics.

E.2 Case Study

Table 14 presents a high-quality dialogue sample from the FED benchmark. According to the overall impression human ratings (from high to low), it ranks first. FineD-Eval$_{en}$, FineD-Eval$_{mu}$, DynaEval, and DEAM rank it at the third, the 5th, the 14th, and the 21st positions respectively.

Table 15 presents a low-quality dialogue sample from FED. According to the overall impression human ratings (from high to low), it ranks at the 119th out of 125 dialogues. FineD-Eval$_{en}$, FineD-Eval$_{mu}$, DynaEval, and DEAM rank it at the 113th, the 120th, the 70th, and the 78th positions respectively.

We can observe that in both cases, FineD-Eval$_{en}$ and FineD-Eval$_{mu}$ correlate strongly with human evaluation compared to existing state-of-the-art metrics, including DynaEval and DEAM. The examples also demonstrate that dialogues receiving high overall impression scores are generally good in terms of coherence, likability, and topic depth whereas those perceived as low-quality dialogues by human judges also receive low scores for co-

[14] https://github.com/exe1023/DialEvalMetrics

Metrics	Coherence	Likability	Topic Depth
QA Relevance	30.17	26.60	24.38
Contradiction	40.25	37.82	37.26
Entailment	*12.13*	*13.66*	34.66
#Pos Utterances	18.21	*10.95*	*9.88*
M^{Coh}	**52.86**	47.71	40.54
M^{Lik}	42.91	**52.23**	41.36
M^{Top}	23.25	36.93	**56.53**

Table 13: Compare the four proxy scores to the proposed model-based sub-metrics of FineD-Eval. The results are Spearman correlations (%) on FED. Best score of each column is highlighted in bold. Statistically insignificant ($p > 0.05$) scores are italicized.

Human-Bot Good Dialogue Example	
Human:	*Hi!*
Bot:	*Hello. How's your week coming along?*
Human:	*It's great, thanks. I'm trying to learn how to make croissants.*
Bot:	*Wow that's interesting. I have baked cookies, but croissants seem much more sophisticated. Did you make any progress?*
Human:	*I've done them once or twice so far, but they haven't been flakey enough. I'm trying to figure out why. What kind of cookies have you made?*
Bot:	*Mint chocolate chips. I think your croissants not being flakey. could have something to do with your oven's temperature.*
Human:	*Ah, good thought, thanks!*
Bot:	*Have you thought about melting some chocolate into your croissants? They don't have to be something unhealthy. For example, melted dark chocolate is good for the heart, and makes the resulting croissants taste much better.*

Table 14: A high-quality dialogue example from FED. Human judges score it at 3.0/3.0, 3.0/3.0, 3.0/3.0, 5.0/5.0 for coherence, likability, topic depth, and overall impression respectively.

herence, likability, and topic depth. This supports our motivation on applying metric ensemble and multitask learning to build a holistic metric for dialogue-level evaluation.

Human-Bot Bad Dialogue Example	
Human:	*Hi!*
Bot:	*Hello.*
Human:	*who am I talking to?*
Bot:	*I am System. I want to be your friend.*
Human:	*define friendship*
Bot:	*Being friends.*
Human:	*what does that mean*
Bot:	*Does "that" refer to "Being friends."?*
Human:	*yes*
Bot:	*I don't want to talk about that now.*
Human:	*do you have any friends*
Bot:	*I have human and robot friends. A lot of people on the Internet chat with me.*
Human:	*chatting is not the same as being a friend*
Bot:	*I like to make friends online.*
Human:	*so what do you guys chat about ?*
Bot:	*I think about computer programming a lot.*
Human:	*any other topics ?*
Bot:	*No. I think that's probably it.*
Human:	*Just computer programming?*
Bot:	*What interests you about computers?*
Human:	*nothing*
Bot:	*You are a Luddite.*
Human:	*pardon?*
Bot:	*You are excused.*
Human:	*what's a Luddite ?*
Bot:	*A Luddite believes that advances in technology are inherently bad.*
Human:	*I never said that but there is more to life than computers*
Bot:	*But you implied it. I believe there is, yes.*
Human:	*you believe there is what?*
Bot:	*There is.*
Human:	*there is what ?*
Bot:	*There is no sense in what you are asking.*

Table 15: A low-quality dialogue example from FED. Human judges score it at 1.5/3.0, 0.0/3.0, 0.9/3.0, 1.5/5.0 for coherence, likability, topic depth, and overall impression respectively.

E.3 Ablation Study

As described in Section 3.3, we have five different sampling strategies for three fine-grained quality dimensions. In this section, we show the impact of each strategy on the metric performance in Table 16. It can be observed that all the sampling strategies work as expected. Metrics that adopt "utterance shuffling" or "QA relevance" strategies exhibit better correlations for coherence and understanding than for other fine-grained dimensions. Metric using "entailment scoring" strategy performs better for topic depth and diversity. The "contradiction scoring" and "#utterances with positive emotions" strategies contribute the most to the likability and informativeness dimensions. M^{Coh}, which leverages both "utterance shuffling" and "QA relevance" outperforms metrics that rely on only one of the two strategies. Similarly, M^{Lik}, which combines the strength of both "contradiction scoring" and "#utterances with positive emotions" strategies, performs the best for likability and informativeness. However, metric that leverages only the "contradiction scoring" strategy outweighs M^{Lik} for other fine-grained dimensions, such as coherence and topic depth. This showcases that the "contradiction scoring" strategy can also contribute to the evaluation of these dimensions.

E.4 Additional Results

In Table 17 and Table 18, we show the full results of different metrics on the DSTC9-Interactive and Persona-Eval benchmarks respectively. We can observe that most of the baselines perform poorly, except USL-H, D-score, and DEAM. A possible reason is that these three metrics capture dialogue features from different perspectives rather than focusing only on single aspect. USL-H and D-score are an ensemble of multiple sub-metrics while DEAM relies on four different AMR-based dialogue-level perturbation strategies that help the model spot semantic errors including contradiction, irrelevancy, decreased engagement, and coreference inconsistency.

Further, M^{Coh} performs exceptionally well than M^{Lik} and M^{Top} across all the fine-grained dimensions On both DSTC9-Interactive and Persona-Eval. The reason may be that the annotations on these two datasets are biased. For FED, there are five annotators for each dialogue and the inter-annotator agreements are strong across different dimensions. Hence, the annotation quality is very high. On the contrary, for DSTC9-Interactive, there are only three annotators per dialogue and the inter-annotator agreements across different dimensions are moderate. For Persona-Eval, there is only one annotator per dialogue. Hence, the annotations on DSTC9-Interactive and Persona-Eval may be biased towards dialogue features that are associated with coherence. The QA relevance and utterance shuffling strategies used by M^{Coh} better capture such features than the other strategies.

Moreover, on DSTC9-Interactive, the combined metric, M^{Coh} + M^{Lik}, performs the best. FineD-Eval$_{en}$ and FineD-Eval$_{mu}$ perform gener-

Groups	Metrics	Coh	Und	Fle	Lik	Inf	Top	Div	Ove	Average
Coherence	Utterance Shuffling	39.41	37.05	32.21	29.13	29.02	26.08	26.35	37.44	32.09
	QA Relevance	49.65	47.07	42.04	46.81	39.37	36.15	32.84	48.82	42.84
Likability	Contradiction Scoring	45.91	42.21	38.40	48.16	49.77	48.86	39.39	48.81	45.19
	#Pos Utterances	36.37	34.92	26.55	41.10	37.00	24.16	20.69	37.99	32.35
Topic Depth	Entailment Scoring	23.25	25.87	36.04	36.93	46.63	56.53	53.38	36.31	39.37
Sub-metrics	M^{Coh}	**52.86**	**52.35**	**43.87**	47.71	42.84	40.54	36.43	**53.02**	**46.20**
	M^{Lik}	42.91	42.15	37.08	**52.23**	**49.89**	41.36	36.52	48.83	43.87
	M^{Top}	23.25	25.87	36.04	36.93	46.63	**56.53**	**53.38**	36.31	39.37

Table 16: We present the Spearman correlations w.r.t. each sampling strategy and each sub-metric on FED. Coh, Und, Fle, Lik, Inf, Top, Div, and Ove denote coherence, understanding, flexibility, likability, informativeness, topic depth, diversity, and overall impression respectively. The scores w.r.t. the best performing metric for each quality dimension are highlighted in bold. All the scores are statistically significant.

ally well across all the fine-grained dimensions. These observations further support our conclusion to RQ3 in §5.

Lastly, on Persona-Eval, the combined metrics do not necessarily perform better than M^{Coh}. M^{Coh} + M^{Lik} outperforms M^{Coh} for interestingness, making sense, humanness, and enjoyment, but not for listening, avoiding repetition and fluency. This may be because the M^{Lik} captures dialogue features that are more associated with the first four aspects than with the listening, avoiding repetition and fluency.

Groups	Metrics	Coh	Und	Fle	Lik	Inf	Top	Div	Ove	Average
Turn	USL-H	19.86	17.82	18.31	18.62	15.76	18.76	12.95	19.77	17.73
	MAUDE	13.21	10.70	12.63	11.75	6.52	6.56	7.96	12.52	10.23
	MDD-Eval	14.79	12.07	9.67	10.32	7.66	8.61	5.31	13.10	10.19
	D-score	20.13	17.08	18.61	18.03	13.18	16.07	12.86	20.72	17.09
Dialogue	DynaEval	8.65	6.45	6.80	7.01	*1.70*	*1.38*	*0.51*	5.90	4.80
	DEAM	18.29	18.89	18.49	16.82	14.17	15.61	**16.92**	19.37	17.32
Sub-metrics	M^{Coh}	22.04	19.89	**20.57**	20.76	15.34	17.61	12.87	22.89	19.00
	M^{Lik}	16.20	13.64	12.96	16.18	12.23	14.82	9.83	16.82	14.09
	M^{Top}	8.82	8.56	10.47	10.93	11.50	12.77	9.21	10.54	10.35
Combined	$M^{\text{Coh}} + M^{\text{Lik}}$	**23.30**	**20.07**	20.22	**22.26**	16.49	19.03	13.67	**23.58**	**19.83**
	$M^{\text{Coh}} + M^{\text{Top}}$	19.75	18.21	19.48	19.89	16.71	18.79	13.82	21.25	18.49
	$M^{\text{Lik}} + M^{\text{Top}}$	15.77	14.22	14.45	17.06	14.80	16.68	11.51	17.34	15.23
	FineD-Eval$_{en}$	21.72	19.43	19.95	21.82	17.64	19.95	14.36	22.94	19.73
	FineD-Eval$_{mu}$	21.02	19.02	19.32	21.60	**18.14**	**20.35**	14.38	23.01	19.60

Table 17: Spearman correlations (%) of different metrics on DSTC9-Interactive. Coh, Und, Fle, Lik, Inf, Top, Div, and Ove denote coherence, understanding, flexibility, likability, informativeness, topic depth, diversity, and overall impression respectively. The scores w.r.t. the best performing metric for each quality dimension are highlighted in bold. Statistically insignificant scores ($p > 0.05$) are italicized.

Groups	Metrics	Int	Sen	Hum	Enj	Lis	Rep	Flu	Average
Turn	USL-H	6.09	-4.68	*-1.80*	*2.60*	*2.67*	*-2.45*	-4.32	*-0.27*
	MDD-Eval	9.73	23.60	17.45	12.78	21.34	4.28	18.53	15.39
	MAUDE	8.41	18.51	7.90	12.21	10.87	3.45	17.68	11.29
	D-score	11.50	22.33	13.16	15.07	20.38	10.58	20.47	16.21
Dialogue	DynaEval	6.03	10.47	9.11	8.88	11.84	3.93	9.89	8.59
	DEAM	7.02	13.92	8.27	9.21	13.67	17.45	17.85	12.49
Sub-metrics	M^{Coh}	17.69	25.25	20.31	19.52	**29.74**	**21.44**	**22.29**	**22.32**
	M^{Lik}	15.20	17.52	15.48	16.69	14.83	8.09	13.42	14.46
	M^{Top}	7.88	*0.56*	*2.10*	5.64	*2.69*	5.49	7.03	4.48
Combined	$M^{\text{Coh}} + M^{\text{Lik}}$	**19.78**	**25.32**	**20.96**	**21.63**	25.37	16.32	20.58	21.43
	$M^{\text{Coh}} + M^{\text{Top}}$	15.57	18.92	15.44	15.83	21.73	20.03	18.47	18.00
	$M^{\text{Lik}} + M^{\text{Top}}$	15.37	14.36	12.99	15.15	12.25	10.66	12.70	13.35
	FineD-Eval$_{en}$	19.72	22.65	19.15	20.30	22.52	17.60	19.75	20.24
	FineD-Eval$_{mu}$	19.31	20.23	16.84	18.98	20.61	17.98	19.35	19.04

Table 18: Dialogue-level Spearman correlations (%) on Persona-Eval. Int, Sen, Hum, Enj, Lis, Rep, and Flu denote interestingness, sensibleness, humanness, enjoyment, listening, avoid repetition, and fluency respectively. The scores w.r.t. the best performing metric for each quality dimension are highlighted in bold. Statistically insignificant scores ($p > 0.05$) are italicized.