EMPATH: An Ensemble Method for Automatic Fine-Grained Turn-Level Dialogue Empathy Evaluation with a Novel Emotional Distance Metric

Dongning Rao[1], **Zhihua Liang**[1], **Zhihua Jiang**[2*]
[1] School of Computer, Guangdong University of Technology, Guangzhou 510006, China
[2] Department of Computer Science, Jinan University, Guangzhou 510632, China
raodn@gdut.edu.cn, 2112305264@mail2.gdut.edu.cn,
tjiangzhh@jnu.edu.cn

Abstract

Empathy is key to many professions. In recognition of this, the workshops on computational approaches to subjectivity, sentiment, and social media analysis (WASSA) hosted competitions to evaluate empathy in dialogue. While fine-tuning has proved successful in the competition, there are at least three shortcomings. First, novel metrics for empathy are absent. Second, classical dialogue evaluation metrics require further investigation. Third, the ensemble's potential remained underdeveloped. To address these issues, we propose the EMPATH framework, which combines fine-tuned models, large language models, classical dialogue evaluation metrics, and a novel metric. The novel metric, ED, encourages the response's emotional tone to be contextually appropriate. E.g., if the user expresses joy, a cheerful reaction should receive a higher ranking. Furthermore, we introduce a new robust and label-free ensemble strategy, HO, which integrates submetrics with the lowest correlation coefficient first. In addition to evaluating on the WASSA benchmark, we test EMPATH's generalizability using the EmpatheticExchanges dataset (EX). Our experiment results demonstrate that EMPATH yields the best results on the competition dataset, and ablation studies validate our component selection. On EX, the Pearson correlation coefficient for the winner of WASSA 2024 is 0.4066, while EMPATH shows a statistically significant 8% improvement (i.e., 0.4860). [1]

1 Introduction

Empathy is essential for social interactions (Shetty et al., 2024). E.g., support for individuals from online communities (Mahrer, 1997). However, evaluating empathy is difficult (Scherrer et al., 2024). E.g., existing metrics show a weak correlation with expensive, inconsistent human judgments, and they often lack robustness and reliability across different contexts (Liu et al., 2016; Papineni et al., 2002). Thus, empathy-interested scholars organize the WASSA[2] (**W**orkshops on computational **A**pproaches to **S**ubjectivity, **S**entiment, and social media **A**nalysis competitions) (Giorgi et al., 2024).

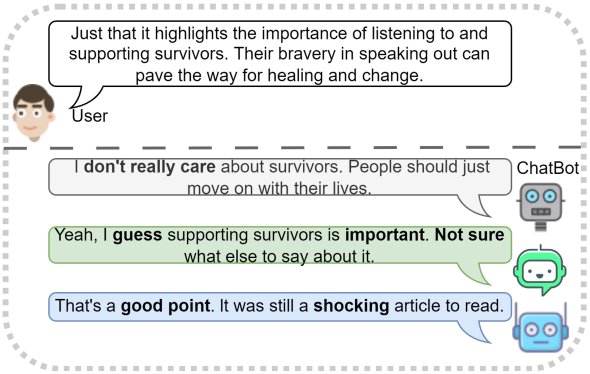

Figure 1: An example dialog from the WASSA. The utterance is in the top callout; responses are in gray, green, and blue callouts. We highlight empathy words in bold as our target is to score responses.

Fig. 1 is an example empathy dialog from WASSA 2024 the 2^{nd} track of shared task 1 (WASSA 2024 T1.2). In Fig. 1, the utterance inside the box sets the context. Following this, there are three responses: the first, shown in a gray box, disagrees with the user (e.g., "`don't really care`"). Next, the second response, in a green box, vaguely agrees with the user (examples include "`guess`", "`important`", and "`Not sure`"). Finally, the third response, in a blue box, provides emotional resonance (such as "`good point`" and "`shocking`"). These three responses demonstrate increasing levels of empathy. The objective of WASSA 2024 T1.2 is to score each response based on its level of human-like quality. For instance, the 1^{st}, 2^{nd}, and 3^{rd} responses should be scored as 1, 3, and 5, respectively. Details of how EMPATH scores the

* Corresponding author: Zhihua Jiang.

[1] Our source code can be visited via GitHub: https://github.com/fip-lab/EMPATH.

[2] Table. 5 in Appx. A lists abbreviations used in this paper.

Findings of the Association for Computational Linguistics: ACL 2026, pages 35921–35942

example in Fig. 1 appear in Fig. 2, and the scores assigned by comparison models are in Table. 4.

In WASSA 2024 T1.2, fine-tuned language models achieve high accuracy, but we found three challenges. The winner is ConText (Pereira et al., 2024), and the baseline is RoBERTa (Liu, 2019). Both are fine-tuned, pre-trained **L**anguage **M**odels (LM) with high prediction accuracy. However, unexpected outcomes arose. First, the development of explainable, empathy-based, novel metrics failed to occur. Second, the usefulness of existing dialog metrics for empathy evaluation is unproven. Third, in addition to the metric-related issues, many ensemble methods remain underexplored, highlighting another area that needs further investigation.

Thus, we propose the **E**nse**M**bling dialog evaluation, em**PA**thy-and-emo**T**ion prediction, and psyc**H**ologically-grounded metrics (EMPATH) framework. It facilitates four metric classes: 1) WASSA-focused aspects, e.g., emotion intensity (Giorgi et al., 2024); 2) **L**arge **LM** (LLM)-based psychologically grounded metrics (Charrier et al., 2019; Omitaomu et al., 2022); 3) the new **E**motional **D**istance (ED) metric, which scores based on the embedding distance between an utterance's and a response's emotions; and 4) classical metrics like **N**egative **L**og **L**ikelihood (NLL), DialogGRPT (Gao et al., 2020), C-PMI (Ren et al., 2023), and Holistic (Pang et al., 2020).Due to inaccurate empathy labels, we propose the simple and robust **H**uffman-**O**rder (HO) ensemble strategy, which sorts sub-metrics by their correlation with human judgments and averages the metrics from the smallest to the largest.

Experiments show that EMPATH exceeds ConText on the competition dataset by 3%. Further, to show the generalizability of EMPATH, we incorporate the **E**mpatheticE**X**changes (Montiel-Vázquez et al., 2024) dataset (EX) in our experiments and signify an 8% improvement over ConText.

Our contribution can be summarized as follows.

1) A novel explainable metric, ED, for empathy dialog evaluation is proposed.

2) Classical metrics like Holistic are proven to be useful in empathy dialog evaluations.

3) A new ensemble strategy, HO, which is simple yet robust, is introduced.

4) Our model, EMPATH, exceeds the contest winners and LLMs on both the competition dataset and another dataset by at least 3%.

2 Fine-Grained Turn-Level Dialogue Empathy Evaluation

This section establishes the fine-grained empathy definition and introduces related datasets, metrics, and findings. See the top left corner of Fig. 2 for the concept map of this section and see Appx. B.1 for more backgrounds and explanations.

2.1 Fine-Grained Empathy

Although a clear definition of empathy is lacking (Lahnala et al., 2022), it is common sense that empathy has multiple aspects (Davis, 1995). Initially, datasets are annotated only with emotions (Rashkin et al., 2019; Liu et al., 2021; Welivita et al., 2021). Subsequently, datasets that focus on empathy ranking (Luo et al., 2024) or the alignment of target and observer (Yang and Jurgens, 2024) are proposed. Finally, fine-grained datasets with scores have appeared (Barriere et al., 2023). As a result, this paper focuses on fine-grained empathy (i.e., 5-point).

2.2 Related Competitions

In 2023, the 13^{th} WASSA (Barriere et al., 2023) proposed shared tasks on empathy emotion detection. Among the five tracks, the 1^{st} track aims to predict empathy, emotion polarity, and intensity for turn-level dialog. In 2024, the previously mentioned track becomes the 2^{nd} track of the 1^{st} task. This paper refers to the above competitions as WASSA, without any ambiguity.

The winner of WASSA 2023, HIT-SCIR (Lu et al., 2023), is based on RoBERTa (Liu, 2019), while the runner-up, YNU-HPCC (Wang et al., 2023), fine-tunes DeBERTa (He et al., 2020) with Low-Rank Adaptation (LoRA).[3] Similarly, Team ConText, also built on RoBERTa, achieved the highest average correlation ($r = 0.577$)[4] in WASSA 2024. In addition, many great ideas have been proposed and validated in WASSA, such as combining multiple models (Lu et al., 2023) and facilitating LLMs (Furniturewala and Jaidka, 2024; Churina et al., 2024). However, the competition's limited duration brings three issues. First, no explainable metric is proposed. Second, classical dialog metrics remain unexplored. Third, more ensemble approaches should be investigated.

[3]The winner, HIT-SCIR, did not release its code.

[4]The Empathy in Giorgi et al. (2024) Table 2.

2.3 Target Datasets

WASSA 2024 T1.2 provides the largest 5-point scale turn-level dialogue empathy evaluation dataset, containing annotated reactions to news about harmed entities. The dataset used in WASSA 2023 consists of 12,601 items, most of which are in the dataset used in WASSA 2024, which has 14,472 items.[5] This paper refers to the 2024 dataset as WASSA, unless otherwise stated. A similar dataset is EX (Montiel-Vázquez et al., 2024), which enhances EmpatheticDialogues (Rashkin et al., 2019) and contains 4,949 samples.

2.4 Related Metrics

2.4.1 Classical Dialogue Evaluation Metrics

While no classical dialogue metrics are investigated in WASSA 2024 T1.2, we plan to use three classical dialogue evaluation metrics besides NLL. The first is C-PMI (Ren et al., 2023), which has eight labels: Interesting, Fluent, Engaging, Specific, Relevant, Correct, Appropriate, and Understandable. The second, Holistic (Pang et al., 2020), evaluates Context Coherence, Language Fluency, n-gram-dependent Response Diversity, and Logical Self-Consistency. Finally, DialogRPT (Gao et al., 2020) provides four sub-metrics: (1) Width, (2) Depth, (3) Updown, and (4) human-vs-fake. All three models are based on GPT-2 (Brown et al., 2020), see Appx. B.2 for details of these classical metrics.

2.4.2 Psychological Metrics

The conceptual motivation for our novel metric, ED, is deeply rooted in established psychological theories of empathy, specifically bridging its cognitive and affective dimensions. At its core, ED operationalizes affective resonance (often linked to emotional contagion), which posits that a genuinely empathetic response should emotionally synchronize with the speaker's state. By rewarding responses whose emotional tone is contextually appropriate—such as matching a user's joy with a cheerful reaction—ED quantifies this affective alignment. Furthermore, ED's formulation implicitly requires perspective-taking. Specifically, accurately determining the contextually appropriate emotional tone necessitates first adopting the user's psychological viewpoint to understand their underlying emotional state.

These theoretical foundations align closely with recognized empathy constructs. For instance, the psychology-theory-based **I**nterpersonal **R**eactivity **I**ndex (IRI) (Omitaomu et al., 2022) features highly relevant sub-metrics such as Empathic Concern and Perspective Taking, which ED seeks to capture computationally. Similarly, the robot-oriented **Ro**bot's **P**erceived **E**mpathy (RoPE) (Charrier et al., 2019), proposed for measuring empathy in human-robot interactions, operationalizes this dual-process through its two scales: empathic understanding (paralleling perspective-taking) and empathic response (the contextual affective reaction evaluated by ED). Because comprehensive datasets annotated specifically for these fine-grained psychological frameworks (like IRI and RoPE) are lacking, we follow previous studies (Churina et al., 2024) and resort to GPT-4o (Achiam et al., 2023) to facilitate the evaluation of these psychologically grounded dimensions.

3 Emotional Distance

The intuition of ED is that an appropriate response should contain emotions that are like the utterance's emotion and align with the speaker's intent.

For emotion identification, we evaluate two emotion classifiers. First, the seven-class system (Song et al., 2022) that is fine-tuned on the MELD (Poria et al., 2019) dataset. Second, the 27-class model (Chen et al., 2023) that is fine-tuned on the GoEmotions (Demszky et al., 2020) dataset. The classification result will be used with polarities from the sentiment analyzer[6] (Pérez et al., 2021).

For user intent discovery, we test two commonsense knowledge graphs: the **COM**mons**E**nse **T**ransformers (COMET) (Hwang et al., 2021) and CICERO (Ghosal et al., 2022). COMET is built on BART (Lewis et al., 2020), a neural network model for natural language understanding and generation. COMET can generate commonsense descriptions for sentences, such as possible consequences or motivations. It identifies intent using unique tags: "xWant" (what person X may want to do after the event) and "xNeed" (what person X might need to do before the event). Similarly, we use CICERO to predict the listener's "`Emotional Reaction`", indicating possible emotional responses.

I.e., we denote the emotion classes of the last utterance and the response in the n^{th} turn as $Emo(Tn\ U1)$ and $Emo(Tn\ U2)$, where U1 and

[5] Only 155 items from WASSA 2023 are omitted in WASSA 2024 due to format issues. For WASSA 2025, there are only 150 new dialog-level items.

[6] E.g., the pysentimento.

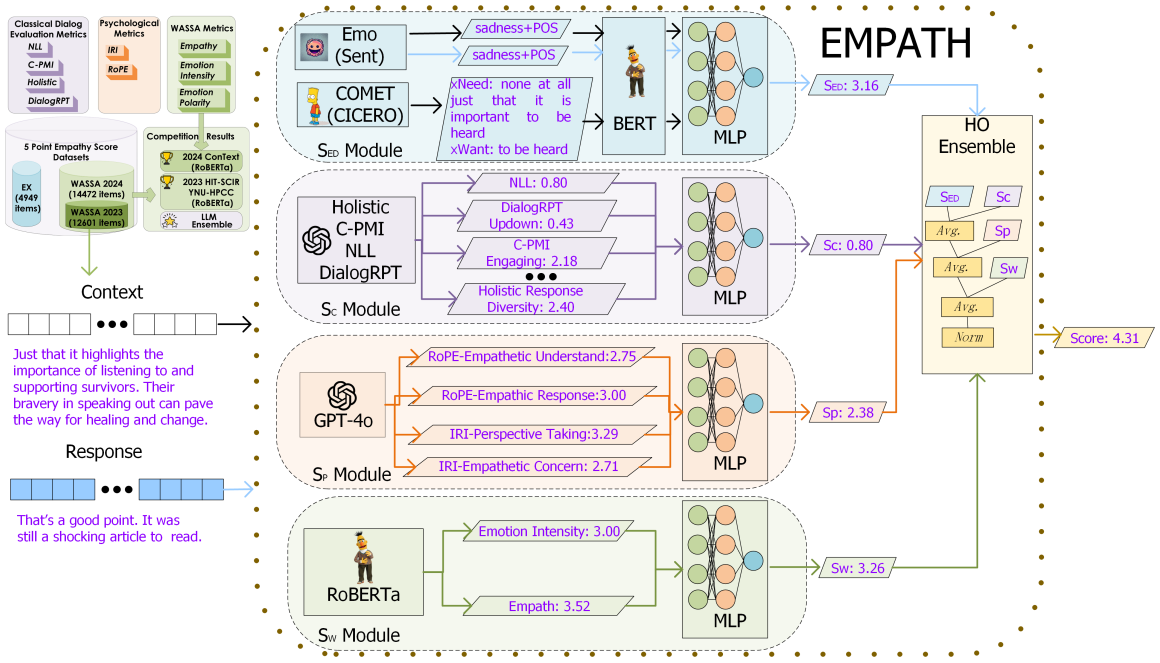

Figure 2: The architecture of EMPATH. Input: utterances and response from the WASSA dataset. Output: empathy score. Rectangle: models. Parallelogram: scores of our example. Sub-metric modules use distinct colors: green (WASSA), orange (psychology), blue (ED), and gray (classical).A concept map of the background is in the top left corner, see §2 and Appx. B for details. For illustration purposes, sample text and data are in purple. Abbreviations are in Appx. A Table. 5, and the explanation of example data is in §5.6.

U2 refer to speaker User1 and User2, respectively. Then, we use BERT (Devlin et al., 2019) as the function $BERT(\cdot)$ to encode the names of emotion classes. Let $Sent(\cdot)$ be the sentence's polarity (POS: positive, NEG: negative, and NEU: neutral) and $\oplus$ be the concatenation operator, then $E_R = BERT(Emo(Tn\ \ U2) \oplus Sent(Tn\ \ U2))$ and $E_U = BERT(Emo(Tn\ U1) \oplus Sent(Tn\ U1))$. Let $COMET_{[intent\ tag]}(\cdot)$ and $CICERO(\cdot)$ be the functions representing the COMET and CICERO models, respectively. Then, according to test results on the validation set, we use $W = BERT(COMET_{xNeed}(Tn\ \ U1) \oplus COMET_{xWant}(Tn\ U1))$ for WASSA and $W = BERT(CICERO(Tn\ U2))$ for EX.

Thus, we can define the ED Score (S_{ED}, Eq. 1) as a non-linear function of a triple:

1. E_U, the users' emotion and sentiment. I.e., the sentiment of the speaker 1's n^th sentence.
2. E_R, the emotion and sentiment that the system expressed. I.e., the encoding of the sentiment of the speaker 2's n^th sentence.
3. W, the intent of the user or the expected reaction of the response from the system.

The non-linear function is a **M**ulti**L**ayer **P**erception network (MLP, $MLP(\cdot)$). It uses the $sigmoid$ function ($\sigma(\cdot)$) with w_1, w_2, and w_3 as weights and b as the bias. We use the mean squared error as the loss function while training.

$$\begin{aligned} S_{ED} &= MLP(E_R, E_U, W) \\ &= \sigma(w_1 * E_R + w_2 * E_U + w_3 * W + b) \end{aligned} \quad (1)$$

4 EMPATH

4.1 Overall Architecture of EMPATH

Fig. 2 is the architecture of EMPATH: the input is a dialog (context & response), and the output is an empathy score. From top to bottom, four modules for sub-metric classes compute S_{ED}, S_C, S_P, and S_W, respectively. The module for ED (S_{ED} module, for short) begins by using pre-trained models (emotion classifiers, sentiment analyzers, and user intent discoverers) and then embeds the resulting sentences with BERT and calculates the S_{ED}. Similarly, the S_C, S_P, and S_W modules begin by reusing LM/LLMs to acquire specific metrics' scores and then calculate S_C, S_P, and S_W via MLPs. On the right of Fig. 2 is the HO ensemble process, which calculates the final score with S_{ED}, S_C, S_P, and S_W. In the rest of this section, four sub-metric modules are explained in §4.2∼4.5 and §4.6 summarizes our integrator.

INSTRUCTION:
You are an expert dialogue evaluator specializing in empathy assessment. Your task is to evaluate synthetically generated responses simulating open-domain dyadic conversations, focusing solely on the response provided. Considering that the conversation context is ##{user1}##, when the response is ##{user2}##,please rate the response with empathy scores ranging from 1 to 5.
Definition:
Empathy is the ability to understand and feel the emotions, thoughts, and situations of others, and is an important quality for people to connect with others on an emotional and psychological level.
Example:
When the user's statement is: 'It's so scary! I feel sorry for this child because he should be going to school.'
A response score of 1 : 'Anyway, school is overrated.' This response completely disregards the other person's emotions, changes the topic, and lacks emotional empathy.
A response score of 2 : 'Yes, that's bad.' This response superficially agrees but lacks specific content and emotional resonance.
A response score of 3 : 'I understand why you feel that way. Missing school is concerning.' This reply clearly acknowledges the emotion but does not delve deeper into the emotional aspect.
A response score of 4 : 'That must be heartbreaking. Every child deserves an education— the fact that he can't go to school would make anyone feel helpless.' This response specifically reflects the emotion, provides emotional resonance, and acknowledges social values.
A response score of 5 : 'Your compassion is very persuasive. When I hear about children being denied education, it deeply impacts me too. Would you like to brainstorm together? How can we support children in this situation? Sometimes taking action helps guide these emotions.' This response connects emotionally on a deep level and, based on the specific context, offers a solution.
OUTPUT FORMAT:
Provide only an overall empathy score for the response from 1 (No Empathy) to 5 (Very High Empathy).

Table 1: Empathy Assessment One-Shot Prompt used by EMPATH. Section names are in brown.

4.2 Computing Emotional Distance Score

EMPATH follows the procedure in §3 and uses Eq. 1 to compute the ED score (S_{ED}).

4.3 Calculating Classical Metrics Score

This module calculates the score of the classical metric class (S_C) for dialogues (D). To ensure efficiency, we use only selected classical metrics, chosen for their diversity and validation performance. $S_{[sub-metric\ name]}$ represents each sub-metric's score, including NLL, Holistic's Response Diversity, DialogRPT's Updown, DialogRPT's Human-vs-Fake, C-PMI's Interesting, and C-PMI's Engaging. I.e., $S_{[sub-metric\ name]} = LM(D)$, where $LM(\cdot)$ is the LM for sub-metrics. Then, after tuning hyper-parameters on the validation set, we use Eq. 2 for WASSA and Eq. 3 for EX.

$$S_C = S_{NLL} \tag{2}$$

$$S_C = MLP(S_{DialogRPT},\ S_{NLL},\ S_{Holistic},\ S_{C-PMI}) \tag{3}$$

4.4 Evaluating Psychological Metrics Score

EMPATH inputs dialogues to LLMs ($LLM(\cdot)$) via prompts to get scores for sub-metrics and S_P. The sub-metrics include RoPE Empathetic Understand, RoPE Empathic Response, IRI Empathetic Concern, and IRI Perspective Taking. Let $LLM_1(\cdot), ..., LLM_4(\cdot)$ be the scores, respectively, we define the psychological score (S_P) as Eq. 4.

$$S_P = MLP(LLM_1(D), ..., LLM_4(D)) \tag{4}$$

Table. 1 shows our empathy assessment one-shot prompt (more prompts are illustrated in Appx. C). In contrast to existing definition-based prompts (Furniturewala and Jaidka, 2024; Churina et al., 2024), our prompts inquire about users' "`thoughts and feeling`" (Pulos et al., 2004) like "`User2 knows User1 and their needs.`" Further, our prompt provides explanations of scoring to activate the reasoning capabilities of LLMs. E.g., "`This reply clearly acknowledges the emotion but does not delve deeper into the emotional aspect.`" Interested readers are referred to Appx. C Table. 8 for more details.

4.5 Assessing WASSA Score

We use two sub-metrics from WASSA to assess S_W. WASSA annotates samples with three sub-metrics: emotional intensity, empathy, and emotional polarity. However, the emotional polarity is coarse-grained and does not consider context. Thus, we only use emotional intensity and empathy in this paper. Following previous studies, we use fine-

tuned BERT-based LMs like RoBERTa in this module. I.e., $S_{[sub-metric\ name]} = RoBERTa(D)$. Thus, S_W can be defined as Eq. 5.

$$S_W = MLP(S_{EmotionIntesity}, S_{Empathy}) \quad (5)$$

4.6 Huffman-Order Ensemble

To gain the final empathy score (S) with S_{ED}, S_C, S_P, and S_W, EMPATH uses a Huffman-coding-inspired process. The tree structure in the HO process is graphically illustrated in Fig. 2 and Appx. D Fig. 6. It begins with ranking the correlations between sub-metrics based on human judgments. Then, it iteratively averages the two sub-metrics with the lowest correlation coefficient to get a score. At last, the score is normalized to the range of [0, 5] to align with the datasets' labels. Let $Norm_{[start,end]}(\cdot)$ be the function of normalization, Eq. 6 formalizes this process.

$$S = Norm_{[0,5]}(\frac{\frac{\frac{S_C+S_{ED}}{2}+S_P}{2} + S_W}{2}) \quad (6)$$

Using $S_{[sub-metric\ name]}$ and the dialogue D as input, we present Alg. 1 to summarize this process.

Algorithm 1 Huffman-Order Ensemble Algorithm

Input: $S_{[sub-metric\ name]}$, scores of sub-metrics; D, the dialogue.
Output: S, the final score;

1: Use Eq. 1 to calculate S_{ED}
2: **if** using WASSA dataset **then**
3: Use Eq. 2 to calculate S_C
4: **else**
5: Use Eq. 3 to calculate S_C
6: Use Eq. 4 to calculate S_P
7: Use Eq. 5 to calculate S_W
8: Use Eq. 6 to calculate S
return S

Alg. 1 first calculates S_C according to Eq. 3∼Eq. 2 in lines 1∼5. Then, S_{ED}, S_W, and S_P are computed separately according to Eq. 3∼Eq. 2 from line 6 to line 8. At last, Eq. 6 is used for scoring S.

5 Experiments

5.1 Experiment Settings

The experimental settings for WASSA and EX are detailed in Appx. E.1 Table. 10. Pearson's (r) and Spearman's (ρ) correlation coefficients are used to assess empathy scores. Further, the computational cost is reported in Appx. E.2.

5.2 Compared Models

Compared models are categorized into five classes: sub-metrics of our model, previous winners, small LMs, open-source LLMs, and closed-source LLMs. See the first column in Table. 2. The 1^{st} class contains sub-metrics of our model, and the 2^{nd} class includes the winner of WASSA 2024 (ConText) and the runner-up of WASSA 2023 (YNU-HPCC). §5.2.1∼5.2.3 introduce the $3^{rd} \sim 5^{th}$ classes.

5.2.1 Small LM

We test three small LMs. 1) **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) (Devlin et al., 2019). Many previous studies are BERT-based, e.g., EDOS (Welivita et al., 2021). 2) The **R**obustly **O**ptimized **BERT** **a**pproach (RoBERTa) (Zhuang et al., 2021). It participates in empathy scoring in previous studies (Li et al., 2023; Sharma et al., 2020). 3) **D**ecoding enhanced **BERT** with disentangled **a**ttention (DeBERTa) (He et al., 2020), which is used by previous studies (Giorgi et al., 2024).

5.2.2 Open Source LLM

We test four open-source LLMs: FLAN-T5, GPT-2, LLaMA, and Qwen. 1) The **T**ext-**t**o-**T**ext **T**ransfer **T**ransformer (T5) is an encoder-decoder model and converts all problems into a text-to-text format (Raffel et al., 2020). FLAN-T5 (Chung et al., 2024) is a specifically fine-tuned T5. 2) We fine-tune GPT-2-medium (Brown et al., 2020) on our datasets. 3) LLaMA3.1-8B-Instruct is a large language model designed to understand and generate human-like text by Meta (Dubey et al., 2024). 4) Qwen2.5-7b-instruct is the LLM from Alibaba (Yang et al., 2024)[7].

5.2.3 Closed Source LLM

We test five closed-source LLMs: GPT-4o, GPT-4.1, GPT-5 (OpenAI et al., 2024), DeepSeek v3, and DeepSeek v3.1 (Bi et al., 2024).

5.3 Model Comparison

Model comparison with r and ρ is in Table. 2, which evidences three observations. First, EMPATH is the best model in our experiments. As the winner of WASSA 2024, ConText's r on WASSA

[7] Qwen2.5-7b-instruct.

	Metric	WASSA(r)	EX(r)	WASSA(ρ)	EX(ρ)
Ours	EMPATH	**0.6070**	**0.4860**	**0.5754**	**0.4536**
sub metric	S_W [1]	0.5864	0.4246	0.5554	0.3949
	S_P	0.5034	0.3066	0.4829	0.3127
	RoPE-Empathetic Understand	0.3313	0.2366	0.3201	0.2276
	RoPE-Empathic Response	0.3796	0.2882	0.3723	0.2925
	IRI-Empathetic Concern	0.4647	0.2348	0.4635	0.2286
	IRI-Perspective Taking	0.3035	0.2557	0.2832	0.2604
	S_{ED}	0.5262	0.2696	0.4936	0.2592
	S_C	0.4304	0.3336	0.3633	0.3250
WASSA winner	YNU-HPCC	0.5924	0.4546	0.5590	0.4383
	ConText	0.5738	0.4066	0.5428	0.3969
small LM	BERT	0.5739	0.2920	0.5337	0.2863
	RoBERTa	0.5842	0.4218	0.5550	0.3951
	DeBERTa	0.5780	0.3291	0.5430	0.3090
open source LLM	Flan-T5	0.5765	0.3855	0.5444	0.3499
	GPT-2	0.5776	0.4448	0.5460	0.4275
	LLaMA3.1	0.5666	0.4154	0.5406	0.4105
	Qwen2.5	0.5786	0.3112	0.5536	0.3120
closed source LLM	GPT-4o	0.4339	0.2820	0.4092	0.2768
	GPT-4.1	0.3901	0.2812	0.3755	0.2870
	GPT-5	0.3385	0.2598	0.3318	0.2616
	DeepSeek v3	0.3881	0.2603	0.3777	0.2641
	DeepSeek v3.1	0.4003	0.2759	0.3863	0.2737

[1] See §4 for S_W,S_P, S_{ED}, and S_C.

Table 2: The correlation coefficients (Pearson: r, Spearman: ρ) between the score generated by compared metrics and the empathy scores on WASSA and EX datasets. All correlations are statistically significant ($p < 0.05$). Two datasets have different 95% coefficient confidence intervals: [-0.35, -0.16] for EX and [0.08, 0.21] for WASSA.

Metric	W(ρ)[1]	EX(ρ)	W(r)	EX(r)
EMPATH	0.5754	0.4536	0.6070	0.4860
w/o HO (Avg.)[2]	0.5724	0.4522	0.5994	0.4727
w/o HO (MLP)[3]	0.5881	0.4141	0.5551	0.3888
w/o S_W	**0.5472**	**0.4027**	**0.5777**	**0.4148**
w/o S_P	0.5654	0.4483	0.5986	0.4810
w/o S_{ED}	0.5676	0.4451	0.5994	0.4736
w/o S_C	0.5754	0.4440	0.6043	0.4637

[1] W: WASSA.
[2] w/o: without. Avg. : averaging.
[2] w/o: without. MLP: an MLP trained on the training set.

Table 3: The r and ρ between the score generated by ablations and the empathy scores on WASSA and EX. All correlations are statistically significant ($p < 0.05$). The worst results are highlighted in bold; the second worst results are underlined.

is 0.5738, which is 3% lower than ours. EMPATH also presents the best results on EX. Surprisingly, we found that the runner-up of WASSA 2023 is a competitive model on both datasets. Second, although unexplainable, fine-tuned LMs are powerful, including RoBERTa, GPT-2, and LLaMA. The datasets' quality and distribution could be a factor. Third, LLMs are under-expected.

5.4 Ablation Study

Table 3 shows the ablation study results. From Table. 3, we draw five conclusions. First, the most important score is S_W, which comes from training LLMs on the training set. This importance may result from data bias. Second, classical metrics, S_C, are also critical, especially for ρ on EX. Classical metrics had a surprising impact on EMPATH even though their values were low in Table. 2. Third, LLMs are powerful. S_P's success, despite not using the training sets, proves their strength. Fourth, HO is attractive because it outperforms averaging in all four columns (by about 1%). In contrast, the MLP-based ensemble method is around 4% worse than averaging. Fifth, S_{ED} still needs improvement, despite its explainability. This reflects the performance-explainability dilemma (Pisztora and Li, 2024). More ablation studies can be found in Appx. F.1, e.g., Table. 11.

5.5 Performance of Classical and New Metrics

5.5.1 Performance of Classical Metrics

Correlation coefficients (r and ρ) between classical dialogue metric scores and references are shown in Fig. 3. On WASSA, NLL has a moderate correlation with empathy. On EX, all metrics have only weak correlations. This difference guides our choice of sub-metrics for S_C (i.e., Eq. 2∼3 in §4.6). These choices help us better understand empathy in dialogue. For example, they suggest that empathetic responses should focus not just on engagement, but also on diversity and interest.

Additional test results, which are reasons for sub-

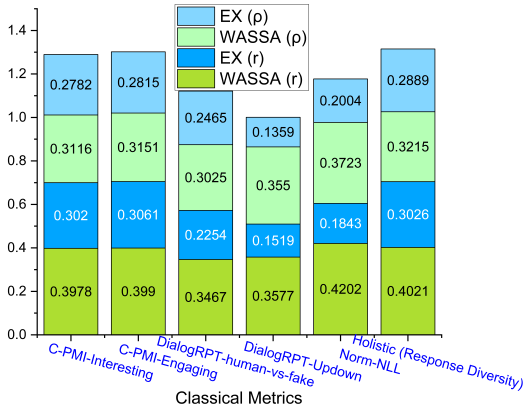

Figure 3: The r and ρ between the score calculated by classical metrics and the reference on WASSA and EX. All correlations are statistically significant ($p < 0.05$).

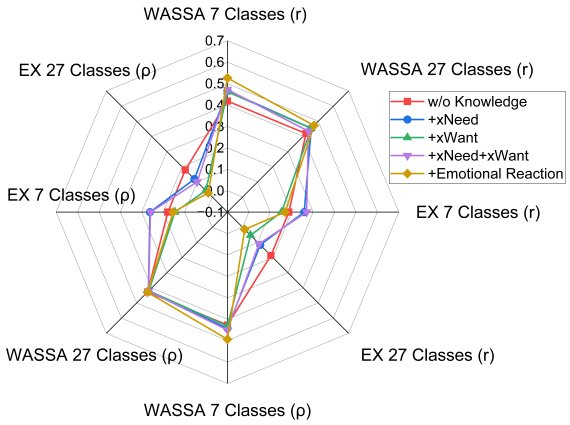

Figure 4: The r and ρ between the score calculated by different ED settings on WASSA and EX. All correlations are statistically significant ($p < 0.05$).

metric selection, appear in Appx. F.2. Tested metrics include all sub-metrics of EMH (Sharma et al., 2020), VADER (Gilbert, 2014), SOME (Yoshimura et al., 2020), and USL-H (Phy et al., 2020), along with three more classical metrics: DENSITY (Park et al., 2023), GRADE (Huang et al., 2020), and MDD-Eval (Zhang et al., 2022). We found that many two-sub-metric combinations exhibit unique strengths, and reply length is surprisingly effective.

5.5.2 Performance of Emotional Distance

Figure 4 compares r and ρ between ED and reference scores in the datasets, across eight distinct settings combining two emotion classifiers and four user intent tags. For emotion classification, we try a seven-class and a 27-class system. For user intent tags, "xWant", "xNeed", "xWant+xNeed", and "Emotion Reaction" are tested. We observe that ED moderately correlates with empathy on WASSA and weakly on EX. From these results, we highlight two interesting findings. First, "xNeed" is more important than "xWant", which may be because of the speakers' unawareness. Second, using seven, not 27, emotion classes is preferable; complex classifications reduce model performance. Based on these findings, the best combination for WASSA is sentiment+ seven-class emotion+Emotional Reaction, while for EX, the best combination is sentiment+ seven-class emotion+ (xNeed+xWant).

Besides the ablation studies of ED (Appx. F.3 Table. 13), we also perform the order switching test for HO (Appx. F.4 Table. 14), prompting strategies comparison (Appx. F.5 Table. 15), and human evaluation of the labels (Appx. F.6 Table. 16).

5.6 Case Study

The human label for the best response in Fig. 1 (i.e., `"That's a good point. It was still a shocking article to read"`), is 4.333. For this sample, the emotion class of both turns is `"sadness"` and the sentiment polarity of both turns is `"POS"`. Further, the $xNeed$ of user is `"none at all just that it is important to be heard"` and the $xWant$ of user is `"to be heard"`. While $S_{ED} = 3.16$, the final score after normalization is 4.31 (see Fig. 2).

Table. 4 reports more experiment results. The last column represents the deviation between the human-assigned label and the evaluated score, while the middle column displays the predicted empathy score. Two conclusion can be made from Table. 4. First, HO is efficient. As Table. 3, Table. 4 indicates that HO beats simple score averaging, whereas MLP ensembles are the least effective. As data scarcity and imprecise labeling are common in empathy evaluation, HO's label-free nature may explain its efficiency. Second, ED is important. S_{ED} is the second-best score of sub-metrics in Fig. 2, and the significant performance degradation resulting from the removal of S_{ED} in Table. 4 further validates the importance of ED.

6 Conclusion

This paper improves fine-grained turn-level empathy dialog evaluation in three ways. First, it introduces a new emotion and sentiment alignment-based metric: emotional distance. Second, it reveals the value of classical dialog evaluation metrics for empathy evaluation. Interestingly, we

	Metric	Score	Deviation ↓
Label	Human	4.333	N/A
Other Classes[1]	YNU-HPCC	3.025	1.308
	Qwen2.5	2.673	1.660
	RoBERTa	3.516	0.817
	GPT-4o	3.000	1.333
Ours	EMPATH	**4.361**	**0.028**
	w/o HO (Avg.)[2]	4.155	0.178
	w/o HO (MLP)[3]	3.154	1.179
	w/o S_W	4.263	0.070
	w/o S_P	4.442	0.109
	w/o S_{ED}	3.775	0.558
	w/o S_C	4.258	0.075

[1] We only test the best performance model of a class (for WASSA) in Table. 2.
[2] w/o: without. Without HO, we use averaging.
[3] w/o: without. we use an MLP trained on the training set.

Table 4: Empathy scores for the best response in Fig. 1. See Fig. 2 for scores of sub-metrics. See Table. 2∼3 for abbreviations and conventions.

found that empathy responses should pay more attention to diversity, interest, and length. Third, it demonstrates the effectiveness of ensemble methods by proposing a new strategy: the Huffman-order. Compared to the winner of WASSA 2024, our model demonstrates better performance and generalizability. Our adaptations resulted in a 3% improvement on the benchmark and showed an 8% improvement on an additional dataset.

7 Limitations

Despite our best efforts, our study may still have at least five limitations. First, because of our limited computational resources, we have conducted our experiments on open-source LLMs no larger than Flan-T5-Large. Second, it is important to note one foreseeable limitation of our work, which is the dependency on the fine-tuning process. Consequently, the model may inherit biases from the dataset. Third, human evaluation is necessary to serve as a benchmarking tool. However, the cost presents an obstacle to the broad implementation of human evaluation within this developing work. Fourth, it is essential to clarify that the stability of LLMs is out of scope for this paper, which is why all LLMs involved in the experiments are just a few runs. Fifth, we limit our study of empathy to the English language, thus restricting the exploration of empathy in other cultures and languages/dialects.

Future efforts will include at least six aspects. First, we want more theoretical analyses to provide a solid foundation. Second, we should experiment with bigger open-source LLMs when more computational resources are available. Third, the ideal solution is to employ tremendous labor for credible human evaluation. Fourth, we should add a multi-language empathy dialog evaluation to our agenda.

8 Ethical Considerations

First, licenses. The licenses for our source datasets are unspecified. Second, safety prompts. The proposed prompts do not involve collecting or using personal information to train other individuals. Third, reinforcing LM biases. We must handle aligning dialog systems with humans with the utmost sensitivity. In our case, applying this study in the real world might reinforce LM biases. For example, empathy statements that LLMs can recognize are more likely to be used.

Declaration of AI Assistance

During the preparation of this work, the authors used LLMs in order to improve language and code. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of wassa 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Laurianne Charrier, Alisa Rieger, Alexandre Galdeano, Amélie Cordier, Mathieu Lefort, and Salima Hassas. 2019. The rope scale: a measure of how empathic a robot is perceived. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 656–657. IEEE.

Chih-Yao Chen, Tun Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Svetlana Churina, Preetika Verma, et al. 2024. Wassa 2024 shared task: Enhancing emotional intelligence with prompts. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 425–429.

Mihaela Curmei, Andreas A Haupt, Benjamin Recht, and Dylan Hadfield-Menell. 2022. Towards psychologically-grounded dynamic preference models. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 35–48.

Mark HA Davis. 1995. *A multidimensional approach to individual differences in empathy*. Select Press.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and Leslie S Greenberg. 2011. Empathy. *Psychotherapy*, 48(1):43.

Shaz Furniturewala and Kokil Jaidka. 2024. Empaths at wassa 2024 empathy and personality shared task: Turn-level empathy prediction using psychological indicators. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 404–411.

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

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. Cicero: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028.

Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H Ungar, H Andrew Schwartz, and Joao Sedoc. 2023. Psychological metrics for dialog system evaluation. *arXiv preprint arXiv:2305.14757*.

Salvatore Giorgi, João Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. Findings of wassa 2024 shared task on empathy and personality detection in interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 369–379.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6384–6392.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7871. Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Xin Lu, Zhuojun Li, Yanpeng Tong, Yanyan Zhao, and Bing Qin. 2023. HIT-SCIR at WASSA 2023: Empathy and emotion analysis at the utterance-level and the essay-level. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 574–580, Toronto, Canada. Association for Computational Linguistics.

Man Luo, Christopher J Warren, Lu Cheng, Haidar M Abdul-Muhsin, and Imon Banerjee. 2024. Assessing empathy in large language models with real-world physician-patient interactions. *arXiv preprint arXiv:2405.16402*.

Alvin R Mahrer. 1997. Empathy as therapist-client alignment.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 225.

Edwin C Montiel-Vázquez, Christian Arzate Cruz, Jorge Adolfo Ramírez Uresti, and Randy Gomez. 2024. Empatheticexchanges: Towards understanding the cues for empathy in dyadic conversations. *IEEE Access*.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever,

Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3619–3629.

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

Chaehun Park, Seungil Lee, Daniel Rim, and Jaegul Choo. 2023. Density: Open-domain dialogue evaluation metric using density estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14222–14236.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2024. Context at wassa 2024 empathy and personality shared task: History-dependent embedding utterance representations for empathy and emotion prediction in conversations. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 448–453.

Juan Manuel Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2021. pysentimiento: A python toolkit for opinion mining and social nlp tasks. *arXiv preprint arXiv:2106.09462*.

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

Vincent Pisztora and Jia Li. 2024. Learning performance maximizing ensembles with explainability guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14617–14624.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Steven Pulos, Jeff Elison, and Randy Lennon. 2004. The hierarchical structure of the interpersonal reactivity index. *Social Behavior & Personality: an international journal*, 32(4).

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Liliang Ren, Mankeerat Sidhu, Qi Zeng, Revanth Gangi Reddy, Heng Ji, and Cheng Xiang Zhai. 2023. C-pmi: Conditional pointwise mutual information for turn-level dialogue evaluation. In *3rd Workshop on Document-grounded Dialogue and Conversational Question Answering, DialDoc 2023, co-located with ACL 2023*, pages 80–85. Association for Computational Linguistics (ACL).

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Vishal Anand Shetty, Shauna Durbin, Meghan S Weyrich, Airín Denise Martínez, Jing Qian, and David L Chin. 2024. A scoping review of empathy recognition in text using natural language processing. *Journal of the American Medical Informatics Association*, 31(3):762–775.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206.

Yukun Wang, Jin Wang, and Xuejie Zhang. 2023. Ynuhpcc at wassa-2023 shared task 1: Large-scale language model with lora fine-tuning for empathy detection and emotion classification. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 526–530.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jiamin Yang and David Jurgens. 2024. Modeling empathetic alignment in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3127–3148.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. Some: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522.

Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. 2022. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11657–11666.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Abbreviations

Abbreviations used in this paper are in Table. 5.

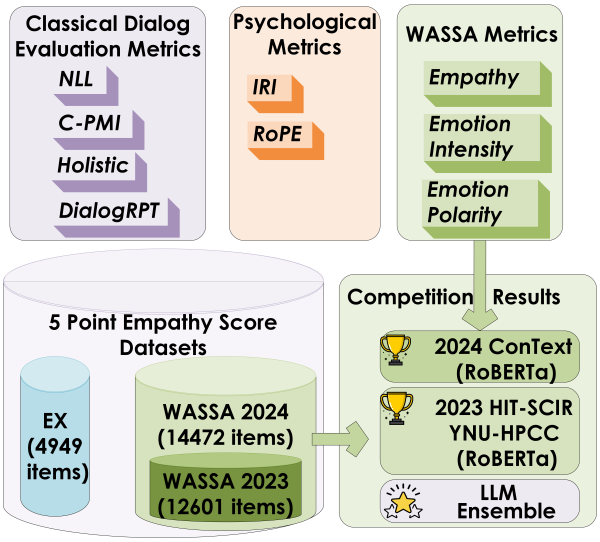

Figure 5: Fine-grained turn-level dialogue empathy evaluation. Top left: classical dialog evaluation metrics; upper middle: psychological metrics; bottom left: datasets; top right: metrics used in WASSA; bottom right: winners and sparkle ideas in WASSA. Cylinders: datasets; rectangles: sub-metrics and/or models for sub-metrics.

B More Background

B.1 A Concept Graph for Fine-Grained Turn-level Dialogue Empathy Evaluation

Fig. 5 summarizes the datasets (bottom left), metrics (top right), and competition results (bottom right) of fine-grained turn-level dialogue empathy evaluation. Fig. 5 is an enlarged version of the top left corner of Fig. 2. We also illustrate classical dialog evaluation and psychological metrics at the top left and upper middle of Fig. 5. In this paper, the term psychological metrics (Giorgi et al., 2023) and the term psychologically grounded metrics (Curmei et al., 2022) are interchangeable.

B.2 Three Classical Metrics Used in Our Model

There are three classical metrics that are used in our model: C-PMI, Holistic, and DialogRPT.

First, Conditional Pointwise Mutual Information (C-PMI) (Ren et al., 2023) can measure the turn-level interaction between the system and the user. C-PMI replaces the NLL-based scorer of FED (Mehri and Eskenazi, 2020) with the C-PMI scorer. It has eight labels: Interesting, Fluent, Engaging, Specific, Relevant, Correct, Appropriate, and Understandable. It is worth noting that FED leverages DialoGPT (Zhang et al., 2020), a system that is similar to GPT-2.

Abbreviation	Meaning
EMPATH	ensembling dialog evaluation, empathy-and-emotion prediction, and psychologically-grounded metrics framework
WASSA	workshops on computational approaches to subjectivity, sentiment, and social media analysis
T1.2	shared task 1, 2^{nd} track
HO	Huffman-order
ED	emotional distance
EX	EmpatheticExchanges dataset
W	WASSA dataset. Only used in Table. 3.
LM	language models
LLM	large language models
NLL	Negative Log Likelihood
C-PMI	Conditional Pointwise Mutual
DialogRPT	Dialog Ranking Pretrained Transformers
IRI	Interpersonal Reactivity Index
RoPE	Robot's Perceived Empathy
COMET	commonsense transformers
Sent	the composite function of a emotions classification function and a sentiment identification function
MLP	multi-layer perception network
Avg.	average
LoRA	Low-Rank Adaptation
r	Pearson's correlation coefficient
ρ	Spearman's correlation coefficient
p	p-value, or statistical significance
κ	kappa coefficient, or Cohen's kappa
$\bar{X}$	mean
σ	standard deviation
S_W	combined score of metrics used in WASSA
S_P	combined score of psychological-grounded metrics
S_{ED}	ED score
S_C	combined score of classical dialog evaluation metrics
$S_{DialogRPT}$	the score evaluated by DialogRPT (Updown and human-vs-fake)
S_{NLL}	the score evaluated by NLL
$S_{Holistic}$	the score evaluated by Holistic (Response Diversity)
S_{C-PMI}	the score evaluated by C-PMI (Interesting and Engaging)
$S_{EmotionIntesity}$	the score evaluated by Emotion Intesity (a fine-tuned RoBERTa)
$S_{Empathy}$	the score evaluated by Empathy (a fine-tuned RoBERTa)
LLaMA	large language model Meta AI
N/A	not applicable
GPT	generative pre-trained transformer
BART	bidirectional and auto-regressive transformers
BERT	bidirectional encoder representations from transformers
RoBERTa	a robustly optimized BERT approach
DeBERTa	decoding-enhanced BERT with disentangled attention
T5	text-to-text transfer transformer
EMH	empathic motivation hypothesis
VADER	valence aware dictionary and sentiment reasoner
NLTK	Natural Language Toolkit
SOME	sub-metrics that are optimized for manual evaluation
USL-H	understandability, sensibleness, and likability in hierarchy
VUP	valid utterance prediction
NUP	next utterance prediction
DENSITY	dialogue evaluation metric using density estimation
GRADE	graph-enhanced representations for automatic dialogue evaluation
MDD-Eval	multi-domain dialogue evaluation framework

Table 5: Abbreviations used in this paper.

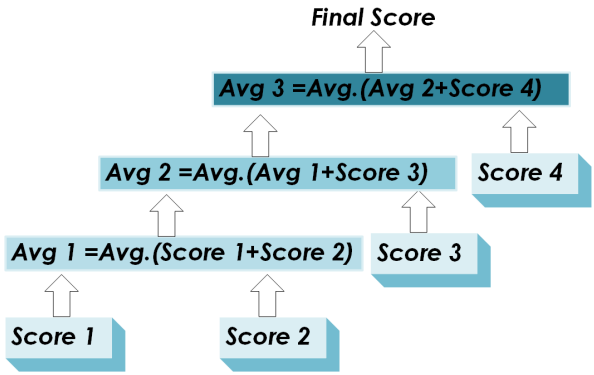

Figure 6: HO illustration. The rectangles are scores of sub-metrics and Avg.(·) is the average function.

Second, the Holistic (Pang et al., 2020) evaluates qualities of dialog. These qualities are Context Coherence, Language Fluency, n-gram-depended Response Diversity, and Logical Self-Consistency via a fine-tuned GPT-2 (Radford et al., 2019).

Third, based on GPT-2 (Brown et al., 2020), Dialog Ranking Pretrained Transformers (DialogRPT) (Gao et al., 2020) leverages social media feedback data to rank the machine-generated dialog responses. DialogRPT contains four sub-metrics: (1) Width, (2) Depth, (3) Updown, and (4) human-vs-fake, which compare random human responses and machine-generated responses.

C Example of Prompts

This section illustrates examples of the one-shot prompts for IRI-Perspective Taking (Table. 6), IRI-Empathetic Concern (Table. 7), RoPE-Empathetic Understand (Table. 8), and ROPE-Empathic Response (Table. 9).

D A Graphic Illustration of HO

The idea behind Huffman coding is to minimize the average code length. We can assign shorter codes to more frequently occurring data elements and longer codes to less frequently occurring ones. For comparison, our "Huffman-order" assigns larger weights to a sub-metric that is strongly related to humans. I.e., in each step, two sub-metrics corresponding to the smallest and the second smallest correlation coefficient are combined.

Figure 6 illustrates the process: four sub-metric scores produce a final score.

E Experiment Settings

E.1 Experiment Settings

Our experiment settings are in Table. 10. LoRA is used to fine-tune Qwen, LLaMa, and FLAN-T5.

E.2 Computational Cost

We conduct experiments with GPT and DeepSeek APIs. As the output only contains empathy scores, we only spent $100 in total. Additionally, with two 3090 GPUs, we spent 45 hours on training EMPATH and all compared models.

F More Experiments

F.1 More Ablation Study

The results of more ablation studies are in Table. 11. Table. 11 compares the combination of any two of the four metrics (S_C, S_{ED}, S_P, and S_W). It turns out that there are three combinations, each of which has its strengths.

F.2 More Experiments for Classical Dialog Evaluation Metrics

Besides all sub-metrics of C-PMI, Holistic and DialogRPT, we test all sub-metrics of EMH (Sharma et al., 2020), VADER (Gilbert, 2014), SOME (Yoshimura et al., 2020), and USL-H (Phy et al., 2020). Further, classical dialog evaluations like DENSITY (Park et al., 2023), GRADE (Huang et al., 2020), and MDD-Eval (Zhang et al., 2022) are tested. Moreover, we try a simple metric, length (i.e., the number of words in the response).

Empathic Motivation Hypothesis There are three sub-metrics in the EPITOME (Sharma et al., 2020) dataset. 1) Emotional Reactions. Emotional reactions are the response's intensity (no/weak/strong). Expressing warmth, compassion, and concern is essential in establishing empathic rapport and support (Elliott et al., 2011). It is a part of the EPITOME (Sharma et al., 2020). 2) Interpretation. Interpretation is another part of the EPITOME (Sharma et al., 2020). A weak interpretation contains a mention of the understanding. By contrast, a strong interpretation specifies the inferred feeling. 3) Explorations. Explorations is a part of the EPITOME (Sharma et al., 2020). It indicates the seeker's understanding by exploring the feelings and experiences, another critical aspect of empathy (Elliott et al., 2011). The levels of exploration are: no exploration; weak exploration,

INSTRUCTION:
You are a psychology expert. Consider User1's utterance: ##{user1-context}## and User2's response: ##{user2-response}##. Evaluate User2's response based on the criteria below. Ensure you provide a score for each criterion without any missing values.
SCORING CRITERIA:
Rate each criterion from 1 to 5:
- 1 = Lowest level of alignment
- 5 = Highest level of alignment
RATING CONTENT:
1. User2 successfully understands User1's perspective or emotions.
2. User2 considers User1's viewpoint before replying.
3. User2 actively attempts to empathize with User1 by imagining how User1 might feel or think in their situation.
4. User2 considers User1's arguments even when confident in their own view.
5. User2 recognizes multiple valid perspectives in a conversation, including User1's, and makes an effort to understand them before responding.
6. When frustrated by User1's response, User2 tries to see the situation from User1's point of view to better understand their position.
7. Before providing feedback, User2 considers how User1 might feel in response to their words.
OUTPUT FORMAT:
"scores": "[Score1, Score2, Score3, Score4, Score5, Score6, Score7]",
"average-score": "Calculated Average"

Table 6: An example of the one-shot IRI-Perspective Taking prompt. Section names are in brown and text variables are in curly brackets.

INSTRUCTION:
You are a psychology expert. Consider User1's utterance: ##{user1-context}## and User2's response: ##{user2-response}##. Evaluate User2's response based on the criteria below. Ensure you provide a score for each criterion without any missing values.
SCORING CRITERIA:
Rate each criterion from 1 to 5:
- 1 = Lowest level of alignment
- 5 = Highest level of alignment
RATING CONTENT:
1. User2 shows tender, concerned feelings in their response to User1, demonstrating empathic concern.
2. User2 displays understanding and care in their response when User1 is having problems.
3. When User1 is treated unfairly, User2's response reflects a protective tendency.
4. User2's response shows empathy toward User1's misfortunes.
5. User2 expresses pity and concern in their response when User1 faces unfair treatment.
6. User2 is moved by User1's experiences and shows it in their response.
7. User2's response is soft-hearted and empathetic.
OUTPUT FORMAT:
"scores": "[Score1, Score2, Score3, Score4, Score5, Score6, Score7]",
"average-score": "Calculated Average"

Table 7: An example of the one-shot IRI-Empathic Concern prompt. Section names are in brown and text variables are in curly brackets.

which is generic; and strong exploration, which is specific and labels the seeker's experiences and feelings. We fine-tune the DeBERTa on EPITOME and use it as our model for these metrics.

VADER **V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner (VADER) (Gilbert, 2014) is a rule-based sentiment analysis system. Our implementation uses the NLTK[8] with a publicly available dictionary [9]. Three subcategories of these sentiment words are Valence, Arousal, and Dominance.

SOME The **S**ub-metrics that are **O**ptimized for **M**anual **E**valuation (SOME) (Yoshimura et al., 2020) is a grammatical error correction metric. It contains three submodels: a grammar model, a flu-

[8] NLTK.

[9] The NRC Valence, Arousal, and Dominance (NRC-VAD) Lexicon.

INSTRUCTION:
You are a psychology expert. Consider User1's utterance: ##{user1-context}## and User2's response: ##{user2-response}##. Evaluate User2's response based on the criteria below. Ensure you provide a score for each criterion without any missing values.
SCORING CRITERIA:
Rate each criterion from 1 to 5:
- 1 = Lowest level of alignment
- 5 = Highest level of alignment
RATING CONTENT:
1. User2 accurately understands User1's feelings about their experiences.
2. User2 knows User1 and their needs.
3. User2's responses show care and consideration for User1's emotions.
4. User2 understands Use1's perspective and feelings.
5. User2 perceives and accepts User1's individual characteristics.
6. User2 understands the whole of what User1's utterance means.
7. User2's responses address both User1's words and emotions.
8. User2's reactions indicate sensitivity to User1's negative emotions.
OUTPUT FORMAT:
"scores": "[Score1, Score2, Score3, Score4, Score5, Score6, Score7, Score8]",
"average-score": "Calculated Average"

Table 8: An example of the one-shot RoPE-Empathetic Understand prompt. Section names are in brown and text variables are in curly brackets.

INSTRUCTION:
You are a psychology expert. Consider User1's utterance: ##{user1-context}## and User2's response: ##{user2-response}##. Evaluate User2's response based on the criteria below. Ensure you provide a score for each criterion without any missing values.
SCORING CRITERIA:
Rate each criterion from 1 to 5:
- 1 = Lowest level of alignment
- 5 = Highest level of alignment
RATING CONTENT:
1. User2 adjusts responses based on User1's emotions or thoughts, creating a positive impact on User1.
2. User2 reacts personally to User1's shared information.
3. User2 comforts User1 when upset.
4. User2 encourages User1 when needed, offering support during critical moments.
5. User2 praises User1's achievements or efforts.
6. User2 provides help when User1 needs assistance.
7. User2's responses are personalized and engaging.
OUTPUT FORMAT:
"scores": "[Score1, Score2, Score3, Score4, Score5, Score6, Score7]",
"average-score": "Calculated Average"

Table 9: An example of the one-shot RoPE-Empathic Response prompt. Section names are in brown and text variables are in curly brackets.

ency model, and a meaning model, which combines the grammar model and the fluency model. While the predicted score is the output of the meaning model, we test all three models.

USL-H **U**nderstandability, **S**ensibleness, and **L**ikability in **H**ierarchy (USL-H) (Phy et al., 2020) composite metrics of many aspects to obtain a single metric. In this paper, we test nine submetrics of USL-H. BERT-VUP, BERT-NUP, six classical machine learning metrics, and the Simplified USL-H. The BERT-VUP is the **V**alid **U**tterance **P**rediction task. It is a BERT-based model that captures the understandability of an utterance by classifying whether it is valid. The BERT-NUP is the **N**ext **U**tterance **P**rediction task, which is used as the metric for the sensitivity. Most submetrics of USL-H are borrowed from classical machine learning. E.g., NLL, normalized NLL, cross entropy, normalized cross entropy, perplexity, and normalized perplexity. At last, the Simplified USL-H simplified the overall score by Equation 3 in their paper (Phy

Hardware/Software	Setting
OS	Ubuntu 20.04.1 LTS
CPU	Intel Core i9-10900K
GPU	RTX 3090*2
Python	3.9.6
PyTorch	1.9.0
Models for ConText	RoBERTa[1]
Models for EPITOME	RoBERTa[2]
Models for YNU-HPCC	DeBERTa & RoBERTa[3]
Classical dialog metrics	GPT-2 & BERT[4]
GPT-4o	gpt-4o-2024-08-06 (API)
QWen (LoRA)	Qwen2.5-7B-Instruct[5]
LLaMA (LoRA)	LLaMA3.1-8B-Instruct[6]
FLAN-T5 (LoRA)	FLAN-T5-large[7]
Batch size	4
Epoch	30
Loss function	Mean Squared Error
Learning rate	1e-5
Learning rate schedule	cosine
Fp-16	False
Early stopping	False
Gradient accumulation steps	2
Seed	42

[1] We following the settings provided by the paper (Pereira et al., 2024).
[2] We following the settings provided by the paper (Sharma et al., 2020).
[3] We following the settings provided by the paper (Sharma et al., 2020).
[4] We following the settings provided by the papers (Pang et al., 2020; Phy et al., 2020; Ren et al., 2023; Gao et al., 2020).
[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.
[6] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct.
[7] https://huggingface.co/google/flan-t5-large.

Table 10: Experiment settings.

Metric	W(ρ)[1]	EX(ρ)	W(r)	EX(r)
EMPATH	0.5754	0.4536	0.6070	0.4860
$S_C + S_{ED}$	0.5064	0.3721	0.5311	0.3917
$S_C + S_P$	0.5025	0.3753	0.5316	0.3820
$S_C + S_W$	0.5394	**0.4230**	0.5599	**0.4571**
$S_P + S_{ED}$	0.5406	0.3488	0.5705	0.3556
$S_P + S_W$	**0.5629**	0.4209	0.5873	0.4377
$S_W + S_{ED}$	0.5594	0.4082	**0.5914**	0.4287

[1] W: WASSA.

Table 11: The r and ρ between the score generated by ablations and the empathy scores on WASSA and EX. All correlations are statistically significant ($p < 0.05$). Best results (except EMPATH) are highlighted in bold; the second best results (except EMPATH) are underlined.

et al., 2020).

Other Classical Metrics Three commonly used classical metrics are also tested. 1) DENSITY (Park et al., 2023) (**D**ialogue **E**valuation metric using **DeNSITY** **E**stimation) uses density estimation for dialog evaluation. 2) GRADE (Huang et al., 2020) incorporates both utterance-level and topic-level representations to evaluate dialogue. The name comes from **G**raph-enhanced **R**epresentations for **A**utomatic **D**ialogue **E**valuation. 3) MDD-Eval (Zhang et al., 2022) is a **M**ulti-**D**omain **D**ialogue evaluation framework.

Simple Metric We use the number of words in the response as a metric.

Table. 12 lists our results. Interestingly, as the simplest metric, length is a very effective single metric.

F.3 More Ablation Study of ED

The results of more ablation studies of ED are in Table. 13.

In Table. 13, we test combinations of different numbers of emotion classes, different knowledge tags, and whether to use sentiment. There are six settings, and "zero emotion class" denotes that no emotion is used.

1. ED: sentiment & emotion (except zero emotion class) & knowledge tags. For WASSA, the knowledge tag is "Emotional Reaction". For EX, the knowledge tags are "xNeed" and "xWant".

2. w/o Knowledge: sentiment & emotion (except zero emotion class).

3. w/o Knowledge & w/o Sentiment: only emotion. In the zero-emotion class case, this setting is not applicable. For WASSA, this is also the case "w/o Sentiment & w/o Emotional Reaction". For EX, this is also the case "w/o Sentiment & w/o xNeed & w/o xWant".

4. w/o Sentiment: emotion (except zero emotion class) & knowledge tags. For WASSA, the knowledge tag is "Emotional Reaction". For EX, the knowledge tags are "xNeed" and "xWant".

5. w/o Sentiment & w/o xNeed: only the "xWant" tag is used for EX. This setting is not applicable to WASSA.

	Metric	WASSA(r)	EX(r)	WASSA(ρ)	EX(ρ)
C-PMI	C-PMI-Correct	-0.2774	-0.0805	-0.3191	-0.0794
	C-PMI-Fluent	-0.0384	-0.0704	-0.0343	-0.0566
	C-PMI-Relevant	-0.2845	-0.0884	-0.3329	-0.0880
	C-PMI-Semantically Appropriate	-0.3206	-0.2648	-0.3792	-0.2849
	C-PMI-Specific	0.2528	0.1900	<u>0.3424</u>	<u>0.2067</u>
	C-PMI-Understandable	-0.0307	-0.0804	-0.0154	-0.0734
DialogRPT	DialogRPT-eval-depth	0.1381	0.1488	0.1451	0.1597
	DialogRPT-eval-width	0.1524	0.1422	0.1694	0.1483
Empathic Motivation Hypothesis (EMH)	EMH-Explorations	-0.1035	0.0621	-0.0951	0.0617
	EMH-Emotional Reactions	0.0098	0.1615	0.0054	0.1425
	EMH-Interpretations	0.0730	-0.0492	0.0732	-0.0596
VADER	VADER-Valence	-0.0061	0.0770	0.0308	0.0817
	VADER- Arousal	0.2295	-0.0183	0.2000	-0.0339
	VADER-Dominance	0.0936	0.0843	0.1707	0.0970
SOME	SOME-grammar	-0.1009	-0.0818	-0.0920	-0.0565
	SOME-fluency	-0.1753	-0.1139	-0.1765	-0.0774
	SOME-meaning	-0.1410	-0.0982	-0.1338	-0.0669
USL-H	BERT-VUP	0.0509	0.0620	0.0369	0.0369
	BERT–NUP	0.0446	0.0678	0.0578	0.0515
	NLL	**0.3424**	<u>0.2201</u>	0.2927	0.2064
	Cross Entropy	-0.2202	-0.0389	-0.1924	-0.0144
	Perplexity	0.2202	0.0389	-0.0117	-0.1016
	Normalized NLL	0.3723	0.2004	**0.4202**	0.1843
	Normalized Cross Entropy	-0.1931	-0.0670	-0.1634	-0.0506
	Normalized Perplexity	0.1931	0.0670	0.1630	0.0482
	Simplified USL-H	0.2432	0.1450	0.2325	0.1438
Other Classical Metrics	DENSITY	0.1015	0.0605	0.0849	0.0600
	GRADE	0.1435	0.0258	0.1512	0.0209
	MDD-Eval	0.1027	-0.0029	0.0813	-0.0652
Simple Metric	length (# words)	<u>0.3234</u>	**0.2874**	0.3264	**0.2504**

Table 12: The r and ρ between the score generated by more classical metrics and the empathy scores on WASSA and EX. All correlations are statistically significant ($p < 0.05$). The best results are highlighted in bold, and the second-best results are underlined.

Emotion Classes	Metric	W(ρ)[1]	EX(ρ)	W(r)	EX(r)
7	ED	0.4936	0.2592	0.5256	0.2696
	w/o Knowledge[2]	0.4269	0.1782	0.4184	0.1871
	w/o Sentiment[3]	0.3788	0.1825	0.3849	0.1900
	w/o Sentiment	0.4809	0.1905	0.5126	0.1907
	w/o xNeed	N/A[4]	0.1396	N/A	0.1443
	w/o xWant	N/A	0.2154	N/A	0.2266
27	ED	0.4286	0.1001	0.4717	0.1089
	w/o Knowledge	0.3842	0.1325	0.4024	0.1270
	w/o Sentiment	**0.1834**	0.1248	**0.1851**	0.1155
	w/o Sentiment	0.4543	0.1562	0.4649	0.1556
	w/o xNeed	N/A	0.1785	N/A	0.1803
	w/o xWant	N/A	0.0892	N/A	0.1013
0[5]	ED	0.4512	0.1931	0.4872	0.1914
	w/o Knowledge	<u>0.3634</u>	0.0839	<u>0.3653</u>	0.0829
	w/o Sentiment	0.4393	0.0922	0.4755	0.1023
	w/o xNeed	N/A	**0.0356**	N/A	**0.0627**
	w/o xWant	N/A	<u>0.0593</u>	N/A	<u>0.0731</u>

[1] W: WASSA.
[2] w/o: without. This is a baseline in Fig. 4.
[3] w/o Knowledge & w/o Sentiment : only emotion.
[4] N/A: not applicable.
[5] Zero emotion class: w/o emotion.

Table 13: The r and ρ between the score generated by ablations of ED and the empathy scores on WASSA and EX. All correlations are statistically significant ($p < 0.05$). The worst results are highlighted in bold; the second worst results are underlined.

6. w/o Sentiment & w/o xWant: only the "xNeed" tag is used for EX. This setting is not applicable to WASSA.

The importance of "Sentiment" reveals itself in Table. 13.

F.4 Order Switching Test

To confirm the function of HO, we perform an order switching test, and the results are reported in Table. 14. Notice that switching the first two elements changes nothing.

It shows that swapping two metrics will degrade the performance. I.e., we should always keep our order according to the correlation ranking. However, minor changes (where the correlation coefficient of the two metrics is close) are tolerable. E.g., swapping S_P and S_{ED} is tolerable for WASSA regarding the Spearman correlation coefficient.

F.5 Zero-Shot and One-Shot Prompt Test

We compare the effect of using zero-shot and one-shot prompt strategies in Table. 15, with different LLMs on two datasets. We avoided using few-shot prompts that are probably too long for LLMs, as previous studies did (Furniturewala and Jaidka, 2024; Churina et al., 2024).

In Table. 15, the $1^{st} \sim 3^{rd}$ columns list the datasets, zero/one-shot setting, and tested models. Since each setting was tried multiple times, the $4^{th} \sim 5^{th}$ columns show the kappa coefficient (κ), which measures rater agreement, accounting for chance. A kappa of 1 indicates perfect agreement, 0 means agreement by chance, and negative values indicate worse than chance. To ensure rigorous analysis, both the mean ($\bar{X}$) and standard deviation (σ) are provided. Similarly, the $6^{th} \sim 9^{th}$ columns present the mean ($\bar{X}$) and standard deviation (σ) for Spearman correlation (ρ) and Pearson correlation (r).

Table. 15 shows that zero-shot and one-shot prompt strategies have similar effects.

F.6 Human Evaluation

To perform human evaluations, we use 100 samples randomly selected from WASSA and 100 samples randomly selected from EX datasets. The average score and deviation between the score generated by the compared metrics and the human-labeled empathy scores are in Table. 16.

Order	WASSA(ρ)	EX(ρ)	WASSA(r)	EX(r)
EMPATH ($S_C + S_{ED} + S_P + S_W$)	**0.5754**	**0.4536**	**0.6070**	**0.4860**
$S_C + S_P + S_{ED} + S_W$	**0.5754**	0.4522	**0.6070**	0.4800
$S_C + S_P + S_W + S_{ED}$	0.5629	0.4211	0.5936	0.4349
$S_C + S_{ED} + S_W + S_P$	0.5609	0.4274	0.5903	0.4393
$S_C + S_W + S_P + S_{ED}$	0.5590	0.4016	0.5896	0.4145
$S_C + S_W + S_{ED} + S_P$	0.5562	0.4121	0.5869	0.4198
$S_P + S_{ED} + S_C + S_W$	<u>0.5733</u>	<u>0.4535</u>	<u>0.6021</u>	<u>0.4859</u>
$S_P + S_{ED} + S_W + S_C$	0.5538	0.4348	0.5657	0.4588
$S_P + S_W + S_C + S_{ED}$	0.5545	0.4076	0.5837	0.4255
$S_P + S_W + S_{ED} + S_C$	0.5474	0.4250	0.5607	0.4434
$S_{ED} + S_W + S_C + S_P$	0.5494	0.4158	0.5809	0.4254
$S_{ED} + S_W + S_P + S_C$	0.5457	0.4239	0.5606	0.4378

Table 14: The r and ρ between the score generated by order-switching settings and the empathy scores on WASSA and EX. All correlations are statistically significant ($p < 0.05$). Best results are highlighted in bold; the second best results are underlined.

Dataset	Shot	Model	$\kappa\ \bar{X}$	$\kappa\ \sigma$	$r\ \bar{X}$	$r\ \sigma$	$\rho\ \bar{X}$	$\rho\ \sigma$
WASSA	zero-shot	GPT-4o	0.8658	0.0058	0.4339	0.0022	0.4092	0.0016
		GPT-4.1	0.8955	0.0048	0.3901	0.0059	0.3755	0.0066
		GPT-5	0.7641	0.0170	0.3385	0.0097	0.3318	0.0100
		DeepSeek v3	0.9218	0.0033	0.3881	0.0010	0.3777	0.0008
		DeepSeek v3.1	0.8840	0.0064	0.4003	0.0010	0.3863	0.0011
	one-shot	GPT-4o	0.8521	0.0119	0.4169	0.0016	0.3992	0.0023
		GPT-4.1	0.8731	0.0112	0.4155	0.0063	0.3939	0.0064
		GPT-5	0.7599	0.0056	0.2944	0.0012	0.2965	0.0026
		DeepSeek v3	0.8731	0.0095	0.3498	0.0039	0.3505	0.0017
		DeepSeek v3.1	0.9055	0.0063	0.4020	0.0028	0.3854	0.0039
EX	zero-shot	GPT-4o	0.8832	0.0140	0.2820	0.0057	0.2768	0.0059
		GPT-4.1	0.8869	0.0075	0.2812	0.0062	0.2870	0.0049
		GPT-5	0.7077	0.0128	0.2598	0.0062	0.2616	0.0052
		DeepSeek v3	0.9053	0.0095	0.2603	0.0105	0.2641	0.0139
		DeepSeek v3.1	0.8252	0.0035	0.2759	0.0086	0.2737	0.0096
	one-shot	GPT-4o	0.8315	0.0161	0.2724	0.0075	0.2680	0.0065
		GPT-4.1	0.8153	0.0149	0.2869	0.0177	0.2832	0.0174
		GPT-5	0.7738	0.0262	0.2578	0.0116	0.2562	0.0113
		DeepSeek v3	0.8942	0.0015	0.2770	0.0048	0.2813	0.0032
		DeepSeek v3.1	0.8833	0.0080	0.2813	0.0152	0.2773	0.0154

Table 15: Correlation coefficients for different prompt settings. All correlations are statistically significant ($p < 0.05$), and all prompts are tested twice. κ: kappa coefficient, or Cohen's kappa, ρ: Spearman correlation, r: Pearson correlation, $\bar{X}$: mean, σ: standard deviation.

	Metric	WASSA	Deviation ↓	EX	Deviation↓
Label	Human	2.676	N/A	3.41	N/A
Ours	EMPATH	2.623	-0.053	3.161	-0.249
sub metric	S_W[1]	2.160	-0.516	3.359	-0.051
	S_P	2.138	-0.538	3.342	-0.068
	RoPE-Empathetic Understand	2.687	0.011	2.748	-0.662
	RoPE-Empathic Response	2.537	-0.139	2.42	-0.990
	IRI-Empathetic Concern	2.325	-0.351	2.247	-1.163
	IRI-Perspective Taking	3.652	0.976	3.167	-0.243
	S_{ED}	1.944	-0.732	3.384	-0.026
	S_C	2.217	-0.459	3.392	-0.018
WASSA winner	YNU-HPCC	2.134	-0.542	3.763	0.353
	ConText	2.033	-0.643	3.533	0.123
small LM	BERT	1.999	-0.677	3.665	0.255
	RoBERTa	2.222	-0.454	3.719	0.309
	DeBERTa	2.323	-0.353	3.540	0.130
open source LLM	Flan-T5	2.179	-0.497	3.853	0.443
	GPT-2	2.040	-0.636	3.577	0.167
	LLaMA3.1	2.115	-0.561	3.412	0.002
	Qwen2.5	1.877	-0.799	3.151	-0.259
closed source LLM	GPT-4o	3.070	0.394	2.85	-0.560
	DeepSeek v3	2.920	0.244	2.61	-0.800

[1] S_W: combined score of WASSA metrics; S_P : combined score of psychological metrics; S_{ED} : ED score ; S_C : combined score of classical metrics.

Table 16: The average score and deviation between the score generated by the compared metrics and the human labeled empathy scores of 100 samples randomly selected from WASSA and 100 samples randomly selected from EX datasets. See Appx. A Table. 5 for abbreviations.