PoE: A Panel of Experts for Generalized Automatic Dialogue Assessment

Chen Zhang, *Student Member, IEEE*, Luis Fernando D'Haro, *Member, IEEE*, Qiquan Zhang, *Member, IEEE*, Thomas Friedrichs, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Chatbots are expected to be knowledgeable across multiple domains, e.g. for daily chit-chat, exchange of information, and grounding in emotional situations. To effectively measure the quality of such conversational agents, a model-based automatic dialogue evaluation metric (ADEM) is expected to perform well across multiple domains. Despite significant progress, existing ADEMs tend to perform well only on data that are similar to its training data (overfit to its training domain). This calls for a domain-generalized metric that can assess dialogues of different characteristics. To this end, we propose a *Panel of Experts* (PoE), a multitask network that consists of a shared transformer encoder and a collection of lightweight adapters. The shared encoder captures the general knowledge of dialogues across domains, while each adapter specializes in one specific domain and serves as a domain expert. To validate the idea, we construct a high-quality multi-domain dialogue dataset leveraging data augmentation and pseudo-labeling. The PoE network is comprehensively assessed on 16 dialogue evaluation datasets spanning a wide range of dialogue domains. It achieves state-of-the-art performance in terms of mean Spearman correlation over all the evaluation datasets. It exhibits better zero-shot generalization than existing state-of-the-art ADEMs and the ability to easily adapt to new domains with few-shot transfer learning.**

***Index Terms*—Adapters, automatic dialogue evaluation, multi-domain generalization, multitask learning.**

Manuscript received 22 March 2022; revised 18 October 2022 and 1 February 2023; accepted 18 February 2023. Date of publication 1 March 2023; date of current version 22 March 2023. This work was supported in part by the 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 under Grant 192 25 00054, in part by the Human Robot Collaborative AI through AME Programmatic Funding Scheme under Grant A18A2b0046, in part by the Robert Bosch (SEA) Pte Ltd through EDB's Industrial Postgraduate Programme – II (EDB-IPP), Project title: Applied Natural Language Processing, in part by the National Natural Science Foundation of China under Grant 62271432, in part by the Internal Project Fund from Shenzhen Research Institute of Big Data under Grant T00120220002, in part by the European Commission through Project ASTOUND under Grant 101071191 – HORIZON-EIC-2021-PATHFINDERCHALLENGES-01, and in part by Programa Propio - Proyectos Semilla: Universidad Politécnica de Madrid under Grant VSEMILLA22LFHE. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhijian Ou. *(Corresponding author: Chen Zhang.)*

Chen Zhang and Qiquan Zhang are with the Human Language Technology Group at Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: chen_zhang@nus.edu.sg; qiquanzh@nus.edu.sg).

Luis Fernando D'Haro is with the Speech Technology Group, Universidad Politécnica de Madri, 28040 Madrid, Spain (e-mail: luisfernando.dharo@upm.es).

Thomas Friedrichs is with the Robert Bosch (SEA) Pte Ltd, Singapore 573943 (e-mail: thomas.friedrichs@sg.bosch.com).

Haizhou Li is with the Shenzhen Research Institute of Big Data, School of Data Science, Chinese University of Hong Kong, Shenzhen 518172, China, also with the University of Bremen, 28359 Bremen, Germany, also with the National University of Singapore, Singapore 119077, and also with the Kriston AI, Xiamen 361005, China (e-mail: haizhouli@cuhk.edu.cn).

Digital Object Identifier 10.1109/TASLP.2023.3250825

I. Introduction

THE research advancement on open-domain dialogue systems, a.k.a. chatbots is guided by evaluation. The evaluation of chatbots is a complex task as the conversations carried out by chatbots can be about any topic, and of very different characteristics, such as daily chit-chat [1], knowledge exchange [2], emotion disclosure [3], and personal interests [4]. Especially, as chatbots are increasingly expected to perform in multiple domains [5], [6], the corresponding evaluation methods ought to be equally versatile. While human judges have no issue in assessing such a wide range of topics given proper instructions, it is too costly to perform human evaluation at every stage of system development [7]. This prompts us to develop an automatic evaluation metric, which highly correlates with human evaluation under different evaluation scenarios.

Recently, there is a rising interest in model-based reference-free automatic dialogue evaluation metrics (ADEMs), that has advantage over the commonly used reference-based untrained metrics such as BLEU [8] and F-score, which are shown to correlate poorly w.r.t. human evaluations [9]. Most of the reference-free ADEMs are trained on human-human dialogue corpora in a weakly supervised fashion. Specifically, a model is trained to classify a dialogue response as either positive or negative given its dialogue context,[1] which consists of several consecutive utterances from a human-human dialogue. During training, a true dialogue response in the context is considered as a positive response, whereas the negative responses are obtained via different semantic or syntactic perturbation strategies [10], [11]. During inference, the trained model is used to score responses given their dialogue contexts.

The recent model-based ADEMs [7], [10], [11], [12], [13], [14] have demonstrated a strong correlation with human evaluation on different dialogue evaluation datasets. However, their generalizability across different dialogue domains is questionable. A recent survey [15] shows that the state-of-the-art ADEMs obtain fair in-domain performance, i.e., good correlations on dialogue data similar to their training data. However, when evaluated on evaluation data different from their training data, the ADEMs tend to perform poorly. This issue has been frequently

[1]Sentences irrelevant, semantically inappropriate or incoherent w.r.t. a dialogue context can serve as negative responses.

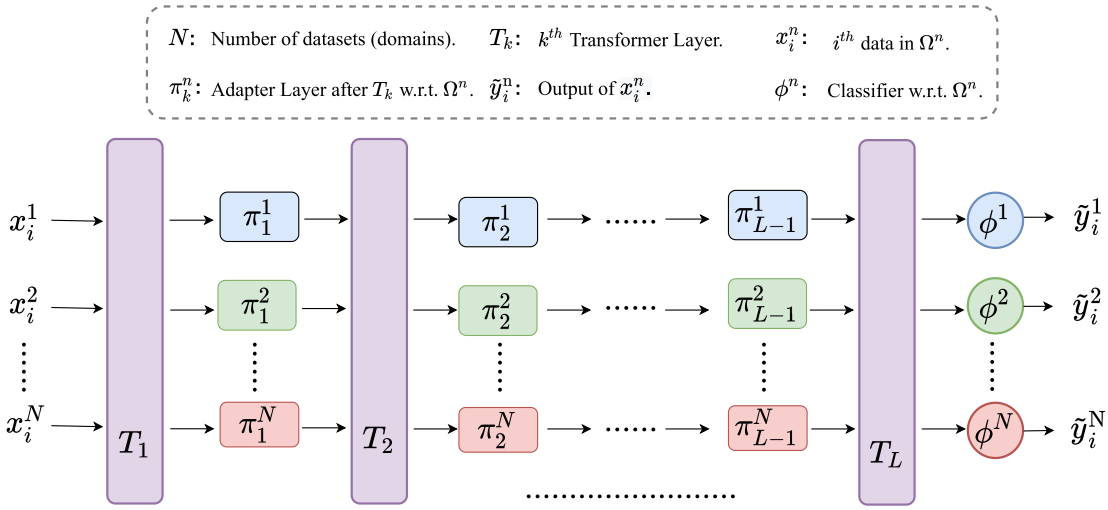

Fig. 1. System architecture of a Panel of Experts (PoE). A transformer encoder T consists of L layers (purple rectangles). Different colors (blue, red, and green) denote domain-specific modules. $\{\pi^n\}_{n=1}^{|N|}$ are the N different domain-specific adapters. Each domain-specific π^n has $L-1$ adapter layers, $\{\pi_1^n, \pi_2^n, \ldots, \pi_{L-1}^n\}$, that are injected in between every two consecutive transformer layers. $\{\phi^n\}_{n=1}^{|N|}$ are the domain-specific classifiers after the final transformer layer, T_L. T is shared by all the domain-specific modules.

raised in recent works on dialogue evaluation [16], [17], [18]. In addition, most ADEMs only consist of a single network, for example, a pre-trained transformer encoder, such as BERT [19] or RoBERTa [20] with a classification layer on top. They don't employ a specific mechanism for domain generalization. Hence, it is believed that an adequate network architecture is required for multi-domain dialogue evaluation.

We propose a network architecture as a single-model metric with a mechanism to handle domain generalization. The network makes a unified decision with multiple domain specific experts, thus is referred to as a Panel of Experts (PoE). It consists of a pre-trained transformer encoder [19], [20] and a set of adapters [21] (Fig. 1) which is shared across domains. The adapters are lightweight task-specific modules interleaved between the layers of the pretrained transformer encoder. Each adapter serves as a domain expert in evaluating a specific category of dialogues. PoE is also flexible when performing out-of-domain evaluation tasks. For instance, we can either average the prediction scores of all adapters (late fusion) or average the parameters of the adapters to derive a single adapter (early fusion) for decision making. Furthermore, we may adapt PoE to new domain with few-shot transfer learning, without the need of full model training or finetuning.

To provide a high-quality multi-domain dataset for this study, we construct a training dataset from five commonly-used and high-quality human-human dialogue corpora leveraging data augmentation and pseudo labeling [22]. We compare our PoE metric with the state-of-the-art metrics and two strong baselines trained on our constructed multi-domain dataset.

In this paper, we make the following contributions: (1) We bridge the gap between existing model-based reference-free ADEMs and a strong multi-domain automatic dialogue evaluation metric, which can also effectively handle out-of-domain evaluation. More specifically, we realize this by constructing a high-quality multi-domain dataset for training the ADEMs and proposing PoE, a novel automatic dialogue evaluation metric based on transformer adapters. To our knowledge, PoE is the first multitask model that targets evaluation across a wide range of dialogue domains. (2) We empirically show that PoE outperforms existing state-of-the-art ADEMs as well as strong baselines on a large collection of evaluation datasets that covers different dialogue domains. In total, there are 11 in-domain and 5 out-of-domain evaluation datasets. (3) The implementation of PoE, datasets, and pretrained checkpoints will be released to the public, allowing researchers and practitioners to use or adapt them for their own evaluation tasks.

The remaining sections are organized as follows: Section II discusses the related work. In Section III, we formally define the dialogue evaluation task. In Section IV, we explain the proposed PoE metric. Section V describes the methods to construct the multi-domain dialogue dataset for training automatic dialogue evaluation metrics. The experiment preliminaries are outlined in Section VI. Section VII presents the experimental results and detailed analyses, which include in-domain, out-of-domain, and few-shot transfer analysis. The last section concludes the paper and outlines the future work.

II. Related Work

A. Automatic Dialogue Evaluation Metrics (ADEMs)

There are a number of commonly used evaluation metrics, that are simple, reference-based, and non-trainable. One category is the word-overlap metrics, such as BLEU [8], ROUGE [23], and METEOR [24]. This category of metrics assigns a score to the dialogue response based on its word or n-gram overlap with the corresponding human-written references. The other category is the embedding-based metrics, such as Greedy Matching [25] and Embedding Average [26]. By leveraging static word vectors, the embedding-based metrics move beyond surface-level matches and focus more on the semantic similarities between a dialogue response and the corresponding references. Despite

their simplicity, both categories are often criticized for their poor correlation with human evaluation [9]. The crux is that there are many possible responses to one given context in an open-ended dialogues [27].

Addressing the problem of multiple possible responses, the study of ADEMs shifts from reference-based approaches towards model-based reference-free ones [15]. Examples of recent model-based reference-free metrics include BERT-RUBER [12], PONE [13], USR-DR [7], GRADE [14], and MaUdE [10]. There are several common characteristics among these ADEMs: (1) To evaluate whether or not a response is appropriate with respect to a dialogue context as opposed to one or more references. (2) To employ a pre-trained language model [19], [20] to improve the ADEMs' classification capability (3) To avoid human annotations by using context-response data derived from a single human-human dialogue corpus. Examples are DailyDialog [1], PersonaChat [4] or TopicalChat [2]. Despite much success, the model-based reference-free approaches face several challenges. A major one is their inability to generalize to dialogue data beyond what they are trained on.

In this paper, we propose to tackle the problem from both the algorithm and data perspectives by developing a novel ADEM based on transformer adapters [21]. We also construct a high-quality dataset to facilitate the study. Zhang et al. [16] proposed MDD-Eval which also targets multi-domain dialogue evaluation. However, there are several key differences between both works. First, the underlying algorithms are completely different. The MDD-Eval metric is a single model trained on pseudo-labeled augmented data. It lacks the mechanism to handle domain-specific datasets in a specialized manner. On the other hand, our PoE metric adopts multi-task learning with the parameter-efficient adapter network. Each adapter module serves as a domain expert. Second, even though the training data of both MDD-Eval and PoE are constructed in a similar manner, PoE is more compatible with the data collection pipeline, because it handles increasing number of datasets more efficiently. MDD-Eval or single-model metrics in general require retraining or full-model finetuning when adapting to new domain-specific or task-specific data, whereas for PoE, we can just add new parameter-efficient adapter modules to handle the new datasets. Section VII-D presents the empirical evidence to support our claim. Third, in Section VII, we show that PoE is a much stronger multi-domain evaluation metric than MDD-Eval under the in-domain, out-of-domain and few-shot transfer settings.

It is noted that dialogue quality is multi-faceted in nature [7], model-based ADEMs are often designed for response appropriateness, as well as engagement [28], naturalness [7], adequacy [29], coherence [30], [31], consistency [32], [33], etc. There have been studies on combining different specialized models for multi-dimensional[2] dialogue evaluation, for example, D-score [11], HolisticEval [34], USR [7], and USL-H [35]. As the scope of this paper is on multi-domain dialogue evaluation rather than multi-dimensional dialogue evaluation, we focus on dialogue qualities that are more frequently studied in the literature and highly correlate with the pre-training objective of our proposed metric, the response appropriateness.

[2]Multi-dimension refers to multiple dialogue qualities.

B. Multitask Learning

In multitask learning [36], a model is trained simultaneously with multiple tasks and a shared representation is learned to capture the commonalities among the related tasks [36]. Multitask learning is an effective approach to reduce over-fitting to a particular task and thus, improve generalizability of the models [37], [38]. It has been successfully applied in a wide range of natural language processing tasks, such as semantic parsing [39], sequence labeling [40], language modeling [19], [41], machine translation [42], [43], and dialogue [44], [45].

In the context of open-domain dialogue evaluation, we recently applied multitask learning [11] for a holistic assessment of dialogues whereby the related tasks are designed to evaluate different dialogue qualities, including language fluency, coherence, semantic appropriateness, and logical consistency. Unlike [11], which addresses multi-dimensional evaluation, we study multi-domain evaluation in this paper by designing a hard-parameter sharing network, which consists of multiple transformer adapters sharing an underlying pretrained transformer encoder. Through multitask learning, the shared transformer encoder is expected to adapt and capture the general knowledge of dialogues while the domain-specific adapters capture specific properties with respect to the respective domains.

C. Adapters

Adapters are lightweight task-specific modules interleaved between layers of a pre-trained network [21], [46]. Adapter-based transfer learning performs similarly to full finetuning, but being more parameter efficient. Houlsby et al. [21] demonstrates that on the GLUE benchmark [47], finetuning only the task-specific adapters attains within 0.4% of the performance of full fine-tuning, which requires adapting all parameters of a pretrained model for each task. In a separate study, Stickland and Murray [48] propose a neural architecture that adds task-specific adapters to BERT [19] and train the entire network with a multi-task learning setup. This neural architecture performs similarly to those separately finetuned on the GLUE benchmark.

More recently, Friedman et al. [49] proposes the MADE model for extractive question and answering tasks. MADE consists of a shared transformer encoder, dataset-specific token classifiers and adapters. When training on a mixture of source datasets, all parameters within MADE are jointly optimized. The model has attained strong in-domain and out-of-domain performance. Following this line of thought, we apply the adapter-based multitask network in the multi-domain dialogue evaluation for the first time.

D. Ensemble

Ensemble is a common technique for boosting prediction accuracy. The conventional ensemble involves two steps: (1) making predictions with multiple independent models; (2) integrating the predictions into a final result [50]. In open-domain

dialogue evaluation, prior works, such as USR [7], USL-H [35], and D-score [11], ensemble multiple metrics to boost correlation with the overall human judgment. For PoE, we apply the unweighted average of predictions inferred by different domain-specific adapters to obtain the final prediction score and examine whether such an approach can yield good out-of-domain performance.

Besides the conventional ensemble of model predictions, Friedman et al. [49] propose a simple method to average the parameters of multiple adapters, which achieves good generalization on unseen question answering datasets. Matena & Raffel [51] propose to merge pre-trained language models that are fine-tuned on various text classification tasks via parameter averaging. In the same line of thought, we assess whether parameter averaging of the domain-specific adapters can achieve performance similar to that of PoE while incurring less inference and memory cost. Concurrent to our work, Wortsman et al. [52] explore averaging the weights of multiple models that are fine-tuned with different hyper-parameter configurations on the same task. They demonstrate that simple parameter averaging attains strong performance in image classification. Their work further validates the effectiveness of parameter averaging.

III. Problem Formulation

In this section, we formally define the multi-domain dialogue evaluation task. Assume we have a collection of J dialogue evaluation datasets, denoted as $\mathcal{D}$. An evaluation dataset within $\mathcal{D}$ is denoted as D^j, where $j \in \{1, \ldots, J\}$.

D^j contains I number of dialogue context-response pairs.[3] We denote the context and the corresponding response as c_i^j and r_i^j respectively, where $i \in \{1, \ldots, I\}$. In addition, each (c_i^j, r_i^j) is annotated by several human judges, and each human judge will provide a quality score based on the Likert scale to indicate his/her perception of the quality of (c_i^j, r_i^j). The mean human score w.r.t. (c_i^j, r_i^j) is denoted as q_i^j.

We aim to learn a metric, $M(c_i^j, r_i^j) \rightarrow s_i^j$ where s_i^j is the metric score that reflects the quality of (c_i^j, r_i^j) as perceived by M. To assess the performance of M on D^j, the correlation score between $S^j = \{s_1^j, \ldots, s_I^j\}$ and $Q^j = \{q_1^j, \ldots, q_I^j\}$ are computed. We use ρ^j to represent the correlation score on D^j. Higher ρ^j indicates better performance on D^j. To test the performance of M on the J evaluation datasets, we compute the average correlation $\tilde{\rho} = \frac{1}{J}\sum_{j=1}^{J} \rho^j$. The specific form of correlation we adopt is the Spearman's rank correlation [53], a common statistical measure used for assessing metrics' performance. Spearman's rank correlation determines the monotonic relationship between two variables. It is appropriate for continuous and discrete ordinal variables [54]. In our case, both S^j and Q^j are treated as two continuous variables and their Spearman's rank correlation can be computed as follows:

$$\rho^j = 1 - \frac{6\sum \mathrm{k}_i^{j\,2}}{I(I^2-1)}$$

where k_i^j is the difference between the ordinal rank of s_i^j within S^j and that of q_i^j within Q^j. The range of Spearman's rank correlation is between +1 and -1. +1 means a perfect monotonic association of the two variables. -1 means a perfect negative monotonic association of the two variables. 0 means that there is no association between the two variables.

Suppose that we train M with a collection of N training datasets, denoted as Ω. Each dataset in Ω belongs to a unique dialogue domain, n. Hence, Ω covers N different domains. We denote each training dataset as Ω^n where $n \in \{1, \ldots, N\}$.

We want to assess both the in-domain and out-of-domain performance of M. For in-domain assessment, $\tilde{\rho}$ is computed over the subset of $\mathcal{D}$, of which the data are in-distribution w.r.t. Ω^n. For out-of-domain assessment, $\tilde{\rho}$ is computed over the subset of $\mathcal{D}$, which contain out-of-distribution data w.r.t. Ω.

IV. A Panel of Experts

A. Motivation of a Panel of Experts

Research on domain generalization remains under-explored in the field of automatic dialogue evaluation. According to a recent survey on domain generalization [55], there are mainly three method categories to improve domain generalization. First, data manipulation, such as data augmentation and data generation. Second, representation learning, such as domain invariant representation learning and feature disentanglement. Third, learning strategies, such as ensemble learning and multi-task learning. To our knowledge, only the recent work, MDD-Eval [16] targets domain generalization in automatic dialogue evaluation. The key idea of MDD-Eval falls under the first method category. Yet, purely relying on multi-domain training data is not enough for domain generalization. In fact, the above-mentioned method categories are complementary to each other and can be combined towards better performance [55]. This motivates us to study better learning strategies and network architecture that can further enhance domain generalization of model-based metrics.

In terms of learning strategy, multi-task training is a natural choice to improve domain generalization. Through a shared representation, the learning on a particular domain-specific dataset can be improved by leveraging additional information from other related domains [37]. Furthermore, different from other tasks, such as text classification, open-domain dialogue generation is open-ended in nature. During evaluation, the dialogue data are often different from the training data of the model-based metrics. It is infeasible to re-train the model or conduct full-model fine-tuning whenever there are new dialogue data. it is more parameter-efficient to reuse existing knowledge encoded in the pretrained model and train lightweight adapters [21] for domain adaptation. An additional benefit of adapters than full-model fine-tuning is to reduce catastrophic forgetting [56].

Although multi-task learning and adapters have been proven effective in other tasks, such as text classification [48], image classification [46], and cross-lingual transfer [57], to our knowledge, its efficacy to dialogue evaluation has not been studied in prior works. We are the first to apply the idea to automatic

[3]The context is one or a few consecutive utterances drawn from a human-human dialogue, and the response is generated by a chatbot conditioning on the context.

dialogue evaluation and comprehensively analyze its effectiveness in terms of in-domain, out-of-domain, and few-shot transfer evaluation. In the experiments (Section VII), we demonstrate that our PoE metric attains state-of-the-art capability in evaluating both in-domain and out-of-domain dialogue data on on a diverse set of dialogue evaluation datasets compared to exsiting model-based metrics.

B. System Architecture

As shown in Fig. 1, we formulate a Panel of Experts (PoE) as an automatic dialgoue evaluation metric, which consists of a pretrained transformer encoder, N adapters, and N classifiers. The transformer encoder, denoted as T, contains L number of layers, $\{T_1, \ldots, T_L\}$. Each adapter, denoted as π^n, consists of a series of adapter layers, $\{\pi_1^n, \pi_2^n, \ldots, \pi_{L-1}^n\}$. An adapter layer is interleaved between two consecutive transformer layers. For example, π_{L-1}^n is inserted in between T_{L-1} and T_L. Each classifier, denoted as ϕ^n, is a single-layer feed-forward network followed by a sigmoid activation function. As defined in Section III, we have N training datasets with each cover a unique dialogue domain. $\{\pi^n, \phi^n\}$ learn to classify responses from their respective domain-specific dataset Ω^n. During training, all $\{\pi^n, \phi^n\}_{n=1}^{|N|}$ and T are jointly optimized in a multitask learning manner.

The input data of PoE, x_i^n is a context-response pair from Ω^n. The pair is denoted as (c_i^n, r_i^n). The associated label of x_i^n is denoted as y_i^n, which indicates whether r_i^n is appropriate w.r.t. c_i^n. It is either 1 (appropriate) or 0 (inappropriate). Note that y_i^n is a pseudo label (discussed in Section V-C) instead of the ground-truth human label. Hence, PoE can be seen as a semi-supervised approach, which differs from all existing model-based dialogue evaluation metrics, except our MDD-Eval metric [16]. However, the architecture of MDD-Eval is simple (a pretrained RoBERTa [20] plus a feed-forward classification network) while PoE benefits from the transformer adapter and multitask learning for improved prediction ability and generalizability.

(c_i^n, r_i^n) is concatenated into a single sequence of tokens including special start, end, and separation tokens when fed into the network: "<s> c_i^n </s> r_i^n </s>". The input sequence length is constrained to 512. The network output $\tilde{y}_i^n$ of PoE represents the model's confidence about the appropriateness of r_i^n conditioning on c_i^n:

$$\tilde{y}_i^n = p_{\theta_M}(y_i^n = 1 \mid c_i^n, r_i^n)$$

where M refers to the PoE model parameterized by θ.

C. Training Objective

The training objective of PoE is defined as follows,

$$\underset{\theta_M}{\operatorname{argmin}} \mathbb{E}_{\Omega^n \sim \Omega}[\mathbb{E}_{(x_i^n, y_i^n) \in \Omega^n}[-(y_i^n \log \tilde{y}_i^n + (1-y_i^n)\log(1-\tilde{y}_i^n))]]$$

We minimize the binary cross-entropy loss between $\tilde{y}_i^n$ and y_i^n. The whole network is trained in a multitask manner whereby during training, T and $\{\pi^n, \phi^n\}_{n=1}^{|N|}$ are jointly optimized. $\{\pi^n, \phi^n\}$ are trained to perform their respective domain-specific classification task. More specifically, given a training mini-batch that consists of samples uniformly drawn from any training dataset in Ω, the parameter update of $\{\pi^n, \phi^n\}$ only depends on $(x_i^n, y_i^n) \in \Omega^n$ in the mini-batch. On the contrary, T is optimized with all training instances in the mini-batch. Hence, $\{\pi^n, \phi^n\}$ are domain-specific while T is domain-independent. In this way, T learns to adapt to the multi-domain dialogue dataset and captures a general representation that encodes regularities w.r.t. dialogue data of different domains, while the adapter modules learn to capture the unique characteristics of various dialogue domains, thus serve as the domain experts.

D. Inference Process

We evaluate a trained PoE on the dialogue evaluation task defined in Section III. Given a context-response input pair (c_i^j, r_i^j) from evaluation dataset D^j, PoE will run the forward pass N times in parallel and output N confidence scores denoted as $\{\tilde{y}_i^1, \ldots, \tilde{y}_i^N\}$. The final metric score, s_i^j, is computed in the following manner:

$$s_i^j = \begin{cases} \tilde{y}_i^n & \text{if } D^j \text{ and } \Omega^n \text{ share the same domain} \\ \frac{1}{N}\sum_{t=1}^{N} \tilde{y}_i^n & \text{otherwise} \end{cases}$$

If an evaluation dataset is in-distribution w.r.t. the training data of PoE, we directly apply the confidence score of the corresponding classifier. Here, a single expert makes the decision. For out-of-distribution evaluation datasets, we perform unweighted averaging on all the confidence scores. In this case, all experts jointly make the decision. Hence, our proposed metric is dubbed as PoE: a **P**anel **o**f **E**xperts.

In out-of-distribution evaluation, the inference involves running the model N times. A simplified strategy is to derive a single adapter and a classifier by averaging the parameters of all N adapters and N classifiers. In practice, we take the arithmetic mean, denoted as ϕ', of the parameters of the individual classifiers $\{\phi^1, \phi^2, \ldots, \phi^N\}$. The parameters of the single adapter module, denoted as π_l' for layer l, are the arithmetic mean of the parameters of $\{\pi_l^1, \pi_l^2, \ldots, \pi_l^N\}$. s_i^j is obtained with the single module, $(\{\pi_1', \pi_2', \ldots, \pi_{L-1}'\}, \phi')$. We justify the use of the arithmetic mean of the parameters of the adapters and classifiers to form a single model by considering the fact that all the adapters share the same configuration as in [21]. Furthermore, the weights of the adapters and classifiers are all initialized with the same uniform distribution.

We denote PoE after parameter averaging as PoE-avg. PoE-avg combines the domain-specific knowledge of multiple experts (adapters) into a single expert (the adapter after parameter averaging) in an early-fusion manner. It serves as a light-weight variant of PoE and is expected to perform on par with PoE in both in-domain and out-of-domain evaluation.

E. Few-Shot Transfer Learning

We also consider a transfer learning setup whereby PoE-avg is finetuned with a small number of human-annotated instances from the target dialogue evaluation dataset. The reason for conducting few-shot transfer learning on PoE-avg instead of PoE

is that we often do not have prior knowledge on new dialogue evaluation datasets. Hence, we may need to finetune all adapters in PoE instead of a domain-relevant one. With PoE-avg, we just need to conduct the transfer learning with a single-adapter setup.

Since the data instances are annotated with continuous human ratings rather than discrete labels, we adopt mean squared error as the optimization objective. Details on the setup of few-shot transfer learning experiments are outlined in Section VI-C.

V. Multi-Domain Dialogue Training Dataset

The success of the multitask training of PoE relies on a large-scale, high-quality, and multi-domain dialogue dataset. However, it is not trivial to construct such a dataset. In the study of model-based automatic dialogue evaluation, many efforts are devoted to neural architecture design, however, the development of high-quality training data are not given sufficient attention. There are two major challenges constructing an adequate training dataset.

First, existing model-based metrics heavily rely on response sampling strategies [10], [11]. Yet, the sampling strategies do not always produce data of good quality. For instance, the commonly-used random utterance selection strategy[4] tends to introduce over-simplistic and false-negative responses. The over-simplistic responses refer to responses that neither semantically nor syntactically overlap with the dialogue context. As pointed out in [16], a large number of such over-simplistic responses mislead the metrics by associating response appropriateness with only content similarity, thus, introducing unwanted bias. Furthermore, the false negatives can lead the model to misclassify appropriate responses.

Second, plausible ways to improve the data quality include the implementation human-in-the-loop quality control or creating data with crowd-sourcing. However, these approaches are costly and time-consuming considering that we need a large-scale multi-domain dataset.

Overcoming the challenges, by extending [16], we apply semi-supervised learning techniques to automatically construct a multi-domain context-response dataset through a 3-step workflow, (1) multiple human-human dialogue corpora (Section V-A), (2) a set of dialogue response augmentation techniques (Section V-B), and (3) a pretrained evaluation model for data pseudo labeling and quality control (Section V-C).

A. Human-Human Dialogue Corpora

Five human-human dialogue corpora are selected for collecting the context-response pairs, as summarized in Table I, which are DailyDialog [1], ConvAI2 [58], TopicalChat [2], EmpatheticDialogue [3] and REDDIT [59]. We selected the five corpora for the following reasons, (1) They have been used in the studies of open-domain dialogue and are of good quality. (2) Each dialogue corpus is collected with a specific goal, hence in one unique domain. For instance, ConvAI2 is about persona-guided conversations while DailyDialog focuses on typical topics discussed in our daily life. (3) They are of a suitable data size for our data collection process.

TABLE I
Human-Human Dialogue Corpora Statistics

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
#context-response pairs	76,052	7,069
EmpatheticDialog	**training**	**validation**
#dialogues	19,529	2,768
#utterances	84,158	12,075
#words	1,127,355	174,786
#avg utterances per dialogue	4.31	4.36
#avg words per dialogue	57.73	63.15
#context-response pairs	64,629	9,307
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
#context-response pairs	235,820	14,566
TopicalChat	**training**	**validation**
#dialogues	8,627	538
#utterances	188,357	11,660
#words	4,374,304	273,331
#avg utterances per dialogue	21.83	21.67
#avg words per dialogue	507.05	508.05
#context-response pairs	179,730	11,122
REDDIT	**training**	**validation**
#dialogues	91,919	12,023
#utterances	644,429	82,927
#words	8,104,273	1,044,756
#avg utterances per dialogue	7.01	6.90
#avg words per dialogue	88.17	86.90
#context-response pairs	523,044	65,192

We only use the training and validation split of the dialogue corpora, because a part of the test split of the dialogue corpora is used for evaluation in our experiment.

1) DailyDialog [1]: The DailyDialog dataset contains high-quality and human-written conversations that cover a wide range of generic topics, such as relationships, ordinary life, and work. The conversations in DailyDialog are mainly for information exchange and social bond enhancement.

2) EmpatheticDialogue [3]: The EmpatheticDialogue dataset is created for developing dialogue agents that can recognize feelings in the conversation partner and reply accordingly. The conversations are grounded in emotion situations whereby a person describes his or her personal experiences and feelings. The conversation partner acknowledges his or her feelings and then provides appropriate responses.

[4]This strategy refers to the random selection of an utterance from a different dialogue as the inappropriate response w.r.t. the current dialogue context.

3) ConvAI2 [58]: ConvAI2 is an extended dataset of the Persona-Chat [4] corpus, which is about exchanging persona information, i.e., the conversations in ConvAI2 are grounded by the personas of the interlocutors. The conversations are about two interlocutors trying to be engaging, to learn about the other's interests, discuss their own interests, and find common ground information [58]. Each persona contains at least 5 sentences describing the corresponding role. In total, there are 1155 possible personas for training. Topic shifts are common within the conversations in ConvAI2 as the interlocutors are continually introducing new information about themselves along the conversation.

4) TopicalChat [2]: TopicalChat is a knowledge-grounded human-human conversation dataset, which contains conversations between two interlocutors exchanging knowledge information. The underlying knowledge spans across 8 broad topics, including fashion, politics, books, sports, general entertainment, music, science & technology, and movies. Each conversation is associated with a Washington Post article and the top three entities by frequency of occurrence in the article. Depending on the configuration, the two interlocutors have access to different knowledge snippets w.r.t. the three entities. The knowledge sources include Wikipedia, Reddit and Washington Post articles.

5) REDDIT [59]: The REDDIT dataset is built on discussions on the Reddit social media platform. There are many different subreddits available, with conversations largely different in topics, language styles, and participation patterns. In total, 109 conversations of at least 3 turns are collected with the median conversation containing 7 utterances. The conversations are extracted from the 2,018 conversational exchanges in the Casual Conversations forum (r/CasualConversations), a community of 607 K conversationalists discussing a variety of topics. Unlike other dialogue corpora, the conversational style in REDDIT is more causal and less organized. In addition, the topics in REDDIT dialogues are more diverse and time-dependent, i.e., the topics may differ a lot when the conversations are carried out at different times.

B. Response Augmentation Techniques

Due to the one-to-many nature of open-ended conversations, the evaluation metrics will benefit from training on multiple appropriate and inappropriate responses per dialogue context. From a human-human dialogue in the five dialogue corpora, we extract 1 to 4 consecutive utterances as the dialogue context, the follow-up utterance serves as the corresponding appropriate response. To generate multiple appropriate and inappropriate responses per context, we need to rely on various response augmentation strategies:

1) Syntactic & Semantic Negative Sampling: Inspired by [10], the following perturbation techniques are applied on the original appropriate responses to generate syntactic negative responses: (1) word-drop (a random portion of tokens in the original response, up to 50%, is dropped). (2) word-shuffle (the order of tokens in the original response is shuffled). (3) word-repeat (randomly repeat words in the original response). For collecting semantic negative responses, we follow the common strategy of randomly sampling an utterance from a different dialogue (within the same corpus) to replace the original appropriate response w.r.t. a dialogue context. To tackle the aforementioned limitations of such random sampling strategy, we leverage additional augmentation techniques as outlined in the subsequent sections and model-in-the-loop quality control measure (Section V-C).

2) Back-Translation: Given a dialogue context-response pair extracted from a human-human conversation, Back-Translation [60] is applied to generate paraphrases of the original response. In the actual implementation, we adopt the pretrained WMT'19 English-German and German-English ensemble model to perform back-translation.

3) Generation From State-of-the-art Dialogue Systems: We rely on state-of-the-art dialogue systems including DialoGPT [61] and BlenderBot [6] to generate a set of appropriate responses with different semantic meanings conditioned on a dialogue context. These systems have been pretrained on a large amount of conversation data and they demonstrate strong ability in generating fluent and on-topic responses.

4) Automatic Generation of Adversarial Responses: Motivated by [62], the mask-and-fill strategy is adopted. There are two steps: (1) masking, where one or a few tokens of a response (up to 15% of the tokens) are replaced with mask tokens. (2) infilling, a pretrained infilling language model [63] is adopted to replace the mask tokens in the response with new tokens conditioned on a random dialogue context instead of the original context. For example, if named entities in a response are masked out, the infilling process conditioned on a random context may introduce different named entities that are not consistent with the original context. Such a response may seem to be appropriate in terms of surface lexical features, but in fact, it is semantically inappropriate.

Another strategy is to randomly sample an utterance from the dialogue context and then perform syntactic perturbations on the sampled response. This strategy intends to generate adversarial inappropriate responses that share a certain degree of content similarity with the corresponding contexts.

Both strategies are intended to automatically construct adversarial negative responses and reduce the reliance on random sampling strategy for introducing semantically negative responses.

C. Pseudo Labeling & Quality Control

To avoid excessive false-negative or false-positive data instances in automatically-constructed dataset, we need to put in place a mechanism to filter out low-quality samples. Note that human validation for quality control on a large-scale dataset is costly. We adopt a strong model to provide pseudo labels [22] to all the context-response pairs candidates during data augmentation in Section V-B.

Instead of training a model on a human-generated dataset from scratch as what we did in [16], we leverage the "Dialogue Evaluator with BERT (DEB)" model released by Sai et al. [64]. The rationale is that DEB is first pretrained on a large-scale Reddit dataset (767 M Reddit dialogue) and then, finetuned with the human-generated DailyDialog++ dataset. DEB can

TABLE II
SUMMARY OF THE POE EVALUATION DATASETS, THAT ARE NEITHER INVOLVED IN TRAINING NOR TUNING

Name	#Instances	Avg.#Utts.	Avg.#Ctx/Hyp Words	Domain	#Annotations	Dimension	Neural architecture of the dialogue systems
PUR [7]	300	9.3	98.4 / 12.0	PersonaChat	5.4K	Maintains Context	Transformer/LSTM Seq2Seq, Memory Network
CGR [14]	600	3.0	24.4 / 11.3	PersonaChat	3K	Relevance	Transformer Seq2Seq, DialoGPT, BERT/Transformer Ranker
PZH [66]	900	5.1	48.8 / 11.5	PersonaChat	3.6K	Appropriateness	Random Sampling, LSTM Seq2SeqAttn, and GPT-2
PDS [65]	4,829	4.0	36.00 / 11.6	PersonaChat	77K	Appropriateness	LSTM Seq2SeqAttn, BlenderBot, DialoGPT and GPT-3
DGR [14]	300	3.0	26.0 / 10.8	DailyDialog	3K	Relevance	Transformer Seq2Seq/Ranker
DZH [66]	900	4.7	47.5 / 11.0	DailyDialog	14.4K	Appropriateness	Random Sampling, LSTM Seq2SeqAttn, and GPT-2
DGP [67]	500	4.9	49.9 / 10.9	DailyDialog	2.5K	Overall	LSTM Seq2Seq, Conditional VAE
TUR [7]	360	11.2	236.3 / 22.4	TopicalChat	6.5K	Maintains Context	Transformers Seq2Seq
TDS [65]	4,500	4.0	50.6 / 15.9	TopicalChat	72K	Appropriateness	LSTM Seq2SeqAttn, BlenderBot, DialoGPT and GPT-3
EGR [14]	300	3.0	29.0 / 15.6	Empathetic	3K	Relevance	Transformer Seq2Seq/Ranker
RDS [68]	9,990	3.5	35.3 / 11.2	Reddit	30K	Relevance	RNN/LSTM Seq2Seq, Memory Network, Pointer-generator
FDT [69]*	375	10.4	87.3 / 13.3	Other	17K	Relevance	Meena, Mitsuku
HUM [70]*	9,500	3.9	17.0 / 6.1	Other	57K	Relevance	Random Sampling
ESL [71]*	1242	2.0	7.05 / 11.81	Other	13K	Overall	BlenderBot, DialoGPT, HRED, Transformer/LSTM Seq2Seq
NCME [71]*	2461	2.0	7.34 / 8.57	Other	33K	Overall	BlenderBot, DialoGPT, HRED, Transformer/LSTM Seq2Seq
ConTurE [72], [73]*	1066	3.8	21.67 / 10.99	Other	3.2K	Overall	State-of-the-art systems including Plato and DialoGPT

Part of the information are obtained from [15] and [65]. * denotes the out-of-domain evaluation datasets W.R.T. The training data. The computation of "Avg.#Utts" includes both the context and the response. PUR, CGR, PZH, PDS, DGR, DZH, DGP, TUR, TDS, EGR, RDS, FDT, and HUM refer to Persona-USR, CONVAI2-Grade, Persona-Zhao, Persona-DSTC10, DailyDialog-Grade, DailyDialog-Zhao, DailyDialog-Gupta, Topical-USR, Topical-DSTC10, Empathetic-Grade, Reddit-DSTC7, FED-Turn, and humod respectively.

generalize across domains due to large-scale pretraining while being capable of accurate estimation of response appropriateness due to learning from manually-crafted data. Yeh et al. [15] also proves it to be a strong automatic dialogue evaluator on a large number of turn-level evaluation datasets.

Concretely, for a context-response pair, DEB provides a soft pseudo label that indicates its confidence on the appropriateness of the response w.r.t. the context. The soft pseudo label is a probability distribution over two classes (appropriate and inappropriate). A confidence threshold of 90% is adopted to exclude pairs classified by DEB with low confidence.

In the end, for each dialogue corpus, we collected 400 K context-response pairs for training and 40 K pairs for validation. Both training and validation split are class-balanced, i.e., they contain equal number of appropriate and inappropriate context-response pairs. In total, the multi-domain dataset contains 2 M context-response pairs in the training split and 20 K pairs in the validation split.

VI. EXPERIMENT PRELIMINARIES

A. Evaluation Datasets

We assess PoE on 16 dialogue evaluation datasets. The selection of evaluation datasets is guided by the recent comprehensive survey on automatic dialogue evaluation metrics [15] as well as the "Automatic Evaluation" shared task of DSTC10[5] [65]. Table II summarizes the essential characteristics of all evaluation datasets. Some evaluation datasets contain annotations along multiple evaluation criteria. For example, the FED-Turn [69] dataset contains annotations along 9 different fine-grained criteria, such as relevance, interestingness, fluency, etc. Since multi-dimensional evaluation is beyond the scope of this work, we only consider response appropriateness in our analysis. For evaluation datasets without annotations along response appropriateness, the criterion that is closest to response appropriateness is considered, such as context relevance or overall quality. As shown in Table II, the "Dimension" column contains the criteria we consider for our correlation analysis.

[5]The Tenth Dialog System Technology Challenge (DSTC10).

Moreover, the evaluation datasets can be categorized with their respective dialogue domains (as indicated in the "Domain" column). Those belonging to the "Other" domain are considered out-of-domain evaluation datasets while the rest are the in-domain evaluation datasets.

B. Baselines

We compare PoE with three types of systems: (1) published state-of-the-art automatic dialogue evaluation metrics. Specifically, we pick the top-ranked ones that are presented in the comprehensive survey [15]. We include DEB [64], USL-H [35], GRADE [14] and USR [7]. USL-H and USR target multi-dimensional evaluation. Hence, both of them contain multiple models with each focus on a specific dialogue quality. Since we only target the response appropriateness, the USR-DR component of USR and the BERT-NUP component of USL-H are adopted respectively. Additionally, we also report the results of the best team in the Track 5.1 of the DSTC10 shared task [65]. The rationale is that the shared task proposes a meta-evaluation benchmark that covers all evaluation datasets used in this paper except ConTurE, which is more recent than the rest. We want to benchmark PoE against the most recent state-of-the-art automatic dialogue evaluation metric.

(2) To showcase the advantages brought by the PoE metric alone, instead of the multi-domain training dataset we have collected, we compare PoE against a strong single-model baseline that is trained on the same multi-domain dataset as PoE. The model consists of a pretrained transformer encoder (same as PoE) and a single feed-forward classification network. We denote the baseline as Single-T. In fact, Single-T is similar to the recently proposed MDD-Eval metric [16] since both have the same architecture. Their differences are: (a) Single-T is trained on more data compared with MDD-Eval (2 M vs 600 K). (b) MDD-Eval is optimized with three different losses while Single-T is optimized with only the cross-entropy loss. We empirically find that Single-T performs on par with MDD-Eval. On some evaluation datasets, it even outperforms MDD-Eval. We include both Single-T and MDD-Eval as a baseline in our experiment.

(3) To show the advantage of a Panel of Experts (PoE), we compare PoE with a collection of individual domain-specific models, that is referred to as a Collection of Experts (CoE). PoE is optimized with multitask learning with a shared architecture across domains, while CoE is not. An individual model in CoE has the same architecture as Single-T, but trained on one domain-specific subset of the multi-domain dataset.

C. Experiment Setup

Following [15], the results w.r.t. existing state-of-the-art metrics are computed with the best checkpoints released by the authors. For Single-T, CoE, and PoE, we repeat the training 10 times with different random seeds to reduce the effect of randomness on model performance. The mean Spearman correlations over the 10 runs are reported for each evaluation dataset. In addition, we perform William's T test [74] for pairwise significance tests. In all the tables, we use † on PoE variants if they significantly outperform Single-T, CoE, and MDD-Eval ($p < 0.05$).

We adopt the RoBERTa-base model [20] as the pretrained transformer encoders in PoE and baselines.This is because RoBERTa has been proven as a powerful text encoder that are beneficial for the automatic dialogue evaluation task in prior works [7], [11], [66], [75]. In addition, we want to have a fair comparison with the existing state-of-the-art metrics, which use either BERT-base or RoBERTa-base except DEB, which is based on BERT-large [19].

Since the training task for PoE is a binary classification task, we adopt accuracy to determine the model performance. The checkpoint with the best accuracy on average over the five validation datasets is picked to perform the dialogue evaluation task. For the dialogue evaluation task, we adopt Spearman correlations to assess the performance of the ADEs. All experiments are conducted on a single Tesla V100 GPU of 16 GB memory.

Following [49], all our experiments are implemented with PyTorch [76], HuggingFace Transformers [77] and the adapter-transformers library [78]. For training PoE, we adopt AdamW optimizer [79] with a constant learning rate of 5e-6. We set the training batch size to 32. The number of training epochs is set to 3. The model is evaluated every 2000 steps. If the average validation accuracy does not improve for ten consecutive checkpoints, we stop the training process. In addition, each mini-batch consists of training instances uniformly drawn from each of the training datasets at run time. After the model is fully optimized with multitask training, we freeze the transformer encoder and continue finetuning each adapter separately on their respective training/validation dataset for 10 more epochs. During finetuning, a constant learning rate of 1e-5 is adopted. The model is evaluated every 1024 steps and if the corresponding validation accuracy does not improve for three consecutive checkpoints, we stop the process.

The training hyperparameters of Single-T is the same as those of PoE, except that Single-T doesn't have the adapter finetuning process. For CoE, each domain-specific model is trained exactly in the same manner as Single-T. The only difference is that Single-T is trained on the multi-domain dataset (same as PoE) while the domain-specific models of CoE are trained on their respective in-domain datasets.

We conduct the few-shot transfer learning experiments on both PoE-avg and Single-T. For each evaluation dataset, we randomly sample K% of the data and K is set to 10%, 20%, 30%, and 40% respectively. Then, we split the K% sample set into half with one half for model finetuning and the other half for validation. The target label of each data instance is the average human annotation score. We adopt a training batch size of 2 and set the learning rate to 1e-5. The model is evaluated with Spearman correlation on the validation set. If the correlation doesn't improve for 10 consecutive epochs, we stop the process. After the finetuning process, we evaluate the model on the full evaluation dataset. All the few-shot experiments are repeated 10 times with different random seeds. The analysis in Section VII-C is based on the mean Spearman correlations over all the 10 trials.

VII. Results & Analysis

We would like to answer the following questions: (1) How does PoE perform for in-domain data (Section VII-A)? (2) How does PoE generalize to out-of-domain evaluation (Section VII-B)? (3) How does PoE adapt to unseen domains with few-shot transfer learning (Section VII-C)? (4) Can PoE be applied to downstream dialogue tasks, such as response selection (Section VII-D)? (5) How efficient is PoE compared to Single-T and CoE (Section VII-E)?

A. In-Domain Performance of PoE

First, USL-H, GRADE, and USR are domain-specific metrics as they are trained on specific dialogue corpora. USL-H and GRADE are trained on context-response pairs that are based on DailyDialog [1] while USR is trained on context-response pairs that are derived from TopicalChat [2]. In Table III, it can be observed that USR performs significantly better on average across all the TopicalChat-related evaluation datasets (row 14) than USL-H and GRADE. On the other hand, USL-H and GRADE perform significantly better than USR on average across all the DailyDialog-related evaluation datasets (row 13). PoE can outperform USL-H and GRADE on the DailyDialog domain as well as USR on the TopicalChat domain. In addition, PoE also outperforms CoE, which are trained on domain-specific datasets, across all five domains (row 12–16). These observations confirm that PoE is effective for the multi-domain evaluation task.

Second, to provide a direct and fair comparison between PoE and the domain-specific baselines (which are trained on single-domain data, while PoE is trained on multi-domain data), we re-train USL-H, GRADE, and USR on the same multi-domain data as PoE and present their in-domain performance in columns USL-H+, GRADE+, and USR+ of Table III respectively. It can be observed that the correlation scores of all three baselines generally improve over their respective domain-specific variants. This shows that better multi-domain evaluation performance can be partially attributed to better training data. However, PoE still outperforms USL-H+, GRADE+, and USR+ by a large margin. The large performance gap is due to (1) the

TABLE III
SPEARMAN CORRELATIONS (%) ON 11 IN-DOMAIN EVALUATION DATASETS

Row	Datasets	USL-H*	USL-H+	GRADE*	GRADE+	USR*	USR+	DEB	Team 5	MDD+	CoE+	Single-T+	PoE-avg	PoE
1	PUR	47.57	54.02	38.31	42.98	56.25	46.06	43.52	49.63	55.03	56.11	60.72	61.85†	**63.23**†
2	PZH	61.83	61.56	57.45	63.04	51.74	59.63	56.99	61.32	59.74	66.74	66.97	**67.36**	67.30
3	PDS	39.75	45.30	41.48	44.40	38.65	41.07	38.67	43.92	37.74	45.26	44.90	44.84	**45.53**
4	CGR	56.30	57.79	57.05	58.28	48.51	51.18	50.43	**58.43**	44.30	57.10	56.06	57.04	57.48
5	DGR	9.13	19.00	25.40	20.76	7.14	16.38	36.29	33.42	28.08	23.04	26.45	**36.64**†	34.16†
6	DGP	61.42	56.43	59.62	52.00	35.76	53.44	57.86	**63.25**	57.13	55.80	61.30	61.35	61.77
7	DZH	53.11	45.15	53.73	48.93	35.41	47.83	52.23	57.54	56.42	51.89	**58.26**	58.24	56.13
8	RDS	28.34	38.31	34.12	38.13	37.03	40.61	37.85	34.34	40.12	39.89	42.19	**44.41**†	43.20†
9	TUR	11.88	35.13	13.84	34.06	36.50	38.24	15.14	37.61	**55.82**	41.45	41.26	41.42	43.72
10	TDS	20.77	27.30	24.59	28.71	27.41	28.92	29.85	27.93	30.52	33.05	31.70	**33.30**	33.20
11	EGR	33.40	42.72	34.28	40.19	34.08	44.27	39.56	30.57	37.58	43.87	44.05	46.00	**46.36**†
12	AVG (PC)	51.36	54.67	48.57	52.18	48.79	49.49	47.40	53.33	49.20	56.30	57.16	57.77	**58.39**
13	AVG (DD)	41.22	40.19	46.25	40.56	26.10	39.22	48.79	51.40	47.21	43.58	48.67	**52.08**†	50.69
14	AVG (TC)	16.32	31.21	19.22	31.39	31.95	33.58	22.49	32.77	43.17	37.25	36.48	37.36	38.46
15	AVG (EM)	33.40	42.72	34.28	40.19	34.08	44.27	39.56	30.57	37.58	43.87	44.05	46.00	**46.36**†
16	AVG (RE)	28.34	38.31	34.12	38.13	37.03	40.61	37.85	34.34	40.12	39.89	42.19	**44.41**†	43.20†

Scores with P-values > 0.05 are underlined (indicating statistical insignificance). TCU 12-16 correspond to the domain-specific average spearman correlations. The best score for each row is highlighted in bold. Team 5 achieves the first place in the DSTC10 "automatic dialogue evaluation" shared task. POE-AVG denotes poe using the parameter averaging inference method. Metrics that are accompanied by an asterisk are trained on domain-specific datasets. Metrics that are accompanied by "+" are trained on the same data as poe. The rest are domain-independent metrics. † denotes that poe variants significantly outperforms Single-T, COE-T, and MDD-EVAL ($p < 0.05$). PC, DD, TC, EM, and RE refer to personachat, dailydialog, topicalchat, empathetic, and Reddit respectively.

specialized adapter modules of PoE that captures the unique characteristics of different dialogue domains. (2) PoE supports multi-task training, which better exploits the domain-specific features than single-model training. The large performance gap also empirically proves that the dedicated architecture of PoE is essential to the multi-domain dialogue evaluation task.

Third, we compare PoE with DEB. DEB is a classification model with BERT-large as the backbone. It is first pre-trained on roughly 767 M Reddit conversations, then fine-tuned on the DailyDialog++ dataset. Hence, DEB has better generalizability than USL-H, GRADE, and USR. PoE outperforms DEB across all five domains. Especially for the Reddit domain (row 16), PoE attains a remarkable improvement of 5.35% over DEB even though DEB has been pretrained on a large amount of Reddit data. The significant improvement may be due to that (1) DEB has been applied in our dataset construction process (Section V-C). PoE acquires the knowledge of DEB on dialogue evaluation by training on its pseudo-labeled data; (2) Though DEB has more trainable parameters than PoE (340 M vs 132 M) and is pretrained on a much larger dataset (767 M vs 2 M), PoE benefits from its network architecture that captures both general knowledge across domains, and domain-specific knowledge.

Additionally, when compared to CoE, PoE attains superior performance across all domains. The superior performance can be attributed to multitask learning, which serves as an effective tool for boosting model performance through implicit data augmentation and regularization. During the multitask training, the RoBERTa encoder shared by all the adapters in PoE provides general and useful representations for the dialogue contexts and their corresponding responses. Besides the superior performance, PoE is also much more lightweight than CoE in terms of trainable parameters (132 M vs 623 M).

Furthermore, PoE also outperforms Single-T and MDD-Eval across all the dialogue domains. Since both Single-T and PoE are trained on the same data and both share the same type of pre-trained transformer encoder (RoBERTa-base), the performance improvement is attributed to PoE's incorporation of the adapters and multitask learning. Even though both Single-T and PoE possess general knowledge of the multi-domain dialogue data, the different adapter modules in PoE help capture the additional domain-specific knowledge. Moreover, if we compare PoE-avg with Single-T, it can be observed that even though both have approximately the same amount of trainable parameters, PoE-avg achieves better performance than Single-T. The observations confirm that PoE is a superior multi-domain automatic dialogue evaluation model to Single-T. MDD-Eval performs generally well across the five different domains. Yet, it faces the same limitation as Single-T: lack of dedicated network modules to capture additional domain-specific knowledge. Hence, it performs worse than PoE in the in-domain setting.

Additionally, PoE variants can outperform the best team (team 5) in the DSTC10 "Automatic Dialogue Evaluation" shared task [65] on 9 out of 11 evaluation datasets. Remarkably, for the Empathetic (row 15) and the Reddit (row 16) domains, PoE achieves performance gain of approximately 16% and 9% respectively in comparison to team 5. Team 5 employs a metric ensemble approach whereby the prediction scores of five different metrics are combined with weighted averaging. The five metrics target relevance, fluency, engagement, specificity, and topic coherence respectively. When performing evaluation on a particular dataset, the weight of each metric is dynamically determined by the Spearman correlation of the metric scores and the corresponding human annotation scores over a subset of that dataset. Different from Team 5's method, PoE doesn't rely on human annotation scores. In addition, it is a single-model metric instead of an ensemble of multiple metrics.

Moreover, the performance of PoE and PoE-avg is almost identical. This finding confirms our expectation about PoE-avg in Section IV-D. It can also be observed that PoE slightly outperforms PoE-avg on the PersonaChat, TopicalChat, and Empathetic domains (rows 12, 14, and 15) while PoE-avg slightly performs better than PoE on the DailyDialog and Reddit

TABLE IV
SPEARMAN CORRELATIONS (%) ON 5 OUT-OF-DOMAIN EVALUATION DATASETS

Row	Datasets	USL-H*	USL-H+	GRADE*	GRADE+	USR*	USR+	DEB	Team 5	MDD+	CoE+	Single-T+	PoE-avg	PoE
1	FDT	19.42	19.11	14.79	17.97	18.32	31.33	18.09	29.85	26.60	32.55	36.21	36.26	**36.74**
2	HUM	64.39	63.80	56.47	64.11	46.08	59.86	64.88	64.84	53.12	66.02	65.49	67.34	**67.35**†
3	ESL	34.53	36.94	30.01	33.61	11.25	38.59	42.24	40.01	**46.76**	45.95	40.24	42.94	46.37
4	NCME	28.02	28.45	21.86	26.15	5.23	22.24	28.57	29.60	21.81	19.90	26.21	**30.55**†	28.46
5	ConTurE	20.80	25.14	24.85	22.31	23.57	30.82	30.12	-	30.39	31.16	33.16	34.74	**35.36**†
6	AVG	33.43	34.69	29.60	32.38	20.89	36.57	36.78	41.08	35.74	39.12	40.26	42.37	**42.86**†

All the scores are statistically significant. The best score for each row is highlighted in bold. Row 6 corresponds to the average spearman correlations across all datasets except team 5 of which the average spearman correlation is computed with the first four datasets. † denotes that poe variants significantly outperform Single-T, COE-T, and MDD-EVAL ($p < 0.05$).

domains (rows 13, 16). A possible reason is that the dialogues in the DailyDialog and Reddit domains are more informal and their language usage is more common. Hence, the knowledge about other dialogue domains can better transfer to the evaluation of such dialogues, but the reverse may not be the same. As a result, PoE-avg, which carries knowledge about different domains, is more capable of evaluating informal dialogues than PoE, which provides domain-specific predictions.

Lastly, we demonstrate the effectiveness of our data collection pipeline by comparing CoE with USL-H and USR. Concretely, USL-H is the domain expert in evaluating DailyDialog-related dialogues while USR is the domain expert in evaluating TopicalChat-related dialogues. Both USL-H and USR are trained with dialogue context-response data that are developed with the semantic negative sampling strategy introduced in Section V-B. There is no quality control on the training data of USR and USL-H. On the other hand, CoE, a collection of domain-specific models that have similar architecture as USL-H and USR, is trained with context-response data that are obtained after our quality control process. We can observe in Table III that CoE outperforms USL-H by roughly 2% in terms of the average Spearman correlation over all DailyDialog-related evaluation datasets (row 13). It outperforms USR by roughly 4.5% in terms of the average Spearman correlation over all TopicalChat-related evaluation datasets (row 14).

B. Out-of-Domain Evaluation With PoE

A major limitation of existing model-based dialogue evaluation metrics is their inability to generalize to new domains beyond the training data. This is evidenced by the out-of-domain performance of USL-H, USR, and GRADE in Table IV. For example, USL-H and GRADE perform poorly on the TopicalChat domain while USR performs poorly on the DailyDialog domain.[6]

We can observe that by training on the augmented multi-domain data, the performance of the domain-specific metrics, such as USL-H, GRADE, and USR improves on out-of-domain evaluation datasets. For example, as shown in Table IV, the average Spearman correlation of USR increases from 20.89% (USR*) to 36.57% (USR+). Hence, data augmentation is useful to domain generalization of dialogue evaluation metrics. Yet, PoE significantly outperforms the domain-specific metrics despite that they are trained on the same multi-domain data. This indicates that purely relying on data augmentation techniques is not enough for domain generalization.

Additionally, PoE and PoE-avg significantly outperform Single-T, and MDD-Eval in three out of the five evaluation datasets. MDD-Eval and Single-T are trained on the same multi-domain data as PoE. Their network architecture contains a single RoBERTa encoder and a classifier without specialized adapter modules. The superior performance of PoE indicates that combining data augmentation and adapter-based multitask training leads to better domain generalization than metrics that only rely on data augmentation for domain generalization.

Furthermore, we can observe that CoE performs comparably well to Single-T and better than USL-H, GRADE, USR, and MDD-Eval. The observation suggests that ensemble of multiple domain-specific models can also help boost domain generalization of the metrics. The observation that the performance of PoE is better than CoE suggests the advantage of applying multitask learning on top of model ensemble over pure ensemble of the domain-specific experts.

Moreover, we compare PoE to DEB and Team 5. DEB is pretrained on large-scale Reddit conversations and it leverages a large pre-trained language mode, BERT-large [19]. The metric proposed by Team 5 is based on ensemble of multiple distinct sub-metrics. Both possess good domain generalization, but PoE outperforms them by 6.08% and 1.78% respectively in terms of the average Spearman correlation across the five out-of-domain evaluation datasets. Hence, we can conclude that combining adapter-based multitask learning and model ensemble are complementary. It can lead to better domain generalization.

In summary, the strong out-of-domain performance of PoE can be attributed to two factors. First, PoE leverages data augmentation and multitask training. It learns to capture the regularities within data across domains. The learned knowledge can effectively transfer to unseen dialogue domains. Second, PoE further exploits the advantage of model ensemble, yet in a lightweight manner. In machine learning, we often adopt ensemble models instead of a single model for robust performance. We note that both PoE and CoE take the average of multiple prediction scores for out-of-domain data. However, PoE exploits the lightweight adapters as the hyper-parameters dedicated to effective model generalization, while CoE seeks to generalize with multiple full-fledged domain-specific models.

[6]DailyDialog and TopicalChat have the least overlap in characteristics. One focuses on daily conversations while the other targets knowledge exchanges.

TABLE V
SPEARMAN CORRELATIONS (%) OF MDD-EVAL, SINGLE-T AND POE-AVG AFTER FEW-SHOT TRANSFER LEARNING ON 16 EVALUATION DATASETS WHEN K = 10%, 20%, 30%, AND 40%

	Before Finetuning			K = 10%			K = 20%			K = 30%			K = 40%		
Dataset	MDD	Single-T	PoE-avg	MDD	Single-T	PoE-avg	MDD	Single-T	PoE-avg	MDD	Single-T	PoE-avg	MDD	Single-T	PoE-avg
PUR	54.48	60.33	60.64	54.92	60.25	60.61	55.56	62.00	60.64	59.74	65.12	63.46	60.70	**67.77**	64.84
PZH	59.97	68.57	68.57	64.18	68.79	68.66	64.80	71.78	69.21	68.44	74.02	73.67	71.15	**75.92**	75.35
PDS	37.13	44.77	44.63	43.86	47.52	50.05†	44.91	49.17	54.54†	50.89	51.40	57.24†	48.21	54.38	**58.48**†
CGR	45.31	54.47	57.04	48.32	52.06	57.30†	43.36	53.24	60.10†	52.22	55.44	63.26†	50.64	57.15	**64.55**†
DGR	18.09	26.12	36.41	26.98	39.52	41.05	30.74	48.17	45.54	32.22	53.59	53.44	43.19	58.88	**61.21**
DGP	57.82	60.06	61.23	57.57	61.93	63.57†	60.72	66.52	65.42	59.86	68.52	65.88	63.67	71.39	**73.05**†
DZH	53.40	57.57	59.32	57.13	58.07	60.87†	56.23	61.80	63.11†	60.88	62.94	64.63	62.68	66.50	**68.73**
RDS	28.91	42.71	42.69	36.65	46.11	49.01†	37.64	49.33	51.49†	40.81	49.62	52.82†	44.08	52.88	**53.96**
TUR	51.43	39.89	42.36	61.57	41.99	42.14	59.68	49.29	52.92	58.51	54.11	60.99†	64.41	60.35	**64.51**†
TDS	28.44	31.18	33.14	28.47	34.40	40.32†	38.24	37.68	43.87†	38.91	42.23	46.92†	42.97	42.75	**50.34**†
EGR	37.66	45.33	48.01	44.37	40.23	47.30†	41.55	45.56	47.71	26.44	47.85	47.28	30.11	48.65	**49.05**
FDT*	26.52	36.07	37.29	29.08	32.28	37.86†	16.73	34.11	38.07†	27.49	40.42	42.15	21.97	42.75	**43.20**
HUM*	52.08	65.05	66.79	57.53	68.04	68.32	60.15	69.10	69.38	61.10	70.32	71.17†	64.23	71.83	**72.25**
ESL*	45.32	35.49	42.67	24.35	45.28	48.68†	38.35	45.80	50.70†	27.55	48.22	52.54†	33.38	51.65	**56.42**†
NCME*	22.71	29.37	29.96	22.04	15.17	30.13†	10.10	26.13	30.04†	12.95	28.42	37.63†	24.13	35.92	**43.62**†
ConTurE*	27.76	33.57	34.78	36.76	40.74	46.24†	49.63	47.80	55.62†	53.56	52.16	58.56†	61.89	54.89	**62.91**†
AVG (PC)	49.22	57.03	57.72	52.82	57.15	59.16†	52.16	59.05	61.12†	57.82	61.49	64.41†	57.68	63.81	**65.81**†
AVG (DD)	43.10	47.92	52.32	47.23	53.17	55.16†	49.23	58.83	58.02	50.99	61.68	61.16	56.49	65.59	**67.67**†
AVG (TC)	39.94	35.53	37.75	45.02	38.20	41.23†	48.96	43.48	48.39†	48.71	48.17	53.95†	53.69	51.55	**57.42**†
AVG (EM)	37.66	45.33	48.01	44.37	40.23	47.30†	41.55	45.56	47.71†	26.44	47.85	47.28	30.11	48.65	**49.05**
AVG (RE)	28.91	42.71	42.69	36.65	46.11	49.01†	37.64	49.33	51.49	40.81	49.62	52.82†	44.07	52.88	**53.96**
AVG (Other)	34.88	40.97	43.00	33.95	40.30	46.25†	34.99	44.59	48.76†	36.53	47.91	52.41†	41.12	51.41	**55.68**†
AVG (All)	40.44	45.99	48.07	43.36	46.71	50.76†	44.27	50.83	53.65†	45.72	53.69	56.95†	49.21	56.71	**60.16**†

The best score for each dataset is highlighted in bold. All the scores are statistically significant. † denotes that POE-AVG significantly outperforms Single-T and MDD-EVAL ($p < 0.05$) after few-shot transfer learning. Datasets that are accompanied by an asterisk are the out-of-domain evaluation datasets.

Finally, the out-of-domain performance difference between PoE and PoE-avg is insignificant as evidenced by their average Spearman correlations across the five out-of-domain datasets (42.86% vs 42.37%). This reinforces the claim in Section IV-D that PoE-avg serves as a lightweight alternative to PoE. An additional advantage of PoE over PoE-avg and other baseline metrics is that PoE is much more flexible when evaluating out-of-domain data, because it allows researchers and practitioners to only select the adapters that are closer to the evaluation domains when running inference.

C. Few-Shot Transfer Learning

Besides its strong in-domain performance and zero-shot generalization, PoE has an additional benefit, the exploitation of few-shot transfer learning for fast adaptation to new dialogue domains. In this section, we analyze the few-shot transfer performance of PoE-avg, Single-T, and MDD-Eval. All the models carry the knowledge of dialogues across multiple domains. Yet, the architecture of PoE-avg is more sophisticated due to the incorporation of the adapter. Table V presents the performance comparison among Single-T, PoE-avg, and MDD-Eval. The performance of Single-T, MDD-Eval, and PoE-avg generally improves as K increases. An absolute improvement of 12% is achieved by PoE-avg in terms of the average Spearman correlations over the 16 evaluation datasets when K = 40% (from 48.07% to 60.16%). Single-T and MDD-Eval also attain 10.72% and 8.77% absolute improvement respectively when K = 40%.

With only 5% of the data for finetuning (K = 10%), few-shot transfer learning brings PoE-avg more than 5% absolute improvement on Persona-DSTC10, Reddit-DSTC7, Topical-DSTC10, and ESL. Remarkably, the improvement on ConTurE is the most significant among all datasets (more than 11% improvement). Additionally, PoE-avg outperforms Single-T and MDD-Eval by 6% and 12.3% respectively in terms of average Spearman correlation across all out-of-domain evaluation datasets (row 22). The observations reinforce that PoE-avg is capable of fast adaptation to new dialogue domains.

In general, PoE-avg significantly outperforms both Single-T and MDD-Eval in most of the datasets for all choices of K. The only exceptions are that (1) when K $\geq$ 20%, Single-T performs better than PoE-avg on Persona-USR and Persona-Zhao; (2) When K $\leq$ 20%, MDD-Eval outperforms PoE-avg on Topical-USR. In most cases, the performance of Single-T becomes comparable with that of PoE-avg only when finetuned on more data. This showcases that PoE-avg is better at few-shot transfer and more data-efficient than Single-T.

With 20% of the data for finetuning (K = 40%), PoE-avg can obtain more than 50% Spearman correlations on 13 out of 16 datasets and more than 60% Spearman correlations on 9 datasets. Especially on most of the out-of-domain evaluation datasets (rows 12–16), the performance improvement is significant. On average, there is an improvement of 12.68% Spearman correlations (from 43.00% to 55.68%). From the observations, we can conclude that few-shot transfer learning with PoE-avg offers us a scalable way for the multi-domain dialogue evaluation task. This is because whenever we need to evaluate new dialogues, we may just need to annotate a few in-domain data instances and then finetune the pretrained PoE-avg model with the annotated data. Subsequently, we can apply the finetuned model for the new evaluation task.

D. Application to Downstream Dialogue Tasks

The purpose of developing automatic dialogue evaluation metrics is to benefit downstream dialogue tasks, such as dialogue generation and response selection. With metrics that strongly correlate with human evaluation, researchers and practitioners can accurately estimate the performance of their models during the development phase. Besides providing an accurate quality

TABLE VI
HITS@1 (%) OF POE VARIANTS AND PREVIOUS METHODS ON PERSONA-CHAT UNDER THE THREE PERSONA CONFIGURATIONS

Models	Dialogue-only	Original	Revised
IR Baseline [4]	21.4	41.0	20.7
Starspace [4]	31.8	48.1	32.2
Profile [4]	31.8	47.3	35.4
KV Profile [4]	34.9	51.1	35.1
IMN [80]	63.8	66.7	64.0
DIM [81]	63.8	78.8	70.7
PoE-direct	51.0	57.6	49.3
PoE-adapter	63.6	78.0	68.3

estimation of model responses, PoE can be directly used for response re-ranking or response selection. In this section, we examine whether PoE is an effective response selector on the Persona-Chat benchmark [4].

The Persona-Chat benchmark consists of 8939 dialogues for training, 1000 for validation, and 968 for testing. Response selection is conducted at every turn of the dialogue. Hence, there are 65719 context-response pairs for training, 7801 for validation, and 7512 for test. For each dialogue context, there is a true positive response, which is the original human response, and 19 distractors. The aim of PoE is to rank the true positive response at the top-1 among the 20 response candidates in terms of contextual relevance. Its effectiveness is measured by the recall of the true positive responses, denoted as hits@1, which is a common metric for evaluating response selection systems. Each dialogue in Persona-Chat is accompanied by two persona profiles and each persona profile consists of 3 to 5 sentences describing the background information of the corresponding interlocutor. There are 955 possible personas for training, 100 for validation, and 100 for test. Since the original persona sentences are similar to the utterances within the dialogues, the authors also provide a set of revised persona profiles. In our experiments, we consider three different settings: (1) response selection without persona profile information (denoted as dialogue-only); (2) response selection with the original persona profile (denoted as original); (3) response selection with the revised persona profile (denoted as revised).

In addition, we apply PoE for response selection in two different ways. First, directly run inference on the test set with the pre-trained PoE model (denoted as PoE-direct). Second, freeze the transformer encoder of PoE and train a new adapter on the Persona-Chat training data. Then, we can run inference on the test set with the new adapter module (denoted as PoE-adapter). We compare the two PoE variants against different ranking models, which include as baselines the current SotA models for retrieval approaches from [4], the Interactive Matching Network (IMN) [80], and the Dually Interactive Matching Network (DIM) [81].

Table VI presents the Hits@1 (%) of PoE variants and previous methods on Persona-Chat. We can make the following observations: (1) even though PoE-direct is not finetuned on the Persona-Chat response selection data, it still performs much better than the ranking baselines under the dialogue-only, original, and revised settings. (2) PoE-adapter outperforms IMN in the original and revised settings. Even though PoE-adapter performs slightly worse than DIM, it is much more parameter-efficient. The total number of trainable parameters of DIM is approximately 6.23 M while that of the PoE-adapter is 1.79 M. IMN and DIM require full model training to reach satisfactory performance whereas, for PoE-adapter, we only need to train the light-weight adapter module and reuse its prior knowledge obtained from pre-training on the multi-domain dialogue data. In addition, we did not conduct hyperparameter search to optimize PoE-adapter's performance on the downstream response selection task. Instead, we simply reuse the hyperparameters that are applied to the dialogue evaluation task. (3) PoE-adapter significantly outperforms PoE-direct under the three persona configurations. This observation is expected as training on task-specific data always improves the model performance on the corresponding downstream task.

The experiment results not only showcase the usefulness of PoE to the response selection task but also demonstrate the claim that PoE is much more flexible than single-model metrics. When adapting to a new task-specific dataset, we can just incorporate a task-specific lightweight adapter into PoE. The number of trainable parameters is much less compared to full-model finetuning.

E. Efficiency Analysis

In this subsection, we compare the efficiency of PoE and the baselines. In terms of training, given 2 million data instances and a training batch size of 32 (62500 steps per epoch), PoE takes roughly 7.5 hours to complete one epoch on a single Nvidia RTX 3090 GPU card. Both Single-T and CoE take around 5.5 hours per epoch. Since model training only requires three epochs and the early stopping strategy is implemented, the additional training time is worthwhile considering the superior performance of PoE.

In terms of inference, the speed of Single-T and that of PoE-avg are similar. The difference between PoE and CoE is negligible. However, the inference time of PoE or CoE is N times as that of Single-T or PoE-avg where N denotes the number of domain-specific adapters or domain-specific models. In our experiments, PoE and CoE take 2.5 minutes to complete inference on 10000 data instances with an evaluation batch size of 32 and a single Nvidia RTX 3090 GPU card. Single-T and PoE-avg take around 30 seconds. Hence, in real-life evaluation tasks, PoE-avg is the best choice considering the fact that it is not only efficient but also achieves comparable performance with PoE in terms of correlations with human evaluation.

F. Beyond PoE-avg

Besides parameter averaging, we explore other ways to collapse the domain-specific adapters and classifiers into a single adapter and a classifier respectively. More specifically, we examine two other methods, max pooling of the parameters and

TABLE VII
PERFORMANCE COMPARISON OF POE-MAX, POE-AVG, AND POE-MIN

Row	Datasets	PoE-max	PoE-avg	PoE-min
1	PUR	62.15	61.85	61.91
2	PZH	66.70	67.36	65.73
3	PDS	44.98	44.84	44.45
4	CGR	56.67	57.04	55.98
5	DGR	34.85	36.64	35.34
6	DGP	60.69	61.35	60.54
7	DZH	57.56	58.24	57.42
8	RDS	43.85	44.41	43.80
9	TUR	41.32	41.42	41.13
10	TDS	33.40	33.30	33.20
11	EGR	45.77	46.00	45.69
12	FDT*	35.24	36.26	32.54
13	HUM*	67.24	67.34	67.19
14	ESL*	43.65	42.94	41.94
15	NCME*	30.38	30.55	29.99
16	ConTurE*	34.59	34.74	34.04
17	AVG (PC)	57.63	57.77	57.02
18	AVG (DD)	51.03	52.08	51.10
19	AVG (TC)	37.36	37.36	37.16
20	AVG (EM)	45.77	46.00	45.69
21	AVG (RE)	43.85	44.41	43.80
22	AVG (Other)	42.22	42.37	41.14
23	AVG (All)	47.44	47.77	46.93

Datasets that are accompanied by an asterisk are the out-of-domain evaluation datasets. All the scores are statistically significant.

min pooling of the parameters, which are denoted as PoE-max and PoE-min respectively. Table VII presents the Spearman correlations of PoE-max, PoE-avg, and PoE-min on the 16 dialogue evaluation benchmarks. It can be observed that the three methods are not significantly different in terms of both in-domain (rows 17–21) and out-of-domain (row 22) evaluation performance. PoE-min performs slightly worse than PoE-avg or PoE-max. Hence, both PoE-avg and PoE-max can be used as efficient alternatives to PoE during inference.

VIII. CONCLUSION AND FUTURE WORK

This paper studies multi-domain dialogue evaluation. We address the poor generalization issue of existing model-based automatic dialogue evaluation metrics and propose a novel metric, that is called a panel of experts (PoE), with transformer adapters under multitask learning. To facilitate the training of PoE, we further construct a high-quality multi-domain dataset via data augmentation and pseudo-labeling. Through extensive and comprehensive experiments on a large collection of evaluation datasets, we demonstrate that PoE strongly correlates with human judgments and outperforms the baselines and existing state-of-the-art evaluation metrics. In addition, PoE exhibits strong zero-shot generalization and few-shot transfer performance. In the future, we will extend PoE for multi-dimensional evaluation by incorporating novel pretraining objectives, such as interestingness and informativeness. In addition, we will extend our study from turn-level evaluation to dialogue-level evaluation.

VIII. LIMITATIONS

Despite PoE achieves the best out-of-domain performance among existing model-based metrics, it is still far from perfection. Due to the unconstrained topics, syntactic and semantic expressions, and styles of open-domain dialogues, it is challenging to generalize to all sorts of dialogue data. A future direction is to leverage large-scale language models (LLM), such as GPT-3 [82] and InstructGPT [83], and the prompt-based few-shot learning paradigm [84]. Such LLMs with few human-labeled examples as demonstrations are proven to be capable of superior generalization to unseen data.

Additionally, we assume that when generalizing to unseen data, all the domain experts play an equal role. Hence, an unweighted average of metric scores from all domain-specific experts is adopted. Such an assumption often does not hold when evaluating out-of-domain data. The weights with respect to the domain-specific experts should be adjusted when assessing different evaluation datasets. One possible solution is to learn a regression function of the weights on a small subset of human-labeled data.

Lastly, we only tackle the overall quality of the generated responses. More fine-grained quality dimensions should be examined at both turn level [11] and dialogue level [85]. An ideal automatic dialogue evaluation metric should generalize across both domains and dialogue quality dimensions. A future direction is to ensemble different dimension-specific variants of PoE to provide a holistic evaluation of the dialogue quality.

REFERENCES

[1] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, 2017, vol. 1, pp. 986–995.
[2] K. Gopalakrishnan et al., "Topical-chat: Towards knowledge-grounded open-domain conversations," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1891–1895.
[3] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5370–5381.
[4] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2204–2213.
[5] K. Shuster, D. Ju, S. Roller, E. Dinan, Y.-L. Boureau, and J. Weston, "The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2453–2470.
[6] S. Roller et al., "Recipes for building an open-domain chatbot," in *Proc. 16th Conf. Eur. Chapter, Assoc. Comput. Linguistics*, 2021, pp. 300–325.
[7] S. Mehri and M. Eskenazi, "USR: An unsupervised and reference free evaluation metric for dialog generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 681–707.
[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[9] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2122–2132.

[10] K. Sinha, P. Parthasarathi, J. Wang, R. Lowe, W. L. Hamilton, and J. Pineau, "Learning an unreferenced metric for online dialogue evaluation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2430–2441.

[11] C. Zhang, G. Lee, L. F. D' Haro, and H. Li, "D-Score: Holistic dialogue evaluation without reference," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2502–2516, 2021.

[12] S. Ghazarian, J. Wei, A. Galstyan, and N. Peng, "Better automatic evaluation of open-domain dialogue systems with contextualized embeddings," in *Proc. Workshop Methods Optim. Eval. Neural Lang. Gener.*, 2019, pp. 82–89.

[13] T. Lan, X.-L. Mao, W. Wei, X. Gao, and H. Huang, "PONE: A novel automatic evaluation metric for open-domain generative dialogue systems," *ACM Trans. Inf. Syst.*, vol. 39, no. 1, Nov. 2020, Art. no. 7.

[14] L. Huang, Z. Ye, J. Qin, L. Lin, and X. Liang, "GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 9230–9240.

[15] Y.-T. Yeh, M. Eskenazi, and S. Mehri, "A comprehensive assessment of dialog evaluation metrics," in *Proc. 1st Workshop Eval. Assessments Neural Conversation Syst., Online: Assoc. Comput. Linguistics*, 2021, pp. 15–33.

[16] C. Zhang, L. F. D'Haro, T. Friedrichs, and H. Li, "MDD-Eval: Self-training on augmented data for multi-domain dialogue evaluation," *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 11657–11666.

[17] J. Zhao et al., "Floweval: A consensus-based dialogue evaluation framework using segment act flows," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Dec. 2022, pp. 10 469–10 483.

[18] E. M. Smith, O. Hsu, R. Qian, S. Roller, Y.-L. Boureau, and J. Weston, "Human evaluation of conversations is an open problem: Comparing the sensitivity of various methods for evaluating dialogue agents," in *Proc. 4th Workshop NLP Conversational AI*, May 2022, pp. 77–97.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, vol. 1, pp. 4171–4186.

[20] Y. Liu et al., "RoBERTa A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

[21] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, vol. 97, pp. 2790–2799.

[22] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Representation Learn., Int. Conf. Mach. Learn.*, 2013, Art. no. 896.

[23] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Assoc. for Comput. Linguistics, Jul. 2004, pp. 74–81.

[24] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Meas. Mach. Transl., Summarization*, 2005, pp. 65–72.

[25] V. Rus and M. Lintean, "An optimal assessment of natural language student input using word-to-word similarity metrics," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2012, pp. 675–676.

[26] J. Mitchell and M. Lapata, "Vector-based models of semantic composition," in *Proc. Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2008, pp. 236–244.

[27] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, vol. 1, pp. 654–664.

[28] S. Ghazarian, R. Weischedel, A. Galstyan, and N. Peng, "Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 7789–7796.

[29] C. Zhang, L. F. D'Haro, R. E. Banchs, T. Friedrichs, and H. Li, "Deep AM-FM: Toolkit for automatic dialogue evaluation," in *Conversational Dialogue Systems for the Next Decade*. Berlin, Germany: Springer, 2020.

[30] N. Dziri, E. Kamalloo, K. Mathewson, and O. Zaiane, "Evaluating coherence in dialogue systems using entailment," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, vol. 1, pp. 3806–3812.

[31] C. Zhang et al., "DynaEval: Unifying turn and dialogue level evaluation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, vol. 1, pp. 5676–5689.

[32] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3731–3741.

[33] Y. Nie, M. Williamson, M. Bansal, D. Kiela, and J. Weston, "I like fish, especially dolphins: Addressing contradictions in dialogue modeling," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, vol. 1, pp. 1699–1713.

[34] B. Pang, E. Nijkamp, W. Han, L. Zhou, Y. Liu, and K. Tu, "Towards holistic and automatic evaluation of open-domain dialogue generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3619–3629.

[35] V. Phy, Y. Zhao, and A. Aizawa, "Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4164–4178.

[36] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," 2020, *arXiv:2009.09796*.

[37] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[38] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.

[39] Q. Qi, X. Wang, H. Sun, J. Wang, X. Liang, and J. Liao, "A novel multi-task learning framework for semi-supervised semantic parsing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2552–2560, 2020.

[40] Y. Wang, Y. Li, Z. Zhu, H. Tong, and Y. Huang, "Adversarial learning for multi-task sequence labeling with attention mechanism," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2476–2488, 2020.

[41] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.

[42] M. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in *Proc. 4th Int. Conf. Learn. Representations*, May 2-4, 2016.

[43] Y. Wang, C. Zhai, and H. Hassan, "Multi-task learning for multilingual neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1022–1034.

[44] A. Rastogi, R. Gupta, and D. Hakkani-Tur, "Multi-task learning for joint language understanding and dialogue state tracking," in *Proc. 19th Annu. SIGDIAL Meeting Discourse Dialogue*, 2018, pp. 376–384.

[45] Z. Li, Z. Li, J. Zhang, Y. Feng, and J. Zhou, "Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2476–2483, 2021.

[46] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 506–516.

[47] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. Int. Conf. Learn. Representations*, 2019.

[48] A. C. Stickland and I. Murray, "BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning," in *Proc. 36th Int. Conf. Mach. Learn., Res.*, 2019, vol. 97, pp. 5986–5995.

[49] D. Friedman, B. Dodge, and D. Chen, "Single-dataset experts for multi-dataset question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6128–6137.

[50] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020.

[51] M. Matena and C. Raffel, "Merging models with fisher-weighted averaging," 2021, *arXiv:2111.09832*.

[52] M. Wortsman et al., "Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *Proc. 39th Int. Conf. Mach. Learn., Res.*, K. Research, S. Chaudhuri, L. Jegelka, C. Song, G. Szepesvari Niu, and S. Sabato, Eds., 2022, vol. 162, pp. 23965–23998.

[53] W. W. Daniel, "The spearman rank correlation coefficient," *Biostatistics: A Foundation for Analysis in the Health Sciences*, Wiley, 1987.

[54] G. W. Corder and D. I. Foreman, *Nonparametric Statistics: A Step-By-Step Approach*. Hoboken, NJ, USA: Wiley, 2014.

[55] J. Wang et al., "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng.*, 2022, early access, May 26, 2022, doi: 10.1109/TKDE.2022.3178128.

[56] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends Cogn. Sci.*, vol. 3, no. 4, pp. 128–135, 1999.

[57] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "MAD-X: An adapter-based framework for multi-task cross-lingual transfer," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 7654–7673.

[58] E. Dinan et al., "The second conversational intelligence challenge (CONVAI2)," in *NeurIPS Competition: From Machine Learning to Intelligent Conversations*. Berlin, Germany: Springer, 2020, pp. 187–208.
[59] A. Ghandeharioun et al., "Approximating interactive human evaluation with self-play for open-domain dialog systems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 13658–13669.
[60] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 489–500.
[61] Y. Zhang et al., "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, 2020, pp. 270–278.
[62] P. Gupta, Y. Tsvetkov, and J. Bigham, "Synthesizing adversarial negative responses for robust response ranking and evaluation," in *Findings Assoc. Comput. Linguistics: Assoc. Comput. Linguistics-IJCNLP*, 2021, pp. 3867–3883.
[63] C. Donahue, M. Lee, and P. Liang, "Enabling language models to fill in the blanks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2492–2501.
[64] A. B. Sai, A. K. Mohankumar, S. Arora, and M. M. Khapra, "Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 810–827, 2020.
[65] C. Zhang, J. Sedoc, L. F. D'Haro, R. Banchs, and A. Rudnicky, "Automatic evaluation and moderation of open-domain dialogue systems," 2021, *arXiv:2111.02110*.
[66] T. Zhao, D. Lala, and T. Kawahara, "Designing precise and robust dialogue response evaluators," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 26–33.
[67] P. Gupta, S. Mehri, T. Zhao, A. Pavel, M. Eskenazi, and J. Bigham, "Investigating evaluation of open-domain dialogue systems with human generated multiple references," in *Proc. 20th Annu. SIGDIAL Meeting Discourse Dialogue*, 2019, pp. 379–391.
[68] M. Galley, C. Brockett, X. Gao, J. Gao, and B. Dolan, "Grounded response generation task at DSTC7," in *Proc. AAAI Conf. Artif. Intell. Dialog Syst. Technol. Challenges Workshop*, 2019.
[69] S. Mehri and M. Eskenazi, "Unsupervised evaluation of interactive dialog with DialoGPT," in *Proc. 21th Annu. Meeting Special Int. Group Discourse Dialogue., 1st Virtual Meeting: Assoc. Comput. Linguistics*, 2020, pp. 225–235.
[70] E. Merdivan, D. Singh, S. Hanke, J. Kropf, A. Holzinger, and M. Geist, "Human annotated dialogues dataset for natural conversational agents," *Appl. Sci.*, vol. 10, no. 3, 2020, Art. no. 762.
[71] S. Lee, H. Lim, and J. Sedoc, "An evaluation protocol for generative conversational systems," 2020, *arXiv:2010.12741*.
[72] S. Ghazarian, B. Hedayatnia, A. Papangelis, Y. Liu, and D. Hakkani-Tur, "User response and sentiment prediction for automatic dialogue evaluation," in *Findings Assoc. Computat. Linguist.*, pp. 4194–4204, May 2022.
[73] C. Gunasekara et al., "Overview of the ninth dialog system technology challenge: DSTC9," *Proc. 9th Dialog Syst. Technol. Challenge Workshop*, 2021.
[74] W. S. Gosset, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.
[75] C. Zhang, L. F. D'Haro, Y. Chen, T. Friedrichs, and H. Li, "Investigating the impact of pre-trained language models on dialog evaluation," in *Conversational AI Nat. Human-Centric Interact.*, pp. 291–306, 2021.
[76] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, and R. Garnett, Eds., Red Hook, NY, USA, vol 32, 2019, pp. 8024–8035.
[77] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 38–45.
[78] J. Pfeiffer et al., "AdapterHub: A framework for adapting transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 46–54.
[79] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.
[80] J.-C. Gu, Z.-H. Ling, and Q. Liu, "Interactive matching network for multi-turn response selection in retrieval-based chatbots," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, New York, NY, USA, 2019, pp. 2321–2324.
[81] J.-C. Gu, Z.-H. Ling, X. Zhu, and Q. Liu, "Dually interactive matching network for personalized response selection in retrieval-based chatbots," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 1845–1854.
[82] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., 2020, vol. 33, pp. 1877–1901.
[83] L. Ouyang et al., "Training language models to follow instructions with human feedback," 2022, *arXiv:2203.02155*.
[84] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," 2021, *arXiv:2107.13586*.
[85] C. Zhang, L. F. D'Haro, Q. Zhang, T. Friedrichs, and H. Li, "FineD-Eval: Fine-grained automatic dialogue-level evaluation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 3336–3355.

Chen Zhang (Student Member, IEEE) is currently working toward the Ph.D. degree (final year) with Electrical & Computer Engineering Department, National University of Singapore (NUS), Singapore. He is also associated with Robert Bosch (SEA) under the NUS-Bosch Industrial Ph.D. Programme. His main research interests include dialogue systems, especially automatic dialogue evaluation and open-domain dialogue generation.

Luis Fernando D'Haro (Member, IEEE) is currently an Associate Professor with the Universidad Politécnica de Madrid (ETSIT, UPM), Spain, and a Member of the Speech Technology and Machine Learning Group. His research mainly focuses on spoken dialogue and natural language processing systems, he has written more than 18 international journal publications specifically on dialog systems. Since 2015, Prof. D'Haro has co-organized the WoChat, DBDC and DSTC series of workshops and challenges, which have the common goal of advancing chatbot systems and their automatic evaluation. He was also a Member of the local organizers for Interspeech in 2014, Human Agent Interaction conference, and the International Workshop on Spoken Dialog System Technology (IWSDS) in 2018, and the General Chair for IWSDS 2020. He was a Senior Member for the Chanel workshop at the Johns Hopkins Summer school (JSALT2020). He is currently the Faculty Advisor for Thaurus Team with the Amazon Alexa Prize Grand Challenge (SGC5), co-organizer of the Track four at DSTC11 challenge (multilingual and robust dialogue metrics evaluation), and PI of the EIC funded ASTOUND Project.

Qiquan Zhang (Member IEEE) received the B.Sc. and Ph.D. degrees in electronic science and technology from the Harbin Institute of Technology, Harbin, China, in 2015 and 2020, respectively. Since March 2021, he has been a Postdoctoral Research Fellow with the Human Language Technology Laboratory, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, under the supervision of Prof. Haizhou Li. He is currently a Postdoctoral Research Associate with the Signal Processing Lab, University of New South Wales, Sydney, NSW, Australia, under the supervision of Prof. Eliathamby Ambikairajah and Prof. Haizhou Li. His research interests include statistical signal processing, speech processing, audio-visual speech processing, speech enhancement, noise estimation, machine learning, and deep learning.

Thomas Friedrichs received the Ph.D. degree in nuclear physics from Technische Universität Braunschweig, Braunschweig, Germany, in collaboration with Institute Laue-Langevin - Grenoble, France, in 1998. He is currently the Director of the IT Strategy and Innovation Asia Pacific, Robert Bosch (SEA) Pte Ltd, Singapore.

Haizhou Li (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from the South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990 respectively. He is currently a Presidential Chair Professor with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China. He is also an Adjunct Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Prior to that, he taught with the University of Hong Kong, Hong Kong, (1988–1990) and South China University of Technology, (1990–1994). He was a Visiting Professor with CRIN in France (1994–1995), Research Manager with the AppleISS Research Centre (1996–1998), Research Director with Lernout & Hauspie Asia Pacific (1999–2001), a Vice President with InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology with the Institute for Infocomm Research, Singapore (2003–2016). His research interests include automatic speech recognition, speaker and language recognition, natural language processing. Dr. Li was the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (2015–2018), a Member of the Editorial Board of Computer Speech and Language since 2012, an elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019 and ICASSP 2022. Dr. Li is a Fellow of the ISCA, and a Fellow of the Academy of Engineering Singapore. He was the recipient of the National Infocomm Award 2002, and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and U Bremen Excellence Chair Professor in 2019.