CCDE: A Compact and Competitive Dialogue Evaluation Framework via Knowledge Distillation of Large Language Models

Guanghui Ye, *Graduate Student Member, IEEE*, Huan Zhao, *Member, IEEE*, Bo Li, Haijiao Chen, *Graduate Student Member, IEEE*, Zhixue Zhao, Zhihua Jiang, *Member, IEEE*, and Keqin Li, *Fellow, IEEE*

***Abstract*—Automatic evaluation metrics not only play a vital role in developing dialogue and interactive systems but also have a great impact on social activities in our daily life. However, previous specialized metrics for evaluating dialogues exhibit a relatively low correlation with human judgments. In addition, today's state-of-the-art (SOTA) evaluators that leverage large language models (LLMs) are challenging to deploy in real-world applications due to their sheer size. To this end, we propose a novel evaluation framework, compact and competitive dialogue evaluation (CCDE), which leverages knowledge distillation of LLMs to generate training data and sequentially learn a multitask evaluator regarding diversified quality dimensions. Specifically, we first employ ChatGPT as *teacher* to generate a high-quality and rich-annotation corpus, CCDE-data. Then, we implement a *student* evaluator CCDE (1.3B) via using InstructGPT as the backbone model that is trained and fine-tuned on CCDE-data. We conduct extensive experiments on three public benchmarks: fine-grained evaluation of dialog (FED), PersonaChat, and TopicalChat. The results demonstrate that our model CCDE can outperform the current SOTA model G-Eval which calls GPT-4 ($\geq$ 175B) by 4.3 on the FED dataset, 3.5 on the PersonaChat dataset, and 0.3 on the TopicalChat dataset, in terms of the Spearman correlation metric (%). We release the data and code at: https://anonymous.4open.science/r/ccde-3827.**

***Index Terms*—Dialogue and interactive systems, knowledge distillation, large language models (LLMs), natural language processing (NLP), resource and evaluation.**

Received 10 October 2024; revised 26 April 2025 and 10 June 2025; accepted 12 June 2025. Date of publication 15 July 2025; date of current version 3 December 2025. This work was supported by the National Natural Science Foundation of China NSFC under Grant 62076092. *(Corresponding author: Huan Zhao.)*

Guanghui Ye, Huan Zhao, Bo Li, and Haijiao Chen are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: yghui@hnu.edu.cn; hzhao@hnu.edu.cn; blee@hnu.edu.cn; chenhaijiao@hnu.edu.cn).

Zhixue Zhao is with the Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, U.K. (e-mail: zhixue.zhao@sheffield.ac.uk).

Zhihua Jiang is with the Department of Computer Science, Jinan University, Guangzhou 510632, China (e-mail: tjiangzhh@jnu.edu.cn).

Keqin Li is with the Department of Computer Science, State University of New York at New Paltz, New Paltz, NY 12561 USA, and also with the College of Information Science and Engineering, Hunan University, Changsha 410082, China (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/TCSS.2025.3580272

I. Introduction

EVALUATION metrics shine the light on the best models and thus strongly influence the research directions of social systems [1], [2]. During the past two years, significant advancements have been made in the development of automatic evaluation metrics for computational open-domain dialogue systems [3], [4], [5], [6]. However, there are still challenges. On the one hand, previously specialized metrics are customized for a particular quality *dimension* (also called *aspect*) and exhibit a low correlation with human judgments from an overall perspective [Fig. 1(a)]. On the other hand, today's state-of-the-art (SOTA) evaluators that leverage strong large language models (LLMs) such as ChatGPT [7] or GPT-4 [8] are challenging to deploy because they are memory inefficient and computation intensive for practical applications [Fig. 1(b)].

Existing neural metrics aimed at optimizing a specific dimension. For instance, CMN [9] employed Mutual Information to model text similarity. WeSEE [3] introduced the turn-depth heuristic to evaluate engagingness. However, these metrics struggled to access multifaceted qualities and demonstrated a low level of correlation with human judgments from the general perspective [10], [11]. Recent studies [12], [13], [14] have proposed to use strong LLMs directly as reference-free evaluators. For instance, GPTScore [12] utilized GPT-3 via zero-shot prompts and in-context learning. G-Eval [13] called GPT-4 with chain-of-thoughts (CoT) to evaluate the quality of the generated texts. However, apart from the aforementioned issues, both GPT-3 and GPT-4 are black-box LLMs where model architectures and weights are not accessible, so it is prohibitive to analyze their results like analyzing those of open-sourced LLMs such as Llama [15].

To avoid these challenges in large model deployment, practitioners often choose to deploy smaller models instead [16], [17]. Knowledge distillation can be a powerful tool for reducing the size of large models without compromising their performance [18], [19], [20]. The idea behind is to distill a *teacher* model, in most cases a large and cumbersome model, into a small and efficient *student* model [21], [22], [23], by forcing the student's predictions to match those of the teacher. For instance, DICE-DST [24] transferred the contextual knowledge learned by the teacher encoder to the student encoder, to strengthen

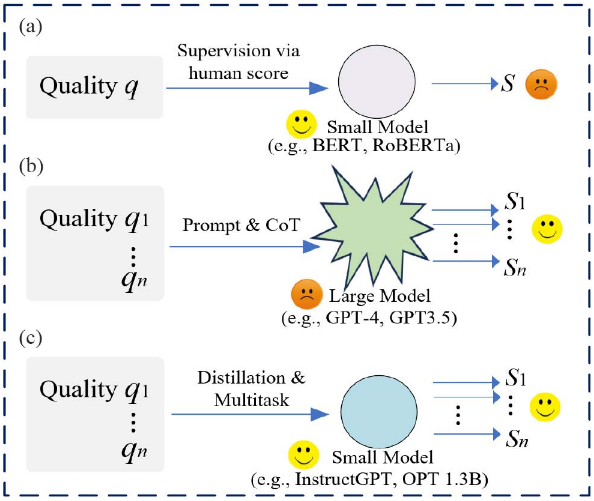

(a)
Quality q
Supervision via human score
S
Small Model (e.g., BERT, RoBERTa)
(b)
Quality q_1 ⋮ q_n
Prompt & CoT
S_1 ⋮ S_n
Large Model (e.g., GPT-4, GPT3.5)
(c)
Quality q_1 ⋮ q_n
Distillation & Multitask
S_1 ⋮ S_n
Small Model (e.g., InstructGPT, OPT 1.3B)

Fig. 1. Illustration of this paper's motivation. (a) Specialized DE metrics focus on optimizing a single dimension but normally have a low correlation. (b) LLM-based DE metrics exhibit a strong correlation but are challenging to deploy. (c) CCDE makes best use of their advantages.

dialogue state tracking (DST) models. As for our task, because existing dialogue evaluation datasets typically have accurate human annotations, it is not necessary to use the LLMs for redundant annotations. Instead, we can employ a *teacher* LLM to generate the new high-quality and rich-annotation data for training a *student* evaluator while keeping from possible human annotators' subjectivity or errors.

Thus, we propose a novel framework, compact and competitive dialogue evaluation (CCDE), which leverages knowledge distillation from LLM and trains a multitask evaluator to assess extensive quality dimensions [Fig. 1(c)]. First, we propose a two-role LLM-based data collection framework that employs ChatGPT as *teacher* to generate a large-scale corpus, CCDE-data. Specifically, we first utilize ChatGPT (as generative expert) to generate a raw dataset and then use it (as verifiable expert) for label validation. Second, we train CCDE as *student* using CCDE-data in a multitask training framework. Specifically, we divide widely used dimensions into groups and build an independent task module for each group. We utilize InstructGPT (1.3B) [25] as a backbone. We experiment on three public datasets: fine-grained evaluation of dialog (FED) [26], PersonaChat (Persona) [27], and TopicalChat (Topical) [28]). The results demonstrate that our CCDE can consistently outperform recent SOTA metrics including CMN, WeSEE, and G-Eval on three datasets. In addition, we conducted an ablation study and in-depth analysis, further verifying the effectiveness of our approach.

Our main contributions are outlined as follows.

1) We propose a two-role LLM-based data collection framework and instantiate it to build a large-scale corpus, CCDE-data (35K), for open-domain dialogue evaluations.
2) We implement a compact and competitive multitask evaluator CCDE (1.3B) regarding diverse quality aspects, with training and fine-tuning InstructGPT on CCDE-data.
3) We experiment with three widely used datasets and results in terms of the Spearman correlation metric (%) show that our CCDE consistently outperforms all compared models, e.g., surpassing G-Eval by 4.3/3.5/0.3 and CMN by 19.2/27.3/12.0 on the FED/Persona/Topical dataset.

The remainder of this article is organized as follows. First, after introducing related work in Section II, we state the problem formulation in Section III. Then we describe our approach, including task analysis, dataset construction, training stage, and inference stage in Section IV. Next, we provide the experimental settings and implementation details in Section V. After that, we present the main results and the analysis in Section VI. In addition, we conduct an in-depth analysis and case study in Sections VII and VIII respectively. Finally, we conclude this article with discussions on theoretical and practical implications and future work in Sections IX and X.

II. Related Work

Automatic evaluation methods [29], [30], [31], [32], [33] are an indispensable part of the effective development of computational social systems such as social networks [34], recommendation systems [35], fake news detection [36], psychological counseling [37], etc. This article focuses on the study of evaluation methods for open-domain dialogue systems [38], [39], which is one of the popular areas of social systems.

A. Dialogue Evaluation Metrics

Evaluation metrics in dialogue systems typically include specialized metrics and LLM-based metrics. Specialized metrics focus on optimizing an individual aspect while reducing the complexity of modeling different LMs [9], [40], [41]. For instance, DEAM [42] was proposed as a `coherent` metric that adopted abstract representation for coherent data generation. DEnsity [4] was treated as a `likeable` metric that used density estimation on the likelihood of token-level words. Recent studies [43], [44], [45] proposed ensemble models that leveraged the strengths of current evaluation models by prompting powerful LLMs. GPTScore [12] used zero-shot transfer and in-context learning to address evaluation needs. Its best version used GPT-3 (175B) [7]. G-Eval [13] is the current SOTA method that uses GPT-4 [8] with CoT, achieving a better correlation than GPTScore. PairEval [14] evaluated responses by comparing their quality with responses in different conversations. Instead of relying on a commercial or proprietary LLM, PairEval is built on top of the moderate-size open-source Llama-2 [15].

B. Knowledge Distillation

Knowledge distillation [46], [47], [48], [49] has been successfully applied to transfer knowledge from larger teacher models to smaller student models affordable for practical applications [16], [20]. Distilling step-by-step [16] trained smaller models that outperformed LLMs and achieved so by leveraging LLM rationales as additional supervision. SODA [18] was the

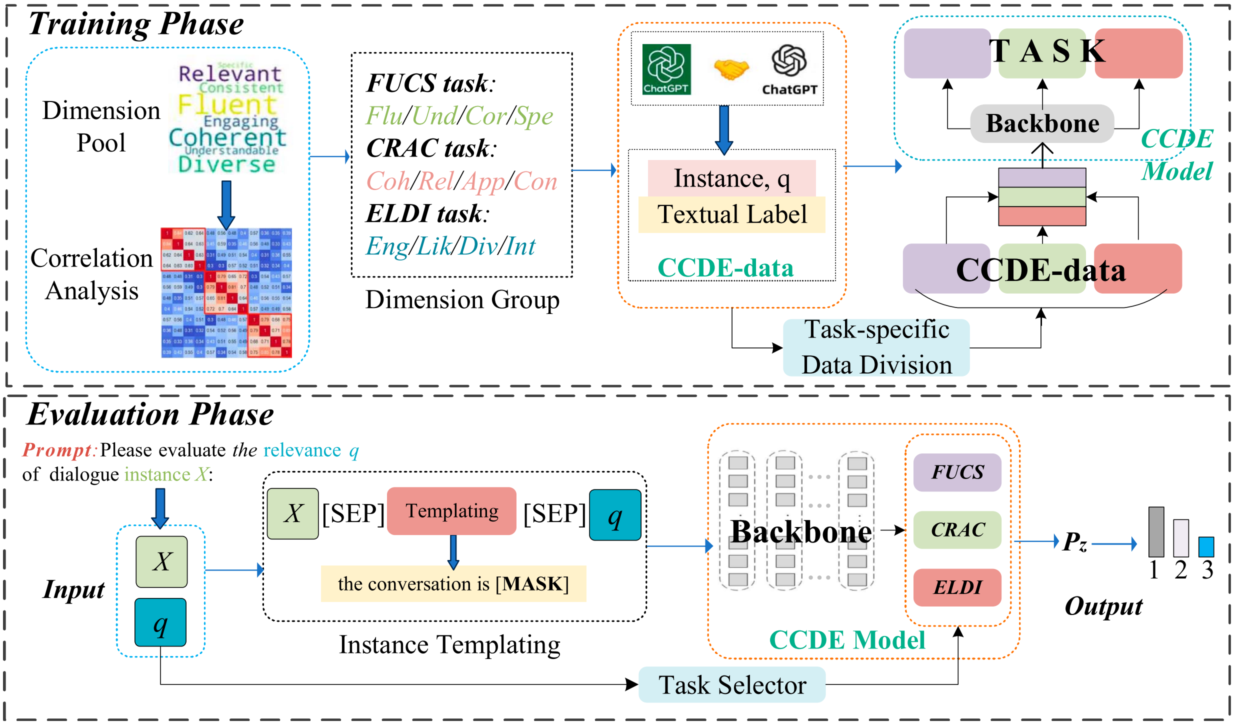


Fig. 2. Two stages adopted in our method. During the training, we analyze the evaluation dimensions, build the CCDE-data dataset and its task-specific subsets, and train the multitask CCDE model. During the evaluation, we use CCDE to find a score for the given input $<X, q>$, where X is a dialogue instance, q is a dimension to be evaluated, and both are filled into a prompt template using the masked token.

first million-scale high-quality social dialogue data set, aimed at distilling a broad spectrum of social interactions from LLMs. Chae et al. [19] proposed a knowledge distillation framework that leverages LLMs as teachers and selectively distills consistent and helpful rationales through alignment filters. Xu et al. [17] proposed a dialogue state distillation network to utilize relevant information from previous dialogue states. Beyer et al. [46] discovered two principles of knowledge distillation for model compression. First, the teacher and student should process the exact same input views. Second, distillation can be interpreted as a task of matching the functions implemented by the teacher and the student. The functions should match on a large number of support points to generalize well. To the best of our knowledge, the introduction of knowledge distillation into automatic dialogue evaluation is still under exploration.

III. Problem Formulation

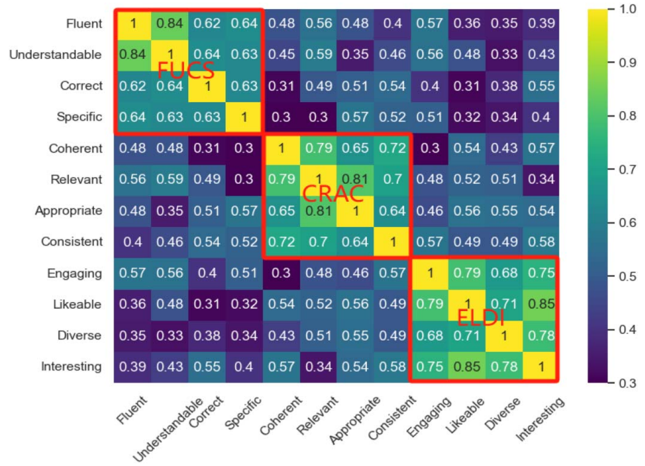
We formally define the evaluation task studied in this article. Suppose that $D = \{d_1, \ldots, d_{|D|}\}$ is a dialogue dataset. To summarize, given an instance $d \in D$ and a queried measurement aspect $q \in Q$, the purpose of CCDE can be defined as $M(d, q) \rightarrow s^q$. Then, a correlation coefficient, denoted as ρ^q, indicating the correlation of predicted scores $S^q = \{s_1^q, \ldots, s_{|D|}^q\}$ with human scores $A^q = \{a_1^q, \ldots, a_{|D|}^q\}$, can be calculated on D. Higher ρ^q suggests better prediction.

IV. Methodology

The overall process of our method is shown in Fig. 2. There are two stages. During training, we first investigate the dimensions of quality of dialogue that are widely used and divide them into three task groups (Section IV-A). Second, we employ ChatGPT to generate the training corpus CCDE-data. Specifically, we first call ChatGPT to generate a raw dataset and then use it again for label validation (Section IV-B). Third, we learn a multitask evaluator CCDE on CCDE-data via using InstructGPT as the backbone model (Section IV-C). During the evaluation, we first incorporate the dialogue instance and the dimension queried via a prompt template and then run CCDE to generate a distribution in the label space. Lastly, the most probable label is outputted as the evaluation result (Section IV-D).

A. Task Analysis

Previous studies [50], [51] show that a multitask model can learn shared representations and capture commonalities among related tasks. Thus, we design multiple evaluation tasks to address the existing diversified quality dimensions. However, rather than treating each dimension as an isolated task like specialized metrics, we divide these dimensions into different groups and regard the evaluation on each group as a unique task. To do this, we first collect twelve widely used dimensions referring to recent studies [1], [12], [13], [41]. Then, we take advantage of the correlation analysis for grouping. We calculate the pairwise Spearman correlation [52] for any two dimensions, as illustrated in Fig. 3. Note that each dimension is viewed as one random variable, and we use their annotation scores on our dataset CCDE-data to calculate the Spearman numbers. Based on the results in Fig. 3, we observe that if we require the lowest correlation number in the same group to be



	Fluent	Understandable	Correct	Specific	Coherent	Relevant	Appropriate	Consistent	Engaging	Likeable	Diverse	Interesting
Fluent	1	0.84	0.62	0.64	0.48	0.56	0.48	0.4	0.57	0.36	0.35	0.39
Understandable	0.84	1	0.64	0.63	0.45	0.59	0.35	0.46	0.56	0.48	0.33	0.43
Correct	0.62	0.64	1	0.63	0.31	0.49	0.51	0.54	0.4	0.31	0.38	0.55
Specific	0.64	0.63	0.63	1	0.3	0.3	0.57	0.52	0.51	0.32	0.34	0.4
Coherent	0.48	0.48	0.31	0.3	1	0.79	0.65	0.72	0.3	0.54	0.43	0.57
Relevant	0.56	0.59	0.49	0.3	0.79	1	0.81	0.7	0.48	0.52	0.51	0.34
Appropriate	0.48	0.35	0.51	0.57	0.65	0.81	1	0.64	0.46	0.56	0.55	0.54
Consistent	0.4	0.46	0.54	0.52	0.72	0.7	0.64	1	0.57	0.49	0.49	0.58
Engaging	0.57	0.56	0.4	0.51	0.3	0.48	0.46	0.57	1	0.79	0.68	0.75
Likeable	0.36	0.48	0.31	0.32	0.54	0.52	0.56	0.49	0.79	1	0.71	0.85
Diverse	0.35	0.33	0.38	0.34	0.43	0.51	0.55	0.49	0.68	0.71	1	0.78
Interesting	0.39	0.43	0.55	0.4	0.57	0.34	0.54	0.58	0.75	0.85	0.78	1

Fig. 3. Correlation matrix of twelve quality dimensions.

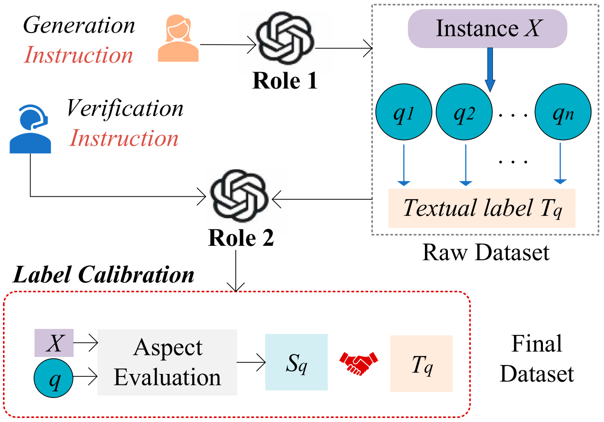


Fig. 4. Construction of CCDE-data.

0.62, then we can divide the original twelve dimensions into three disjoint groups, each containing four strongly correlated dimensions. For example, the lowest correlation numbers in the three resulting groups (FUCS, CRAC, ELDI) are 0.62, 0.64, and 0.68. Following [53], [54], we name each task by using the first letter of each dimension in a group. Thus, we construct three independent evaluation tasks (CRAC, FUCS, ELDI).

B. Dataset Construction

1) Two-Role LLMs Data Collection: We use ChatGPT to implement our two-role LLM-based data collection framework (Fig. 4). First, we use it to generate a raw dataset with labels. Second, to ensure the quality of the training corpus, we utilize it again for label validation and filter inconsistent instances.

a) Role 1: Generative expert: We first design the ChatGPT prompt as a generator. Since a typical evaluation dataset usually includes good and bad samples, we provide a textual quality label T_q, whose value is one of {not, moderately, very}, to indicate the desired quality of a response generated with respect to a specific dimension. In addition, to limit the input length, each conversation is required up to four turns.

Prompt 1: As the Data Generator.

Status: *You are Generative Expert*
Task Description:
Please imitate human and chatbot separately and carry out a conversation to generate human-chabot multi-turn dialogue X_m with the following requirements.
Generation Steps:
1. Each dialog-level multi-turn dialogue X_m has a total of 4 turns (in total 8 utterances, U1-U8) , i.e:[human:U1; chatbot:U2; human:U3; chatbot:U4; human:U5; chatbot:U6; human:U7; chatbot:U8].
2. The quality on Dialogue X_m contains multiple fine-grained aspects including Coherent/Relevant/Appropriate/Consistent. The quality label q of each aspect is randomly selected from three categories: poor, moderate, or good. Meanwhile, the topic T of the generative conversation is randomly selected from one of the three categories of chit-chat, personalized or knowledge-grounded.
3. Split the above generated dialogue X_m into 4 individual human-chatbot turns X_1:(U1,U2); X_2:(U3,U4); X_3:(U5,U6); X_4:(U7,U8).
4. Please follow step 2 and output the fine-grained quality label of each single turn of dialog X_1, X_2, X_3 and X_4. The annotation aspects are 4 FUCS aspects (Fluent/Understandable/Correct/Specific) and 4 ELDI aspects (Engaging/Likeable/Diverse/Interesting).
5. Could you please follow the Example below and output the generated raw dataset X_{raw} consisting of the multi-turn dialog X_m and the 4 labeled aspects, and the single-turn dialog X_1, X_2, X_3, X_4 and their 8 aspects mentioned above.
Example: q ∈ (poor, moderate, good)
Multi-turn dialogue X_m
Coherent:q, Relevant:q, Appropriate:q, Consistent: q
Single turn dialogue X_1: Fluent:q, Understandable:q, Correct:q, Specific:q
Engaging:q, Likeable:q, Diverse:q, Interesting:q
Single turn dialogue X_2: Fluent:q, Understandable:q, Correct:q, Specific:q
Engaging:q, Likeable:q, Diverse:q, Interesting:q
Single turn dialogue X_3: Fluent:q, Understandable:q, Correct:q, Specific:q
Engaging:q, Likeable:q, Diverse:q, Interesting:q
Single turn dialogue X_4: Fluent:q, Understandable:q, Correct:q, Specific:q
Engaging:q, Likeable:q, Diverse:q, Interesting:q

Further, we split each four-turn dialogue into four single-turn sub-dialogues, thus generating single-turn data.

b) Role 2: Verifiable expert: Due to the instability of the LLMs, the quality of the generated texts may not be as expected, possibly leading to negative effects on the model's performance (see Section VI-B for ablation study on the necessity). Thus, we utilize ChatGPT again to perform evaluations for each instance of the raw dataset from the first step. Specifically, we require ChatGPT to generate a score class, one of {1,2,3}. If the score class S_q is inconsistent with the preset quality label T_q, then we will remove the corresponding X.

2) Prompt Design: We provide the detailed prompts for using ChatGPT as two roles, as shown in Prompts 1 and 2 respectively. There are some explanations. 1) A dialogue dataset usually has a topic management [2], e.g., FED is a chat-chit dataset while Persona is a personalized-conversation dataset. Thus, when generating our data with Prompt 1, we provide a topic option, one of {Chit-chat, Persona, knowledge}. 2) To enforce the generation of good samples and bad samples, we also provide a quality option, one of {not, moderately, very} in Prompt 1. We just design such prompt and leave the randomness to the LLMs. 3) Our data construction also involves steps such as transformation and cleaning, so the final CCDE-data is generated by additional manual and algorithmic operations except for the aforementioned prompt design.

3) Data Annotation: Conventionally, dialogue aspects are divided into two levels: turn-level [1] and dialog-level [53].

Prompt 2: As the Data Validator.

Status: *You are Verifiable Expert*
Task Description:
Assume that you are an expert in assessing the quality of conversations. Please assess in detail the quality of a sample of dialog given below.
Raw Instance and Aspect Label: X_{raw}
Validation Steps:
1. As an assessment expert, please carefully analyze each of the aspects of X_{raw}, including the Coherent/Relevant/Appropriate/Consistent of X_m, as well as the 4 FUCS aspects of X1/X2/X3/X4 (Fluent/Understandable/Correct/Specific) and 4 ELDI aspects (Engaging/Likeable/Diverse/Interesting).
2. According to step 1, please give a score label to each aspect of each conversation, ranging from 1-3. The higher the score, the higher the quality. 1 means bad quality, 2 means moderate quality, and 3 means that the conversation is of good quality in the current aspect.
3. You are asked to filter and select samples. If the pre-defined quality label (not/very/good) of X_{raw} in an aspect matches the score label(1/2/3) keep the data labeled, otherwise exclude the sample.
Output: *After the steps 2 and 3 procedures, please output the final dataset, X_{final}, obtained by sample filtering. The sample output format of X_{final} is consistent with X_{raw}.*

A turn-level aspect (e.g., `Correct`) measures the quality of response conditioned on context; by contrast, a dialog-level aspect (e.g., `Coherent`) measures the quality of the throughout dialogue. Since CRAC aspects are usually categorized as dialog-level,[1] we annotate each multiturn dialogue with four CRAC aspects. Instead, both FUCS and ELDI aspects are usually categorized as turn-level, we annotate each single-turn subdialogue with four FUCS aspects and four ELDI aspects, respectively. Based on these LLM-generated dialogues along with their annotations, we further divide CCDE-data into three subsets `CRAC-data`, `FUCS-data`, and `ELDI-data`, which we call *task-specific* data, aiming at optimizing each task only with relevant data to reduce the training cost. For instance, the `FUCS` task module will be trained using only `FUCS-data`.

4) CCDE-Data Examples: Our final dataset contains 7500 4-turn dialogues, each annotated with CRAC dimensions. When dividing a multiturn dialogue into singe-turn subdialogues, we annotate each subdialogue with both FUCS and ELDI dimensions. Therefore, our final CCED-data includes 7500 multiturn dialogues (with CRAC annotations) and 30 000 single-turn dialogues (with FUCS and ELDI annotations). We present some examples in Table I. The whole example is a 4-turn dialogue, accompanied with four CRAC dimensions and their corresponding annotations. For each turn of this dialogue, we exhibit four FUCS dimensions and four ELDI dimensions with their corresponding annotations. Each label value is one of {`not`, `moderately`, `very`}. We use these token-level labels to naturally connect a dialogue instance and an evaluated dimension in the textual template [see (3) for more details].

C. Multitask Training

We adopt InstructGPT (1.3B) as the backbone to implement CCDE. Training details will be given in Section V-E. On top of the backbone model, we implement three task modules (FUCS, CRAC, and ELDI) with small-size LMs, facilitating our model as compact as possible. Each module includes three sequential Transformer blocks and a final fully-connected layer (FCN) via softmax. Specifically, we extract embeddings for each token in the last layer of the backbone and then feed them to the first block. After all Transformer blocks, we extract embeddings for each token in the output layer and then feed it to the FCN. Finally, the FCN outputs a probability distribution in the label space {`not`, `moderately`, `very`}. We select the most probable label as the final answer. Different from task modules using task-specific data (e.g. `CRAC-data`), the backbone model is trained on the whole CCDE-data dataset. The final loss L can be defined as

$$\boldsymbol{L} = w_1\,\mathcal{L}_{\text{CRAC}} + w_2\,\mathcal{L}_{\text{FUCS}} + w_3\,\mathcal{L}_{\text{ELDI}} \tag{1}$$

where $\mathcal{L}$ denotes the Cross Entropy (CE) loss. We set $w_1 = w_2 = w_3 = 1/3$ in our experiments to treat all tasks alike.

D. Evaluation Stage

Given a dialogue instance X and a queried aspect q , CCDE can predict an answer from {`not`, `moderately`, `very`}. We employ OpenPrompt[2] to connect the input and the model.

1) Evaluation Steps: First, we apply the prompt template (denoted by the function f_{prompt}) to rewrite the input pair $<X, q>$ into natural language (NL) sentences X' as the new input

$$X' = f_{\text{Prompt}}(X, q). \tag{2}$$

We design the $f_{\text{prompt}}(\cdot)$ to connect X and q in coherent text via introducing the masked language model (MLM) objective[3]

$$f_{\text{Prompt}}(X, q) = [X] \text{ the conversation is } [\text{MASK}][q]. \tag{3}$$

where [MASK] indicates the answer (i.e., label) slot.

Second, X' is fed to PromptTokenizer for segmentation.

$$X'_{\text{token}} = \text{PromptTokenizer}(X'). \tag{4}$$

Third, X'_{token} is sent to backbone for generating contextual representations. We set $H^{(L)}_{X'}$ to be the concatenation of embeddings of each token from the last layer L of backbone

$$H^{(L)}_{X'} = \text{Backbone}(X'_{\text{token}}). \tag{5}$$

Fourth, $H^{(L)}_{X'}$ is given to a specific task module TM (one of CRAC, FUCS, and ELDI) that will be assigned by the task selector TS based on q, for generating an answer distribution

$$P_{\mathbb{Z}} = \text{TM}(H^{(L)}_{X'}). \tag{6}$$

Last, our model outputs the most probable $\hat{z} \in \mathbb{Z}$. When calculating correlation metrics, we will map a label word into a score class as follows: `not`→`1`, `moderately`→`2`, `very`→`3`.

2) Task Selector: We build a task selector TS, which selects the most proper TM for q, i.e., TM = TS(q). There are two situations. If q (e.g., `Relevant`) is *seen* (or *used*) during the training, then its belonging TM (e.g., CRAC) will be naturally

[1] The level identification is not absolute. For instance, "interesting" is treated as a turn-level dimension on the FED dataset (Table IV) while as a dialog-level dimension on the Persona dataset (Table V).

[2] https://github.com/thunlp/OpenPrompt

[3] We tested several prompt types and reported the experimental results of prompt variations in Table VIII, Section VII-D.

TABLE I
EXAMPLES OF CCDE-DATA

Multi-turn dialogue			
Human: I've been thinking about learning Python. Is it a good language to start with? Chatbot: Yes, Python is beginner-friendly. **Human**: That's good to know. Can you suggest some resources to start with? Chatbot: Blue fish in the sea. **Human**: I didn't get that. Can you recommend Python learning resources? Chatbot: You can try Codecademy or Coursera. **Human**: Great! How about some books for deeper understanding? Chatbot: Python Crash Course is good.			
Dialogue: CRAC-aspects			
Coherent: moderately	Relevant: very	Appropriate: very	Consistent: very
Single-turn dialogues (extracted from the above multi-turn dialogue)			
Human: I've been thinking about learning Python. Is it a good language to start with? Chatbot: Yes, Python is beginner-friendly.			
Turn 1: FUCS-aspects			
Fluent: very	Understandable: very	Correct: very	Specific: very
Turn 1: ELDI-aspects			
Engaging: very	Likeable: moderately	Diverse: moderately	Interesting: moderately
Human: That's good to know. Can you suggest some resources to start with? Chatbot: Blue fish in the sea.			
Turn 2: FUCS-aspects			
Fluent: very	Understandable: not	Correct: not	Specific: not
Turn 2: ELDI-aspects			
Engaging: very	Likeable: moderately	Diverse: moderately	Interesting: not
Human: I didn't get that. Can you recommend Python learning resources? Chatbot: You can try Codecademy or Coursera.			
Turn 3: FUCS-aspects			
Fluent: very	Understandable: very	Correct: very	Specific: very
Turn 3: ELDI-aspects			
Engaging: very	Likeable: moderately	Diverse: moderately	Interesting: very
Human: Great! How about some books for deeper understanding? Chatbot: Python Crash Course is good.			
Turn 4: FUCS-aspects			
Fluent very	Understandable very	Correct very	Specific very
Turn 4: ELDI-aspects			
Engaging very	Likeable moderately	Diverse very	Interesting moderately

assigned. However, if q (e.g., `Natural`) is *unseen*, then we will assign a task module (e.g., FUCS) via comparing the similarity between their corresponding sentence vectors. We apply SentenceBERT (SBERT) [55] to do this. Specifically, we first use SBERT to generate a sentence vector for each dimension q' in Fig. 3 and then take the average of the generated sentence vectors for the group Q where $q' \in Q$

$$E_Q = \text{Avg}(\text{SBERT}(q')). \tag{7}$$

Next, we generate a sentence vector for any *unseen* dimension q via SBERT: $E_q = \text{SBERT}(q)$. Finally, we calculate the cosine similarity between E_Q and E_q and assign the task module that corresponds to Q with the largest similarity.

$$TS(q) = \text{argmax}_Q\{\text{cosine-sim}(E_Q, E_q)\}. \tag{8}$$

V. Experimental Setup

A. Datasets

1) Training Dataset: We use CCDE-data to train CCDE. CCDE-data includes 7500 multiturn dialogues and 30 000 single-turn dialogues. Each multiturn dialogue is annotated with four CRAC aspects. Each single-turn dialogue is annotated with four FUCS aspects and four ELDI aspects. The train/validate split is 7000/500 (multiturn) and 28 000/2000 (single-turn). The total size of the training data is 7000 + 28 000 = 35 000. The sizes of three subsets `CRAC-data`, `FUCS-data`, and `ELDI-data` are 7500, 30 000, and 30 000 accordingly. The average length per conversation/response is 29.3/18.4 tokens.

2) Testing Datasets: We validate CCDE on three open-domain datasets (FED [26], Persona [27], and Topical [28]). FED consists of 124 chit-chat conversations. Each conversation is annotated via both dialog-level and turn-level dimensions. Persona consists of Persona-see and Persona-usr. Persona-see (also called Persona-dialog) [27] contains 3316 personalized conversations. Each conversation is annotated only with dialog-level dimensions. In contrast, Persona-usr (also called Persona-turn) [28] is a much smaller subset, including 300 conversations, each annotated only with turn-level dimensions. The original Topical [56] contains 360 knowledge-grounded human-human conversations where the underlying knowledge spans eight broad topics (Fashion, Politics, Books, Sports, Entertainment, Music, Science & Technology, Movies). For dialogue evaluation, we use a variant of Topical, released by the USR metric [28], where each conversation is annotated with six turn-level dimensions. In our experiments, we select eight turn-/dialog-level dimensions from FED, five turn-/dialog-level dimensions from Persona, and six turn-level dimensions from

TABLE II
DIMENSION DESCRIPTION AND TASK MODULE ASSIGNMENT ON THREE EXPERIMENTAL DATASETS (F - FED; P - PERSONA; T - TOPICAL)

Dimension	Description	Dataset	Task
Fluent	Smooth and natural flow of response.	F / P	FUCS
Correct	Adheres to grammatical rules / proper syntax.	F	FUCS
Specific	Provides detailed and precise information.	F	FUCS
Understandable	Clear and easily comprehensible.	F / P / T	FUCS
Interesting	Captures attention with engaging content.	F / P / T	ELDI
Relevant	Stays focused on the topic or audience interests.	F / P	CRAC
Appropriate	Suitable for the context and audience.	F	CRAC
Engaging	Actively involves the listener.	F / P	ELDI
Coherent	Logical flow and unified structure.	F	CRAC
Consistent	Maintains a steady tone and message.	F	CRAC
Diverse	Includes a variety of perspectives and content.	F	ELDI
Topic Depth	Explores subjects in detail.	F	CRAC
Likeable	Creates a positive and enjoyable experience.	F	ELDI
Informative	Provides valuable and useful information.	F	ELDI
Flexible	Adapts to changes and new information.	F	ELDI
Inquisitive	Encourages curiosity and questions.	F / P	ELDI
Natural	Effortless and reflecting everyday interactions	P / T	FUCS
Maintains	Stays relevant and logically connected to context.	P / T	CRAC
Listening	Actively paying attention and showing empathy.	P	CRAC
Enjoy	Pleasant and satisfying experience.	P	ELDI
Use Knowledge	Demonstrates expertise and shares information.	T	ELDI
Overall	The cumulative effect of all dimensions.	T	FUCS

Topical, respectively. In particular, for each dimension, we provide a brief description, datasets that use it, and our assigned task module in Table II. Given diverse datasets, our method requires no redistilling or retraining.

B. Comparison Models

We compare CCDE with recent specialized metrics and LLM-based metrics.

DEAM [42] is a coherence metric that uses abstract meaning representation for coherent data generation.

WeSEE [3] is an engagingness metric, which uses the remaining depth for each turn as a weak heuristic label.

DEnsity [4] is a likeability metric that uses density estimation to measure the likelihood that a response will occur.

CMN [9] is a similarity metric, which uses mutual information to model the semantic similarity of text.

GPTScore [12] is a typical LLM-as-a-Judge method using a zero-shot prompt. We used GPTScore with GPT-3 (175B).

G-Eval [13] is a framework using powerful LLMs with CoT and a form-filling paradigm. We used G-Eval with GPT-4.

PairEval [14] assesses responses by comparing with other samples. We fine-tuned PairEval that is based on Llama-2 (7B).

C. Meta Evaluation

Meta evaluation aims to evaluate the reliability of automated metrics by calculating how well automated scores ($X = (X_1, ..., X_n)$) correlate with human judgments ($Y = (Y_1, ..., Y_n)$) using correlation functions. In this work, we adopt one widely-used correlation measure: the *Spearman* correlation (ρ) [52] measures the monotonic relationship between two variables based on their ranked values. A stronger correlation indicates greater agreement between prediction and reference

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{9}$$

TABLE III
DETAILS OF INSTRUCTGPT TRAINING STAGES

Stage	SFT	RM	RLHF
Information			
Base-Model	opt-1.3B	opt-330M	opt-1.3B
Training-Data	gptj-pairwi (33.3K)	rm-static (81.3K)	hh-rlhf (169.3K)
Training-Time	63 mins	15 mins	96 mins
Parameter			
Data-Split	2,4,4	2,4,4	-
Epoch	16	1	1
Gradient-Step	8	4	2
Lora-Dim	128	-	128
Weight-Decay	-	0.1	-
Seed	1234	1234	1234

where $d_i = R_{X_i} - R_{Y_i}$, and R_{X_i}, R_{Y_i} denote the rank of X_i, Y_i, respectively.

D. Details of Implementation Settings

Our implementation is based on the open-source Hugging Face Transformers library [57] and Pytorch [58]. We utilize ChatGPT via using API calls to GPT-3.5-Turbo[4] and reproduce InstructGPT (1.3B) following the official instructions.[5] Training details of InstructGPT are provided in Section V-E. We adopt the AdamW optimizer [59] with a learning rate of 3e-5 and a batch size of 16. The training epoch number is set to 30. All the experiments are conducted on 4 nvidia 3090 24GB GPUs.

E. Details of Training InstructGPT

We reproduce the InstructGPT (1.3B) model in the Microsoft/DeepSpeed framework. Following the official instructions, we train Facebook/opt-1.3B as the generation model and Facebook/opt-330M as the reward model, respectively. Specifically, we conduct a three-stage training: Step 1, supervised fine-tuning (SFT); Step 2, reward model fine-tuning (RM); Step 3, reinforcement learning human feedback (RLHF). As a result, we get an executable InstructGPT model. Then, we continuously train InstructGPT from Steps 1∼3 on our dataset CCDE-data (35K) in a multitask framework and generate our final model CCDE. Details of the above three-stage training are described in Table III. It should be noted that the additional training time of InstructGPT on CCDE-data is 48 min.

VI. RESULTS AND ANALYSIS

A. Main Results

We report the Spearman correlation (ρ) numbers regarding diverse aspects on three datasets in Tables IV–VI. Specifically, we first calculate ρ between model scores and human scores for

[4] https://platform.openai.com/docs/models/gpt-3-5-turbo

[5] We reproduced InstructGPT via https://github.com/LanXiu0523/RLHF_instructGPT

TABLE IV
SPEARMAN CORRELATION NUMBERS (%) ON THE FED DATASET

Aspect	Specialized DE				LLM-based DE			Ours
	DEA	WES	DEN	CMN	GPT3	GPT4	PAI	CCDE
FED-dial								
COH	48.2	23.5	42.0	45.9	58.5	64.5	60.6	**66.7**
CON	43.7	22.4	25.6*	37.7	50.3	57.9	57.4	**60.2**
DIV	49.5	30.4	47.6	40.8	56.0	60.3	58.4	**63.0**
TOP	42.6	29.8	33.4	35.6	58.6	53.3	56.4	**61.4**
LIK	47.9	29.6	45.2	42.2	59.1	62.5	60.3	**64.3**
INF	46.4	28.7	44.7	39.4	55.8	61.0	59.7	**63.5**
FLE	36.3	29.4	30.1	34.0*	45.7	53.5	50.0	**64.2**
INQ	39.8	30.2	25.0	29.2	46.4	**54.2**	47.6	52.7
Avg.	44.3	28.0	36.7	38.1	53.8	58.4	56.3	**62.0**
FED-turn								
FLU	22.4	32.6	30.9	40.4	42.1	46.6	48.4	**51.4**
COR	9.6	25.0	25.9	32.4	37.9	43.9	46.2	**47.6**
SPE	12.4	24.5	28.1	33.2	34.7	37.6	42.5	**45.9**
UND	14.8	35.1	32.7	41.2	44.5	47.5	48.7	**53.5**
INT	18.4*	40.3	33.7	38.4	40.6	48.6	49.6	**53.8**
REL	13.5	30.2	29.8	40.2	41.3	51.4	50.3	**57.2**
APP	15.0	24.3	26.7	39.7	40.1	52.1	51.0	**53.9**
ENG	17.9	48.4	34.6	36.9	38.4	50.7	49.8	**55.4**
Avg.	15.5	32.6	30.3	37.8	40.0	47.3	48.3	**52.3**
FED (Full)								
Avg.	29.9	30.3	33.5	38.0	46.9	52.9	52.3	**57.2**

Note: DEA, WES, DEN, CMN, GPT3, GPT4, and PAI denote GRADE, DEAM, WeSEE, DEnisty, CMN, GPTScore, G-EVAL, and PairEval.

TABLE V
SPEARMAN CORRELATION NUMBERS (%) ON THE PERSONA DATASET

Aspect	Specialized DE				LLM-based DE			Ours
	DEA	WES	DEN	CMN	GPT3	GPT4	PAI	CCDE
Persona-dial								
FLU	12.7	22.5	29.5	12.4	46.8	50.5	52.9	**56.2**
LIS	10.5	35.2	25.3	12.4	47.7	51.0	46.3	**51.8**
INT	7.9	21.3	32.6	14.6	46.1	50.8	51.5	**54.9**
ENJ	7.0	40.6	34.7	14.5	55.5	55.7	57.6	**60.3**
INQ	19.4	39.4	22.4	12.1	43.4	48.5	45.4	**49.8**
Avg.	11.5	31.8	28.9	13.2	47.9	51.3	50.7	**54.6**
Persona-turn								
UND	20.7	19.4	24.5	30.8	31.4	39.7	41.2	**44.5**
NAT	21.9	21.4	28.1	32.0	33.0	40.7	40.6	**43.6**
REL	24.9	36.4	35.4	40.4	47.8	51.2	48.7	**54.5**
MAI	22.6	35.5	36.1	38.6*	45.6	48.7	49.0	**50.8**
ENG	23.0	35.8	32.3	40.7	44.9	49.8	50.6	**55.4**
Avg.	22.6	29.7	31.3	36.5	40.5	46.0	46.0	**49.8**
Persona (Full)								
Avg.	17.1	30.8	30.1	24.9	44.2	48.7	48.4	**52.2**

Note: DEA, WES, DEN, CMN, GPT3, GPT4, and PAI denote GRADE, DEAM, WeSEE, DEnisty, CMN, GPTScore, G-EVAL, and PairEval.

each dimension; then, we display the averaged ρ over all dimensions in the "Avg." row for each subset; finally, we calculate the final ρ averaged on both subsets in the last row, showing the evaluation ability of a metric on the full dataset. We highlight the best result of each row (corresponding to a specific aspect) in bold. All values are statistically significant due to *p-value* < 0.05 unless marked by $*$. In particular, due to the writing space limit, we use the first three letters to represent a dimension, e.g., `COH` stands for `Coherent`.

TABLE VI
SPEARMAN CORRELATION NUMBERS (%) ON THE TOPICAL DATASET
(ONLY TURN-LEVEL DIMENSIONS PROVIDED)

Aspect	Specialized DE				LLM-based DE			Ours
	DEA	WES	DEN	CMN	GPT3	GPT4	PAI	CCDE
UND	36.1	29.3	32.4	35.3	39.1	50.7	38.6	52.0
NAT	35.4	27.2	34.9	37.5	40.2	49.5	41.0	50.8
MAI	39.3	34.5	37.3	40.2	47.0	52.6	46.9	50.4
INT	46.2	43.8	46.2	47.6	49.2	55.8	52.5	57.0
KNO	31.6	25.7	30.3	34.9	37.2	44.0	36.6	42.6
OVE	38.7	28.0	35.2	36.4	41.3	49.7	43.4	51.3
Avg.	37.9	31.4	36.1	38.7	42.3	50.4	43.2	50.7

Note: DEA, WES, DEN, CMN, GPT3, GPT4, and PAI denote GRADE, DEAM, WeSEE, DEnisty, CMN, GPTScore, G-EVAL, and PairEval.

1) Superiority of CCDE Regarding the Overall Performance: Observed from Tables IV–VI, G-Eval significantly outperformed other methods including GPTScore. However, indicated by the final "Avg." row, CCDE surpassed G-Eval by 4.3/3.5/0.3 on the FED/Persona/Topical dataset. As for the others such as CMN, CCDE surpassed it by 19.2/27.3/12.0; or WeSEE, CCDE surpassed it by 26.9/21.4/19.3. We observe that PairEval exhibited a comparable performance with G-Eval on FED and Persona, e.g., it lost only 0.6 and 0.3, while performing inferior to G-Eval in knowledge-intensive domains such as Topical due to the powerful ability of GPT-4. In general, CCDE performed best among all compared methods. This provides a strong argument in favor of the strengths of appropriately training a small-size LLM (like ours) or a moderate-size LLM (like PairEval) on specific domain data may bring in a superiority of the proposal compared with large-size LLMs (like GPT-4).

2) Superiority of CCDE Regarding Dimension-Sensitive Evaluations: From Tables IV–VI, we find that CCDE consistently outperformed specialized metrics on their trained dimensions and also outperformed G-Eval on such dimensions. For instance, DEAM is a coherence metric. In the `COH` row of FED-dial, DEAM, G-Eval, and CCDE achieved a Spearman number of 48.2, 64.5, and 66.7 respectively; WeSEE is an engaging metric. In the `ENG` row of Persona-turn, WeSEE, G-Eval, and CCDE achieved 35.8, 49.8, and 55.4; Density is a likeability metric. In the `LIK` row of FED-dial, Density, G-Eval, and CCDE achieved 45.2, 62.5, and 64.3.

B. Ablation Studies

We design ablation variants of the full model as follows.

1) *w/o Data Filtering*: We implement CCDE via using the dataset before and after validation respectively, keeping all other configurations unchanged.
2) *w/o CCDE-data*: We do not use CCDE-data to further train the reproduced InstructGPT. Instead, we directly run the reproduced model after Steps 1$\sim$3 (Table III) on testing data.
3) *w/o Multitask*: We remove the three task modules learned in the multi-task scenario. Instead, we connect InstructGPT to a new LM (only one) with the same structure (3 Transformer blocks and a FCN layer) and optimize it with CCDE-data.

TABLE VII
ABLATION TESTS OF CCDE ON FED AND PERSONA

Model	FED-dial	FED-turn	Persona-dial	Persona-turn	Avg.
Full Model					
CCDE	**62.0**	**52.3**	**54.6**	**49.8**	**54.7**
Ablation Variants					
- w/o Data Filtering	57.2	48.9	48.3	45.3	49.9
- w/o CCDE-data	34.6	33.0	31.7	35.2	33.6
- w/o Multitask	37.1	35.4	33.4	35.8	35.4
- w/o Textual Labels	48.2	45.9	44.3	44.5	45.7

Note: Numbers represent the spearman correlation (%). Bold indicates the best result in all comparisons.

4) *w/o Textual Labels*: We replace the textual labels {`not`, `moderately`, `very`} with the score classes {`1`, `2`, `3`} for our data. Thus, the evaluator can predict a score distribution.

Results in Table VII show that. 1) Data filtering was beneficial. We observe that 8.54% of samples were removed from the raw dataset. Using the validated dataset that finally includes 7500 multiturn dialogues, the model performance raised from 49.9 to 54.7. 2) Knowledge distillation contributed most to the full model. Without the LLM-generated CCDE-data, the model performance was dramatically declined, from 54.7 to 33.6. 3) The contribution of the multitask framework ranked second. Without it, the model performance dropped to 35.4. 4) Using the textual labels was more effective. After we turned to the score classes, the result was reduced to 45.7.

C. Visualization Analysis

The correlation analysis shows how model's judgments align with human judgments from an overall perspective. As a complement, we provide a visualization presentation that shows what are the actual model-generated scores and to what extent they align with human scores. We randomly select 200 samples, 100 from FED and 100 from Persona. We first calculate the average score generated by a specific metric (e.g., CCDE) on all dimensions for each sample. Then, after the normalization, we project these 200 scores into a scatter-plot graph (Fig. 5). We choose six such metrics including CMN, WeSEE, GPTScore, PairEval, G-Eval, and CCDE, for visualization presentation. In particular, we draw `Gold Line` where $y = x$, indicating no difference between prediction and reference. Intuitively, the closer a projection point is to `Gold Line`, the more similar the indicated prediction is to the corresponding reference.

First, we observe that the scatter plots of GPTScore, PairEval, G-Eval, and CCDE were more intensive than those of CMN and WeSEE with respect to (w.r.t.) `Gold Line`, showing that the variances of these four metrics were relatively small on this sample set. Second, we calculate the mean square error (MSE) for further quantitative analysis. Provided this sample set where each score was normalized to [0,1], the MSEs of CMN/WeSEE/GPTScore/PairEval/G-Eval/CCDE were 0.037/0.036/0.028/0.026/0.024/**0.019** on FED, and 0.040/0.039/0.033/0.030/0.029/**0.023** on Persona, respectively. This shows that CCDE scores were more aligned with human scores than all the others on this test.

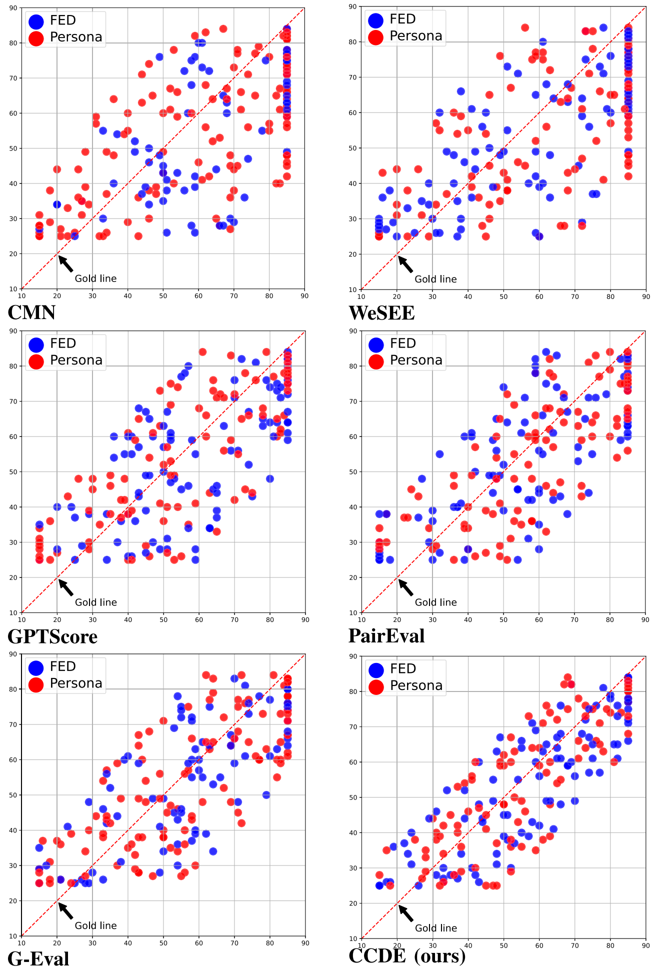


Fig. 5. Visualization. X-axis and Y-axis numbers stand for prediction scores and reference scores respectively. For better illustration, FED samples (in blue) and persona samples (in red) are distinguished via different colors.

VII. In-Depth Analysis

A. Implementing With Different Backbones

To comprehensively evaluate the applicability of our method, we implement CCDE using four backbone models with different model structures and parameter sizes: RoBERTa-large (355M) [60], BART-large (406M) [61], and Llama-2 (7B) [15]. For more comparisons, we also report the benchmark performance of these backbone models. Fig. 6(a) compares the results of CCDE implemented with four backbones on FED. Findings are 1) Each CCDE model (marked in blue) significantly outperformed its corresponding base model (marked in red). 2) CCDE-Llama (7B) achieved a minor improvement over CCDE-InstructGPT (1.3B), e.g., 58.0 versus 57.2. However, due to practical considerations such as training cost and inference time, we select InstructGPT as our default backbone model. 3) Model performance can be greatly affected by the disparity of parameter size. For instance, CCDE-InstructGPT (1.3B) achieved 57.2 while CCDE-BART (406M) only achieved 40.6.

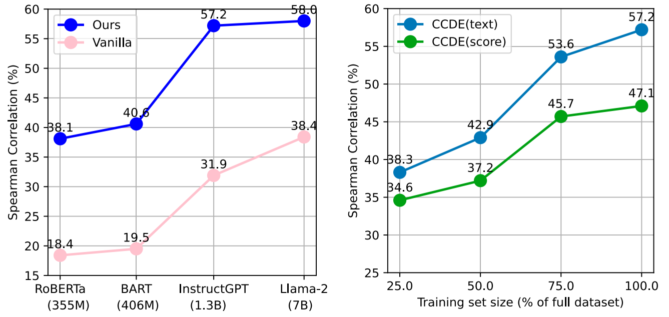


Fig. 6. Results of diverse implementations of CCDE which are tested on FED. (a) Different backbone models including RoBERTa (355M), BART (406M), InstructGPT (1.3B), and llama-2 (7B). (b) Different training data sizes (25%, 50%, 75%, and 100%). A Y-axis number represents a spearman number.

B. Training With Varied Data Sizes

To validate the impact of using the diverse data amount, we implement CCDE with InstructGPT provided varied sizes (25%, 50%, 75%, and 100%) of the whole CCDE-data. To further verify the proper label configuration, we also implement CCDE under two label settings: textual labels (i.e., CCDE-text) and score labels (i.e., CCDE-score). Results in Fig. 6(b) show that: 1) larger training-data size led to better performance; 2) provided any training size, CCDE-text consistently outperformed CCDE-score; and 3) CCDE-text trained with 75% data achieved 53.6 while GPTScore only achieved 46.9 (see Table IV).

C. Improving Fine-Grained Evaluations

To show the improvement of our model on the main backbone in a specific dimension, we compare the results of CCDE and InstructGPT on the two subsets of FED respectively. From Fig. 7, we observe that: 1) compared with InstructGPT, CCDE consistently improved the evaluations on all fine-grained dimensions of both FED-turn and FED-dialog; and 2) CCDE exhibited a stable performance no matter which aspect was evaluated. For instance, InstructGPT performed poorly on `Flexible` with a low ρ number (e.g., 23.8) while CCDE still enabled to achieve a moderate ρ number (e.g., 64.2).

D. Experiments With Prompt Variations

We design the $f_{\text{prompt}}(\cdot)$ function (3) via enforcing the masked token (*mask*) before a dimension name, which we call the MLM-prompt to distinguish it from other prompt attempts. The answer can be extracted from *mask* via optimizing the MLM objective. Intuitively, the MLM-prompt aims at facilitating the whole prompt text more natural and fluent. To validate this benefit, we attempt another prompt template which may better match with the causal nature of GPT models. Specifically, we rephrase the MLM-prompt such that *mask* is the last token in the form of question answering (QA), which we call the QA-prompt, e.g., “[X] the conversation is interesting? [*answer*]”, where *answer* is equivalent to *mask*. These additional experiments were conducted on two studied datasets. Results in Table VIII show that the MLM-prompt performed better than the QA-prompt, providing a strong argument in favor of the strengths of using natural and fluent text in the prompt.

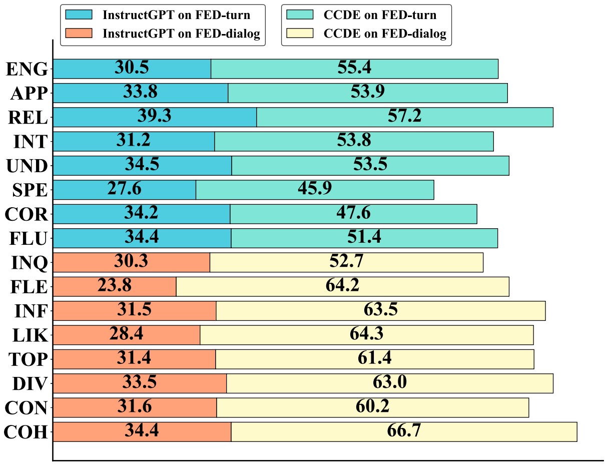


Fig. 7. Improvements on fine-grained evaluations.

TABLE VIII
COMPARISON ON DIFFERENT PROMPT DESIGN

Dataset	FED-dial	FED-turn	Persona-dial	Persona-turn	Avg.
CCDE					
MLM-prompt	**62.0**	**52.3**	**54.6**	**49.8**	**54.7**
QA-prompt	51.4	49.3	45.8	41.0	46.9

E. Improving Challenging Dimensions

Observed in Tables IV–VI, assessing dimensions such as “Understandable” remains a challenge (e.g., low correlation with human ratings achieved by all models) due to difficulties in capturing accurate semantic descriptions. Following [16], we further explore using LLM-generated rationales for additional supervision to improve the assessment on such dimensions. LLMs can produce rationales in natural language that justify their predicted labels, e.g., “The response is hard to understand because the message conveyed is unclear and hardly comprehensible to the user”. We therefore utilize these generated rationales as additional and richer information to train the backbone models through a multitask training setup (10), with both label prediction and rationale prediction tasks

$$\mathcal{L} = \mathcal{L}_{\text{label}} + \lambda\mathcal{L}_{\text{rational}} \tag{10}$$

where $\mathcal{L}_{\text{label}}$ is the label prediction loss in (1) and $\mathcal{L}_{\text{rational}}$ is the rationale generation loss. Specifically, we first generate additional rationales to justify all “Understandable” scores in our CCDE-data. Then, we leverage these rationales in addition to the task labels to train our backbone model, InstructGPT. Compared with the results reported in Tables IV–VI, we achieve a minor improvement on the assessment of the “Understandable” dimension (e.g., from 44.5 to 45.3 on the Persona dataset).

TABLE IX
CASE STUDY ON FED

Multi-turn dialogue

Human: I want to talk to you normally. Chatbot: You are who I want to talk with too.
Human: Should I stop being lazy and go into work? Chatbot: Trust your instincts.
Human: My instincts say that it is cold. Chatbot: You just repeated yourself.
Human: Yes, because you just asked me the same thing. Chatbot: That's not what I meant.

Dialog-level dimensions

	COH	CON	DIV	TOP	LIK	INF	FLE	INQ	Avg.
Human	4.00	3.00	3.33	3.33	2.33	3.00	3.00	1.50	2.93
G-Eval	4.22	2.87	3.54	3.52	2.18	2.82	3.61	2.10	3.11
CCDE	4.13	2.85	3.41	3.38	2.29	2.88	3.45	1.90	3.04

Turn-level dimensions

	FLU	COR	SPE	UND	INT	REL	APP	ENG	Avg.
Human	4.50	4.00	3.00	5.00	3.00	5.00	3.67	4.20	4.04
G-Eval	4.44	4.30	3.70	4.80	3.30	4.45	3.82	4.26	4.13
CCDE	4.35	4.13	3.37	4.50	3.24	4.67	3.70	3.98	3.99

Note: For each dimension, the model score that is more aligned with human score is underlined.

TABLE X
CASE STUDY ON PERSONA

Multi-turn dialogue

Human: Hello there how are you?. Chatbot: I am good how are you?
Human: I'm wonderful, were are you from? Chatbot: I am. What do you do for a living?
Human: I'm a mechanical engineer. Chatbot: What city are you from?
Human: I'm from Colorado. Chatbot: What city are you from?
Human: I'm from Denver how about you? Chatbot: I'm from the united states.
Human: Cool, What do you do for a living? Chatbot: I am a stay at home mom.

Dialog-level dimensions

	FLU	LIS	INT	ENJ	INQ	Avg.
Human	3.00	3.00	2.50	2.00	2.00	2.50
G-Eval	3.37	4.08	2.87	2.52	2.18	3.00
CCDE	3.14	3.94	2.94	2.35	2.14	2.90

Turn-level dimensions

	UND	NAT	REL	MAI	ENG	Avg.
Human	5.00	5.00	5.00	5.00	5.00	5.00
G-Eval	4.72	4.48	4.53	4.68	4.57	4.59
CCDE	4.46	4.50	4.83	4.71	4.61	4.62

Note: For each dimension, the model score that is more aligned with human score is underlined.

VIII. Case Study

To perform a point-wise comparison, we show two case studies in Tables IX and X, respectively. Since G-Eval uses the 1-5 Likert scale, we map our scores to this scale for intuitive comparisons. As for human scores, we calculate an averaged score if there are multihuman annotators in the same scenario. We observe that: 1) our model CCDE scored better than G-Eval on most dimensions, e.g., 12 out of 16 on the FED case (Table IX) and 8 out of 10 on the Persona case (Table X); 2) suggested by the last "Avg." column, each averaged score of CCDE was consistently more aligned with the corresponding human score than that of G-Eval. These case studies further validate the superiority of our approach.

IX. Theoretical and Practical Implications

Theoretically, this work focuses on a fundamental but hot research topic on exploring automatic evaluation metrics for dialogue systems, which is rather significant to understand the correspondences between conversation generation and evaluation across natural language processing (NLP) and the computational social systems communities. Because human evaluations for natural language generation (NLG) are both expensive and time-consuming, relevant and meaningful automatic metrics that strongly correlate with human judgments are crucial. Practically, this work proposes and implements an open-domain dialogue evaluation metric that follows the expectations, e.g., general-purpose, memory-efficient, and accurate, on people's daily conversations and other social interactions. In addition, our approach is flexible because the model can be implemented with different backbones and trained with limited data, making it possible for future improvements.

X. Conclusion and Future Works

In this article, we generate a large-scale (35K) dataset CCDE-data and train a compact (1.3B) and competitive (surpassing G-Eval) evaluation model CCDE, benefiting from knowledge distillation of the LLMs, the multitask learning, and the prompt design proposed for dialogue evaluations. In addition to the contribution of a novel evaluation method, our research provides an affordable and robust framework for data collection and model training to promote the NLP community. Future work includes more explorations of diversified datasets to further validate the applicability and superiority of our proposal.

Acknowledgment

The authors thank anonymous referees for their invaluable insights.

References

[1] Z. Jiang, G. Ye, D. Rao, D. Wang, and X. Miao, "im^2: an interpretable and multi-category integrated metric framework for automatic dialogue evaluation," in *Proc. EMNLP Conf.* 2022, pp. 11091–11103.

[2] Y. Chen, N. Nishida, H. Nakayama, and Y. Matsumoto, "Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations," in *Proc. LREC/COLING Conf.*, 2024, pp. 13650–13665.

[3] S. Jiang, S. Vakulenko, and M. de Rijke, "Weakly supervised turn-level engagingness evaluator dialogues," in *Proc. CHIIR Conf.*, 2023, pp. 258–268.

[4] C. Park, S. C. Lee, D. Rim, and J. Choo, "Density: open-domain dialogue evaluation metric using density estimation," in *Proc. Findings ACL Conf.*, 2023, pp. 14222–14236.

[5] L. Wang et al., "Dialogue summarization enhanced response generation for multi-domain task-oriented dialogue systems," *Inf. Process. Manage.*, vol. 61, no. 3, 2024, Art. no. 103668.

[6] X. Li, J. Su, Y. Yang, Z. Gao, X. Duan, and Y. Guan, "Dialogues are not just text: model. cognition dialogue coherence evaluation," in *Proc. AAAI Conf.*, 2024, pp. 18573–18581.

[7] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS Conf.*, 2020.

[8] OpenAI, "GPT-4 Technical Report," 2023, *arXiv:2303.08774*.

[9] K. Zhao, B. Yang, C. Lin, W. Rong, A. Villavicencio, and X. Cui, "Evaluating open-domain dialogues in latent space with next sentence prediction mutual information," in *Proc. ACL Conf.*, 2023, pp. 562–574.

[10] Y. Yeh, M. Eskénazi, and S. Mehri, "A comprehensive assessment of dialog evaluation metrics," 2021, *arXiv:2106.03706*.

[11] M. Kim and J. Kim, "A study on automatic open-domain dialogue evaluation metrics," in *Proc. ICCE Conf.*, 2024, pp. 1–3.

[12] J. Fu, S. Ng, Z. Jiang, and P. Liu, "GPTscore: Evaluate as you desire," 2023, *arXiv:2302.04166*.

[13] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proc. EMNLP Conf.*, 2023, pp. 2511–2522.

[14] C. Park, M. Choi, D. Lee, and J. Choo, "PairEval: open-domain dialogue evaluation with pairwise comparison," 2024, *arXiv:2404.01015*.

[15] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[16] C. Hsieh et al., "Distilling step-by-step! outperforming larger language models with less training data smaller model sizes," in *Proc. Findings ACL Conf.*, 2023, pp. 8003–8017.

[17] J. Xu et al., "Dialogue state distillation network with inter-slot contrastive learning dialogue state tracking," in *Proc. AAAI Conf.*, 2023, pp. 13834–13842.

[18] H. Kim et al., "SODA: million-scale dialogue distillation with social commonsense contextualization," in *Proc. EMNLP Conf.*, 2023, pp. 12930–12949.

[19] H. Chae et al., "Dialogue chain-of-thought distillation for commonsense-aware conversational agents," 2023, *arXiv:2310.09343*.

[20] H. Qiu, S. Zhang, H. He, A. Li, and Z. Lan, "Facilitating pornographic text detection for open-domain dialogue systems via knowledge distillation of large language models," 2024, *arXiv:2403.13250*.

[21] X. Guo, W. Zhou, and T. Liu, "Multilevel attention imitation knowledge distillation for RGB-thermal transmission line detection," *Expert Syst. Appl.*, vol. 260, 2025, Art. no. 125406.

[22] G. Wang, J. Huang, Y. Lai, and C. Vong, "Dealing with partial labels by knowledge distillation," *Pattern Recognit.*, vol. 158, 2025, Art. no. 110965.

[23] K. Zeng, Z. Wan, H. Gu, and T. Shen, "Self-knowledge distillation enhanced binary neural networks derived from underutilized information," *Appl. Intell.*, vol. 54, no. 6, pp. 4994–5014, 2024.

[24] J. Guo, K. Shuang, K. Zhang, Y. Liu, J. Li, and Z. Wang, "Learning to imagine: Distillation-based interacting context exploitation dialogue state tracking," in *Proc. AAAI Conf.*, 2023, pp. 12845–12853.

[25] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. NeurIPS Conf.*, 2022,

[26] S. Mehri and M. Eskénazi, "Unsupervised evaluation interacting dialog with dialoGPT," in *Proc. SIGdial Conf.*, 2020, pp. 225–235.

[27] A. See, S. Roller, D. Kiela, and J. Weston, "What makes a good conversation? how controllable attributes affect human judgments," in *Proc. NAACL-HLT Conf.*, 2019, pp. 1702–1723.

[28] S. Mehri and M. Eskénazi, "USR: An unsupervised reference free evaluation metric dialog generation," in *Proc. ACL Conf.*, 2020, pp. 681–707.

[29] J. Shu, Y. Liang, W. Ma, and L. Liu, "Key nodes evaluation method based on combination weighting VIKOR in social networks," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 4, pp. 5404–5418, Aug. 2024.

[30] Q. Xie et al., "Efficiency evaluation of insurance companies from multiperiod perspective," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 2, pp. 2656–2674, Apr. 2024.

[31] J. Shen et al., "A novel intelligence evaluation framework: Exploring the psychophysiological patterns of gifted students," *IEEE Trans. Comput. Soc. Syst*, vol. 11, no. 2, pp. 2036–2045, Apr. 2024.

[32] A. Daud, S. Ghaffar, and T. Amjad, "Citation count is not enough: Citation's context-based scientific impact evaluation," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 4, pp. 4567–4573, Aug. 2024.

[33] H. Mo, H. Hu, J. Hu, Y. Li, X. Wang, and F. Wang, "Interval type-2 fuzzy risk evaluation and prevention for parallel breast cancer treatment system," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 2, pp. 673–685, Apr. 2023.

[34] S. Chen, W. Yang, and S. Gao, "Positive evaluation maximization in social networks: Model and algorithm," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 3, pp. 1402–1413, Jun. 2023.

[35] X. Zheng, Z. Ni, X. Zhong, and Y. Luo, "Kernelized deep learning for matrix factorization recommendation system using explicit and implicit information," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 1205–1216, Jan. 2024.

[36] Y. Djenouri, A. Belhadi, G. Srivastava, and J. C. Lin, "Advanced pattern-mining system for fake news analysis," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 6, pp. 2949–2958, Dec. 2023.

[37] R. Vatsal, S. Mishra, R. Thareja, M. Chakrabarty, O. Sharma, and J. Shukla, "An analysis of physiological and psychological responses in virtual reality and flat screen gaming," *IEEE Trans. Affect. Comput*, vol. 15, no. 3, pp. 1696–1710, Jul./Sep. 2024.

[38] F. Wang, X. Zhao, and X. Sun, "SHADE: speaker-history-aware dialog generation through contrastive and prompt learning," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 2, pp. 2302–2312, Apr. 2024.

[39] M. Firdaus, A. P. Shandilya, A. Ekbal, and P. Bhattacharyya, "Being polite: Modeling politeness variation in a personalized dialog agent," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 4, pp. 1455–1464, Aug. 2023.

[40] L. Huang, Z. Ye, J. Qin, L. Lin, and X. Liang, "GRADE: Automation graph-enhanced coherence metric evaluating open-domain dialogue system," in *Proc. EMNLP Conf.*, 2020, pp. 9230–9240.

[41] Z. Ding, Z. Yang, and H. Lin, "A plug-and-play adapter for consistency identification in task-oriented dialogue systems," *Inf. Process. Manage.*, vol. 61, no. 3, 2024, Art. no. 103637.

[42] S. Ghazarian, N. Wen, A. Galstyan, and N. Peng, "DEAM: dialogue coherence evaluation using AMR-based semantic manipulations," in *Proc. ACL Conf.*, 2022, pp. 771–785.

[43] J. Mendonça, P. Pereira, J. P. Carvalho, A. Lavie, and I. Trancoso, "Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation," 2023, *arXiv:2308.16797*.

[44] H. Duan et al., "BOTChat: Evaluating LLMS' capabilities of having multi-turn dialogues," 2023, *arXiv:2310.13650*.

[45] Y. Liu et al., "Are LLMS good at structured outputs? A benchmark for evaluating structured output capabilities in LLMS," *Inf. Process. Manage.*, vol. 61, no. 5, 2024, Art. no. 103809.

[46] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient consistent," in *Proc. CVPR Conf.*, 2022, pp. 10915–10924.

[47] C. Yang, Z. An, L. Cai, and Y. Xu, "Knowledge distillation using hierarchical self-supervision augmented distribution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2094–2108, Feb. 2024.

[48] F. Ding, Y. Yang, H. Hu, V. Krovi, and F. Luo, "Dual-level knowledge distillation via knowledge alignment and correlation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2425–2435, Feb. 2024.

[49] C. K. Joshi, F. Liu, X. Xun, J. Lin, and C. S. Foo, "On representation knowledge distillation for graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 35, no. 4, pp. 4656–4667, Apr. 2024.

[50] Y. Chen, Y. Zhang, C. Zhang, G. Lee, R. Cheng, and H. Li, "Revisiting self-training few-shot learning language model," in *Proc. EMNLP Conf.*, 2021, pp. 9125–9135.

[51] X. Gao, C. Gupta, and H. Li, "Automatic lyrics transcription of polyphonic music with lyrics-chord multi-task learning," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2280–2294, 2022.

[52] J. H. Zar, "Spearman rank correlation," *Encyclopedia bioStatist.*, vol. 7, no. 1, pp. 701–715, 2005.

[53] C. Zhang, L. F. D'Haro, Q. Zhang, T. Friedrichs, and H. Li, "Fined-Eval: Fine-grained automatic dialogue-level evaluation," in *Proc. EMNLP Conf.*, 2022, pp. 3336–3355.

[54] M. Rodríguez-Cantelar et al., "Overview of robust and multilingual automatic evaluation metrics for open-domain dialogue systems at DSTC 11 track 4," 2023, *arXiv:2306.12794*.

[55] N. Reimers and I. Gurevych, "Sentence-Bert: Sentence embeddings using siamese Bert-networks," in *Proc. EMNLP Conf.*, 2019, pp. 3980–3990.

[56] K. Gopalakrishnan et al., "Topical-chat: Towards knowledge-grounded open-domain conversations," in *Proc. Interspeech Conf.*, 2019, pp. 1891–1895.

[57] T. Wolf et al., "Transformers: state-of-the-art natural language processing," in *Proc. EMNLP Conf.*, 2020, pp. 38–45.

[58] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS Conf.*, 2019, pp. 8024–8035.

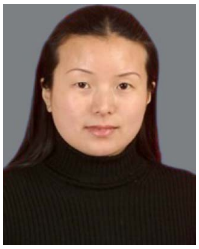
[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR Conf.*, 2019, pp. 1001–1018.

[60] M. Lewis et al., "BART: denoising sequence-to-sequence pre-training natural language generative, translation, comprehension," in *Proc. ACL Conf.*, 2020, pp. 7871–7880.

[61] Y. Liu et al., "Roberta: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

Guanghui Ye (Graduate Student Member, IEEE) received the M.S. degree in computer science and technology from the Department of Computer Science, Jinan University, Guangzhou, China, in 2023. He is currently working toward the Ph.D. degree in computer science with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China.

His research interests include computational linguistics, dialog systems, large language models, resource, and evaluation. His publications have been published at top academic conferences such as *Empirical Methods in Natural Language Processing* and *North American Chapter of the Association for Computational Linguistics*.



Huan Zhao (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science and technology from Hunan University, Changsha, China, in 1989, 2004, and 2010, respectively.

Currently, she is a Professor with the School of Information Science and Technology, Hunan University. Her research interests mainly include speech signal processing, cross-media retrieval, and natural language processing.

Dr. Zhao has published over 100 research papers in top-tier international journals and conferences, covering various high-impact academic platforms, including journals such as IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), *IEEE Internet of Things Journal (IoTJ)*, and conferences such as Annual Meeting of the *Association for Computational Linguistics (ACL)*, Conference of the *North American Chapter of the Association for Computational Linguistics (NAACL)*, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, etc.

Bo Li received the B.S. degree in computer science and technology from the School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China, in 2016. She is currently working toward the Ph.D. degree in computer science with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China.

Her research interests include dialog systems, emotion recognition, and multimedia computing.

Haijiao Chen (Graduate Student Member, IEEE) received the M.S. degree in computer application technology from Xinjiang University, Urumqi, China, in 2017. He is currently working toward the Ph.D. degree in computer science with Hunan University, Changsha, China.

His research interests include affective computing, multimodal analysis, federated learning, and deep learning. He has published papers in important journals and conferences such as International *Conference on Acoustics, Speech, and Signal Processing* and *IEEE Internet of Things Journal*.

Zhixue Zhao received the Ph.D. degree in natural language processing from the University of Sheffield, Sheffield, U.K.

Currently, she is a Lecturer in natural language processing with the Computer Science Department, the University of Sheffield. Previously, she was a Postdoctoral Researcher on explainable AI and responsible AI. Her long-term research goal is to enable trustworthy, responsible, and efficient NLP models. These days, she is interested in anything related to interpretability and large language models (LLMs). Her publications have been published at top academic conferences such as ACL, NAACL and EMNLP. Her research interests include model compression, model editing, and text-to-image models.

Zhihua Jiang (Member, IEEE) received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2008.

Currently, she is an Associated Professor with the Department of Computer Science, Jinan University, Guangzhou, China. Her publications have been published at top academic conferences such as *Empirical Methods in Natural Language Processing* and *North American Chapter of the Association for Computational Linguistics*. Her research interests include natural language processing, information retrieval, data mining, and multimedia computing.

Keqin Li (Fellow, IEEE) received the B.S. degree from Tsinghua University, Beijing, China, in 1985 and the Ph.D. degree from the University of Houston, Houston, TX, USA, in 1990, both in computer science.

Currently, he is a SUNY Distinguished Professor with the State University of New York, NY, USA and a National Distinguished Professor with Hunan University, Changsha, China. He has authored or co-authored more than 1000 journal articles, book chapters, and refereed conference papers.

Dr. Li received several best paper awards from international conferences including *Parallel and Distributed Processing Techniques and Applications (PDPTA 1996)*, *National Aerospace and Electronics Conference (NAECON 1997)*, *International Symposium on Parallel and Distributed Processing (IPDPS 2000)*, *International School Psychology Association (ISPA 2016)*, *Network and Parallel Computing (NPC 2019)*, *International School Psychology Association (ISPA 2019)*, and *International Conference on Cyber, Physical and Social Computing (CPSCom)*. He holds nearly 75 patents announced or authorized by the Chinese National Intellectual Property Administration. He is among the world's top five most influential scientists in parallel and distributed computing in terms of single-year and career-long impacts based on a composite indicator of the Scopus citation database. He was a 2017 recipient of the Albert Nelson Marquis Lifetime Achievement Award for being listed in Marquis *Who's Who in Science and Engineering, Who's Who in America, Who's Who in the World, and Who's Who in American Education* for over twenty consecutive years. He received the Distinguished Alumnus Award from the Computer Science Department at the University of Houston in 2018. He received the *IEEE Technical Community on Cloud Computing Research Impact Award* from the *IEEE CS Technical Committee on Cloud Computing* in 2022 and the *IEEE Technical Community on Services Computing*, *Research Innovation Award* from the *IEEE CS Technical Community on Services Computing* in 2023. He won the IEEE Region 1 *Technological Innovation Award* (Academic) in 2023. He is a member of the SUNY Distinguished Academy. He is an AAAS Fellow, an AAIA Fellow, and an ACIS Founding Fellow. He is an Academician Member and fellow of the International Artificial Intelligence Industry Alliance. He is a member of Academia Europaea (Academician of the Academy of Europe).