Automatic Evaluation of Open-Domain Real Conversations: Combining Encoder-Based, Dialogue-Based Features and Large Language Models Ratings

Cristina Conforto-López, Marcos Estecha-Garitagoitia, Mario Rodriguez-Cantelar, Ricardo de Córdoba, Luis Fernando D'Haro

Information Processing and Telecommunications Center (IPTC) - ETSI de Telecomunicación
Universidad Politécnica de Madrid
Avenida Complutense 30, 28040, Madrid, Spain

Correspondence: luisfernando.dharo@upm.es

Abstract

Conversational AI is a central application of NLP, yet ensuring high response quality remains challenging due to the inherently subjective nature of user satisfaction. Dialogue evaluation can be performed manually—through expert or user ratings—or automatically, using methods that aim to predict quality scores consistent with human judgment.

In this work, we present a reference-free automatic dialogue evaluation system that predicts user ratings from a dataset of real human–chatbot interactions collected during the Alexa Prize Socialbot Grand Challenge 5, combining multiple complementary models to enhance correlation with human scores.

Experimental results indicate that the model that achieves the highest Pearson correlation with users' ratings is an XGBoost regression model that combines different features such as conversation length, engineered flags capturing conversation characteristics, predictions from an Encoder-based Panel of Experts (PoE), and instruction-based outputs from a fine-tuned LLM. The overall Pearson Correlation on the eval set is 0.404, which is competitive with prior work trained on an order of magnitude more dialogues, albeit using different datasets and system configurations.

1 Introduction

Natural Language Processing (NLP) has advanced chatbot development, enabling real-time interactions in customer service and virtual assistants. A key challenge is dialogue evaluation, which can be manual—requiring expert annotators or ratings from users—or automatic based on using algorithms. Automatic methods include reference-based approaches, comparing responses to predefined answers, and reference-free approaches, which assess conversations without a reference.

Developing effective reference-free evaluation models is crucial for improving chatbot performance and user experience, especially in open-domain dialogues where multiple answers are allowed. However, the problem remains challenging due to the subjective nature quality assessment.

In recent years, much research has been done in the field. One of the first learning-based metrics was Automatic Dialogue Evaluator or ADEM (Lowe et al., 2017a), which proposes an alternative to reference-based metrics by training an RNN using a variety of dialogue datasets. The creation of transformers (Vaswani et al., 2017) was a turning point in automatic dialogue evaluators since they enabled a deeper understanding of language.

DialogRPT (Gao et al., 2020) builds upon GPT-2 and is finetuned using data from human votes and interactions done on posts of Reddit. It introduces a fully reference-free metric that focuses on aspects such as relevance or engagement by implementing a modular architecture that focuses on these aspects. Panel of Experts (PoE) (Zhang et al., 2023) is a model built on the RoBERTa architecture that uses small, efficient components called adapters. It uses a "multitask learning" approach, which means it learns several different tasks at the same time. This helps the model: Identify common patterns shared across different tasks, improve general understanding by looking at the "big picture.", and prevent overfitting, ensuring the model doesn't become too specialized in just one area at the expense of others.

Later, Amazon proposed another approach using ratings from users conversing with their Alexa devices (Le et al., 2023). They proposed a combination of a transformer model trained directly from dialogue text and a model trained on turn-level user features such as disinterest or compliments. Given that for this case, the ratings were provided by a non-specialized and random set of users, it becomes harder to achieve higher correlations.

The emergence of LLMs has led to their direct use as dialogue evaluators (Zhang et al., 2024a; Li

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 52–63
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

et al., 2024a; Mendonça et al., 2024), often through prompt engineering to specify the evaluation procedure. While LLMs can capture complex linguistic patterns and generate structured assessments, their performance remains limited when applied to real-user ratings. Recent work, such as USR (Mehri and Eskenazi, 2020) and G-Eval (Liu et al., 2023), has demonstrated promising results using LLM-based or multi-dimensional scoring methods; however, these are generally evaluated on static benchmark datasets, while our approach targets real-user ratings collected from an active chatbot deployment.

This paper introduces a hybrid ensemble model trained on a dataset of chatbot-user conversations rated on a 1–5 scale. Unlike existing datasets with expert annotations, this dataset captures real user subjectivity, introducing evaluation challenges. The paper is structured as follows: Section 3 describes the dataset used in this study that was collected during the Alexa Prize SocialBot Grand Challenge 5 (Johnston et al., 2023). Section 4 outlines the experimental setup, rationale, and procedures for developing the different components of the evaluator. Then, section 5 presents the results and findings. Finally, section 6 summarizes key insights and future research directions.

2 Related Work

Automatic dialogue evaluation has received growing attention, particularly in *reference-free* settings. Early approaches, such as ADEM (Lowe et al., 2017b), employed learning-based metrics using small RNN architectures. More recent systems, including USR (Mehri and Eskenazi, 2020) and G-Eval (Liu et al., 2023), introduced structured human annotations and graph-based prompt templates to assess dialogues along dimensions such as coherence, fluency, and engagement.

While USR applies hierarchical modeling and G-Eval leverages LLM prompting with score rationales, our approach differs in several key ways. We combine encoder-based models with dialogue-level evaluators and integrate flag-based features extracted directly from real-user conversations. Furthermore, whereas USR and G-Eval are trained on expert-labeled public datasets, our models are trained on user-provided ratings collected during live deployment, enabling more realistic and context-specific assessments.

Recent studies have continued to advance automatic dialogue evaluation, particularly in real-user and reference-free settings. For example, (Lee et al., 2025) introduced RealTalk, a benchmark designed to better capture the challenges of evaluating real-world conversations, highlighting the limitations of traditional static datasets. (Ito et al., 2025) conducted a survey of reference-free metrics, underscoring the importance of robustness across diverse dialogue scenarios. Within the context of LLM-based evaluators, (Chiang et al., 2024) presented Chatbot Arena, a large-scale human-feedback benchmark for comparing conversational models, while (Zhang et al., 2024b) provides a comprehensive study on the application of LLMs for automatic dialogue evaluation, probing the large advances on using LLMs-as-judges.

3 Database

The dataset consists of conversations collected with real Alexa users by one of the participant teams during the Alexa Prize SocialBot Grand Challenge 5. These dialogues are anonymized transcriptions that were collected through spoken communication between users and their Alexa devices as the primary mode of interaction. The audio signal itself was not accessible for use in this work. Additionally, and depending on the final users' device capabilities, a screen displayed relevant images based on the conversation topic.

The dataset captures a wide range of user-driven discussions, spanning topics such as sports, movies, daily life, science, or geography. Users from diverse backgrounds and locations engaged in these conversations, contributing to the dataset's variability. Each user had full control over the conversation length and could rate the chatbot's performance at the end of the interaction. Ratings ranged from 1 (poor quality) to 5 (highly satisfactory).

From the total number of dialogues collected during our participation, we sampled a subset of dialogues. First, we removed those without user ratings ("rated-data"), resulting in around 16,000 conversations (290k turns). Next, we excluded interactions with only one or two turns, as we found in our initial analysis that the ratings for these dialogues leaned towards extremes (highest/lowest satisfaction). Finally, only rated conversations with at least five turns were retained ("filtered-data"), reducing the dataset to 13,000 conversations (260k turns). This reduction was driven primarily by the removal of incomplete, duplicated, or low-information turns that did not contribute meaningful signal to the analysis. Importantly, the filtering process was not topic-based and therefore did not preferentially exclude specific types of interactions. As a result, the retained subset preserves the diversity and coverage of the original dataset while improving overall data quality. We therefore do not expect the reduction to result in the loss of relevant or interesting issues, but rather to enhance the robustness and interpretability of the reported results.

To perform our experiments, the dataset was split first into two parts: a) an Eval set consisting of 10% of the data (∼1300 dialogues) reserved for final evaluation to ensure an unbiased assessment of model performance (i.e., allowing a comparison between the results for the test sets and the eval set), and b) a 90% of the data used to perform 5-Fold Cross-Validation. This 90% dataset (∼12000 dialogues) was divided into the following subsets: (a) Train: 3 out of the 5 Folds were used for training or fine-tuning the model, (b) Dev: 1 Fold used to evaluate performance and guide hyperparameter selection, and (c) Test: the remaining 1 Fold for assessing model performance on unseen data after training.

4 Methodology

As outlined in the introduction, automatic dialogue evaluation methods can be classified along two main axes: *reference-based vs. reference-free* and *dialogue-level vs. turn-level*. Given our dataset's characteristics, we established the following model requirements: 1.) Reference-free: Our dataset lacks reference responses, 2.) Numerical outputs: The model must produce continuous scores to match the dataset's rating format, 3.) Multi-domain capability: The model should handle diverse topics and domains, 4.) Fine-tuning support: Adaptation to dataset-specific characteristics is necessary, 5.) Strong correlation: The model should achieve high Pearson correlation between predicted and actual ratings, and 6.) Local deployability: Data confidentiality must be preserved.

The remainder of this section describes the selected models: - Pre-LLM state-of-the-art encoder-based models (Section 4.1) - Zero-shot and fine-tuned LLMs (Sections 4.2.1 and 4.2.2) - Regression models using dialogue-derived flag-based features (Sections 4.2.3 and 4.3)

Finally, we describe the regression-based combination model that integrates predictions from all approaches (Section 4.4).

4.1 Encoder-based models

For pre-LLM SotA models, we selected the Panel of Experts (PoE) model (Zhang et al., 2023) which is intended for turn-level.[1] PoE is trained on five distinct datasets: DailyDialog (Li et al., 2017), ConvAI2 (Logacheva et al., 2019), TopicalChat (Gopalakrishnan et al., 2023), EmpatheticDialogue (Rashkin et al., 2018), and Reddit (Huryn et al., 2022). Each dataset is associated with a dedicated adapter, enabling multi-domain adaptability and the modeling of diverse evaluation dimensions (e.g., overall score, engagement, naturalness).

Since PoE is turn-based, it processes interactions in the format: `rating ||| user 1 input ||| user 2 response`. Instead of evaluating the entire dialogue, it focuses on the appropriateness of user 2's response given user 1's input.

Since our dataset provides dialogue-level ratings rather than turn-level annotations, we adapted PoE using a cyclic turn-evaluation approach. The input format for training was modified to: `rating ||| context ||| chatbot response`, where context includes a fixed number of preceding turns (minimum of 5), and response is the chatbot's subsequent utterance. This adaptation aligns PoE with our dataset's rating granularity while preserving its evaluation methodology[2].

For fine-tuning, we generated three subsets per fold, each varying in context length: (a) 3-turn, (b) 4-turn, and (c) 5-turn contexts. Further detail on the distribution of the turns for each case can be seen in appendix A.1.

Contexts longer than five turns were excluded due to the model's 512-token input limit, which is typically exceeded with extended interactions.

After restructuring the data to match PoE's input format, we obtained a total of 183,000 turns, distributed in sets of 3, 4 or 5 turns and then distributed according to each Fold distribution, i.e., 60% for training, 20% for testing, and 20% for development. We hypothesize that longer contexts will improve performance, as additional conversational history provides more context for rating assignments. Our experiments aim to validate this hypothesis and

[1] FinED-Eval (Zhang et al., 2022a), which is intended for evaluation at dialogue level, was initially considered, but discarded after preliminary experiments showed that fine-tuning did not yield improvements over PoE.

[2] Since the dataset lacks turn-level annotations, this approach was the only feasible way to apply PoE to our data. To enable a fairer comparison with dialogue-based models, we experimented with multiple context lengths, allowing the turn-level evaluation to approximate dialogue-level reasoning.

quantify its impact on evaluation accuracy.

4.2 Large Language Models - LLMs

Considering the large improvements in using LLMs as judges for automatic dialogue evaluation (Li et al., 2024b; Gu et al., 2024; Zheng et al., 2023), we consider it important to test their capabilities in real-user settings. Therefore, we established additional criteria for selecting suitable models for our study: (a) The LLM must be instruction-tuned, as such models demonstrate superior performance in evaluation tasks (Dai et al., 2024), and (b) To comply with computational constraints, models larger than 8 billion parameters were excluded, as they are infeasible to fine-tune or deploy on our current hardware. Based on these criteria, we selected two models for different steps of the methodology.

Qwen 2.5 7B Instruct (Yang et al., 2024), an open-source model by Alibaba, is available on Hugging Face. Its 7B parameter size aligns with our computational constraints, and its instruction-tuning enhances evaluation performance. At the time of selection, it ranks among the top lightweight open-source models in the Judge Arena[3], making it the primary model for our experiments. This will be the model used to perform the Prompt-engineering 4.2.1 and Finetuning 4.2.2 experiments. The main reason for this choice is that, at the time, its performance as a judge outperformed all other lightweight models, achieving results comparable to proprietary models with hundreds of billions of parameters.

On the other hand, we selected Llama 3.2 3B Instruct (Grattafiori et al., 2024), an open-source model by Meta available on Hugging Face. The Llama series also show strong performance in Judge Arena and LLM evaluator benchmarks. This version was chosen for its recent release—which incorporates updated training data and techniques—while maintaining a small parameter footprint, enabling efficient deployment on our hardware. It is the model we used for the Flag Extraction experiment (Section 4.2.3).

4.2.1 Prompt-engineering

Our objective is to apply prompt engineering techniques to design effective instructions that guide the model to accurately perform the task—specifically, generating a rating for a given dialogue.

We adopt an iterative approach, starting with a simple prompt and incrementally refining it to emphasize aspects that improve model performance. To minimize variability across steps, we employ a zero-shot setup (i.e., no in-prompt examples). In addition, among the configurable LLM parameters, we focus on temperature, keeping other settings fixed. Temperature controls output randomness: lower values yield more focused and deterministic responses, while higher values increase diversity. To ensure reproducibility and reduce variance, we use a low temperature and iteratively engineer the prompt to maximize correlation with human-provided ratings.

4.2.2 Finetuning

This experiment focuses on fine-tuning the selected LLM to adapt its internal parameters for evaluating dialogues in our dataset. Although LLMs can perform various tasks in a zero-shot manner due to their training on large, diverse datasets, fine-tuning can enhance performance for specific tasks.

For fine-tuning, we use the LLaMA-Factory framework (Zheng et al., 2024). To accelerate experimentation, we combined the prompts identified in section 4.2.1, that maximized Pearson correlation. We fine-tuned the model using this prompt and the dialogues from our database. Low-Rank Adaptation (LoRA) (Hu et al., 2022) was applied to improve computational efficiency. Five models were trained, one per data fold, and evaluated on the respective test set.

4.2.3 Dialogue-based Feature extraction

During the Alexa Prize competition, we identified indicators strongly correlated with human ratings and feedback and developed an automated method to extract them.

The resulting set of features, referred to as flags, was defined through manual analysis of conversations with diverse ratings to identify recurring user patterns. Flag extraction was performed using prompt-based instructions, following the methodology described in Section 4.2.1. For each flag, the LLM received an instruction prompt and the conversation as input to determine whether the characteristic was present.

Each flag was validated against manually annotated labels on 50 conversations, and prompt or definition refinements were applied when discrepancies were observed. In addition to the LLM-based flags, conversation length in number of turns was included as a feature and extracted directly.

Flag extraction was carried out using Llama 3.2

[3] https://huggingface.co/spaces/AtlaAI/judge-arena

3B Instruct. Its smaller size enables faster inference, and using a different model from the downstream Qwen 2.5 7B Instruct reduces potential self-referential bias.

Six binary flags were defined: confusion, angry, engaging, loop, toxic, and correction. Their prompts and occurrence rates are reported in Table 7 and Appendix A.4. The most frequent flags are loop and correction, followed by angry, while engaging and confusion occur at similar rates. Toxic interactions are rare.

4.3 Regression Model using Dialogue-based Features

This experiment applies classical machine learning models for rating prediction. In this case, it takes the detected dialogue features (section 4.2.3) to derive numerical scores using regression models. The selected models are classical machine learning algorithms suited for regression in a supervised learning framework: Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016).

The input features for these models include numerical characteristics extracted from each conversation, primarily sourced from the flag extraction experiment (Section 4.2.3), along with an additional feature: the total number of dialogue turns.

4.4 Models Combination via Regression

The final approach aims to enhance overall performance by combining the different described approaches (i.e., the encoder-based model from Section 4.1, LLMs from Section 4.2, and dialogue features from Section 4.2.3). By integrating the best predictions from these methods, the goal is to maximize Pearson correlation with the true ratings.

The combination strategy involves training a regression model using classical machine learning techniques; in this case, the same tested in Section 4.3, where the input features are predictions from the various models. Specifically, the following model combinations will be explored: (a) a regression model using dialogue features and PoE predictions, (b) a regression model using dialogue features and fine-tuned LLM predictions, and (c) a regression model incorporating PoE predictions, dialogue features and fine-tuned LLM predictions.

5 Results

5.1 Encoder-based models

This section presents the results of the Encoder-based (PoE) model experiment. The pre-trained PoE model includes 5 adapters, each trained on a different dataset as described in Section 4.1. For this experiment, we fine-tuned all 5 adapters with our data, combining the diverse knowledge and context they provide with the characteristics of our new dataset. Each adapter was fine-tuned and evaluated in parallel, and then averaged to contribute to the final output rating. The learning rate was set to 1×10^{-5}, and training was conducted for up to 10 epochs, although early stopping based on the patience parameter typically resulted in convergence after an average of 2 epochs per model.

The fine-tuning data followed a cyclic format (see section 4.1 and appendix A.1), where each dialogue was divided into multiple inputs consisting of a fixed number of context turns and a corresponding response. The model predicted a rating for each individual turn (since the dataset is at the dialogue level, the same overall score was applied to all turn scores within a given dialogue). The final dialogue-level score is calculated as the average of all turn-level ratings considering the corresponding context length, allowing for direct comparison with human ratings.

The results are summarized in Table 1, which shows two models: **PoE-base** (the pre-trained model before fine-tuning) and **PoE-ft** (the model after fine-tuning). Fine-tuning was performed with context lengths of 3, 4, and 5 turns. The table reports the average Pearson correlation across the 5-fold test sets for each context length.

The PoE-base model exhibits very low correlation across all context sizes, indicating poor performance on the dataset. In contrast, PoE-ft shows significant improvement, with Pearson correlation increasing as the number of context turns increases. This suggests that incorporating more dialogue context improves the model's ability to predict ratings more accurately. The best performance was achieved by PoE-ft with a 5-turn context, which aligns with our hypothesis that more context would enhance model performance. While the improvement is modest, the trend is consistent: increasing context length improves Pearson correlation. Specifically, the 5-turn variant outperforms the 3-turn version by approximately 7%.

Model	# Turns	Pearson Correlation
PoE-base	3 turns	-0.046 ± 0.022
	4 turns	-0.037 ± 0.020
	5 turns	-0.047 ± 0.023
PoE-ft	3 turns	0.289 ± 0.010
	4 turns	0.302 ± 0.011
	5 turns	**0.309 ± 0.009**

Table 1: PoE results on the test set

5.2 Large Language Models (LLMs)

5.2.1 Prompt engineering

The prompt engineering experiment aimed to iteratively refine a prompt that enables the LLM to accurately perform the task using zero-shot learning (**LLM-zero shot**). The model used for this experiment was Qwen 2.5 7B Instruct. Multiple approaches were explored, tested, and subsequently discarded throughout the iterative process.

The initial approach involved designing prompts that emphasized different conversational dimensions typically evaluated in open-domain dialogues (Mehri et al., 2022), as well as using seed prompts available from different papers (Zhao et al., 2024; Zhang et al., 2024b; Mendonça et al., 2023; Zhang et al., 2022b). The rationale was to identify which traditional aspects (e.g., naturalness, coherence, fluidity, etc.) users prioritized when rating the chatbot. However, after testing various combinations, no dimension emerged as significantly more relevant, leading to the abandonment of this approach.

Our solution was to design a prompt that instructs the LLM to output the key aspects associated with the user's provided rating, rather than manually defining the conversational dimensions. By analyzing 30 dialogues per rating level, we derived descriptions for each score. The final prompt, based on these insights, achieved an average Pearson correlation of 0.212 ± 0.028. The prompt is provided in appendix A.2.

Despite our solution of defining score-level descriptors, we observed there was a model's tendency to predict ratings primarily between 2 and 3, rarely assigning 1 or 4, and never assigning 5. This behavior likely resulted from two factors: (1) the model rarely assigned the lowest rating, even for low-quality conversations, and (2) due to large advancements in generative AI, the LLM may have generated scores consistent with the evaluated chatbot being underperforming relative to modern benchmarks, producing rarely higher ratings (i.e., the LLM exhibited a conservative bias, likely due by its pretraining on higher-quality dialogues).

Finally, several other variations were tested but discarded due to lack of improvement. These include refining the prompt to evaluate only the chatbot's performance (i.e., focusing on chatbot turns) or replacing numeric ratings with textual labels.

5.2.2 LLM Fine-tuning

Once an optimal prompt was identified in the prompt engineering experiment, the next step was to fine-tune the LLM using this prompt (**LLM-ft**). Based on preliminary experiments, the selected model for this experiment was Qwen 2.5 7B Instruct, fine-tuned using LoRA.

The final hyperparameters and the average Pearson correlation obtained are summarized in Table 6 in appendix A.3. The learning rate and number of epochs were jointly optimized, resulting in a low learning rate paired with a higher number of epochs to enable gradual learning. The LoRA rank was set to an intermediate value to balance underfitting and overfitting, given the limited training set size. LoRA alpha was set to twice the rank, the standard configuration enhancing effectiveness, while LoRA dropout was kept low to maintain an optimal trade-off between training efficiency and generalization.

The fine-tuned model achieved a 21% improvement in average Pearson correlation compared to prompt-based evaluation alone. However, despite this improvement, the overall correlation remains lower than that obtained in the PoE-ft experiment with a 5-turn context (Table 1).

5.2.3 Extraction of Dialogue-based Features

5.3 Results using Dialogue-based Features

In this experiment, various numerical features are used to train classical machine learning models for rating prediction. The selected features include the binary flags previously extracted (section 5.2.3 and the conversation length, resulting in a total of seven features: *confusion*, *angry*, *engaging*, *loop*, *toxic*, *correction*, and *conversation length*.

Two types of regression models were trained: Support Vector Machines (SVM) and XGBoost. To identify the optimal hyperparameters for the models and training data, we performed GridSearch (systematic search). Table 8 shows the best hyperparameters with the corresponding average Pearson correlation. XGBoost emerged as the slightly best-performing model. The final Pearson correlation for XGBoost (0.245) is lower than the results from PoE-ft with 5 turns (0.309, see Table 1) and the finetuned LLM (0.257 in Table 2), but it surpasses the zero-shot performance observed in the prompt-engineering experiment (0.212).

5.4 Combined Model Results via Regression

In this experiment, classical machine learning regression models were trained using the best results from previous experiments. The combinations involved training a regression model (as in the regression experiment) with predictions from other models as additional inputs. Two models were tested—Support Vector Machines (SVM) and XGBoost—to determine the best fit for the training data using GridSearch (parameter sweep) for hyperparameter optimization.

Three input combinations were tested, each based on results from prior experiments. For each experiment, the predictions with the highest Pearson correlation were selected. The best results for the Encoder-based models came from the PoE-ft-5-turns model, while for LLM predictions, the best results came after fine-tuning the model.

For **PoE regression**: The best model combined dialogue features, conversation length, and PoE-ft-5-turns predictions. XGBoost outperformed PoE-ft-5-turns by nearly 10%, with the inclusion of PoE-ft-5-turns predictions and numeric features refining the model's dialogue rating ability.

For **LLM regression**: The best model combined flags, conversation length, and LLM-finetuned predictions. XGBoost outperformed LLM-ft by 18%, with the flags and conversation length providing additional valuable information.

For **PoE+LLM regression**: The best model combined flags, conversation length, PoE-ft-5-turns predictions, and LLM-finetuned predictions. XGBoost again delivered the best result, yielding an average Pearson correlation of 0.346. This performance surpasses PoE-ft-5-turns by 12% and the LLM-finetuned by 34%. With respect to the performance improvement over LLM regression the improvement is substantial, increased by 14%, while the improvement over PoE regression is more modest, at 2%. Although PoE+LLM regression shows a higher average Pearson correlation, the difference is not statistically significant. Comparative results are shown in Table 11, with detailed results for each regression algorithm in appendix A.5.

5.5 Final evaluation

Table 2 presents the comparative results for the best models from each of the previous experiments, evaluated on both the test and eval sets. The test results represent the average across the 5-folds, with approximately ∼2,400 dialogues per fold, while the eval set, consisting of ∼1,300 dialogues, was specifically reserved for final evaluation to assess performance on fully unseen data. This ensures that the adjustments made during the experiments were not influenced by the test set results. As shown, the results on the eval set follow a similar trend to those on the test set, with slightly higher Pearson correlation, especially for the final model that combines all the information.

Regarding the individual experiments, **PoE-ft-5-turns** achieved the highest performance, outperforming both LLM-ft and Feature-based approaches. Given the broader scope and larger training dataset of LLM-ft, it would be expected to yield superior results. However, PoE-ft-5-turns demonstrated better performance, likely due to the dataset characteristics. PoE, developed in 2022, was originally trained on dialogues more aligned with the chatbot used in this study, whose conversations, collected in 2023, may exhibit lower quality compared to current genAI models. In contrast, LLM-ft was trained on higher-quality data, making it less suited to adapting to the chatbot's dialogue style, thereby slightly reducing its overall performance.

Among the combination experiments, PoE + LLM achieved the highest Pearson correlation, as anticipated. Each model captures rating information from distinct perspectives: PoE adopts a turn-level approach, LLM operates at the dialogue level, and regression incorporates additional conversational features. Their integration enhances performance by leveraging complementary insights.

The performance of PoE + Dialogue features is close to that of PoE + LLM + Dialogue features, while LLM+regression lags behind. This suggests that incorporating LLM-ft predictions into PoE+regression has a limited impact. As shown in Table 2, PoE-ft-5-turns alone outperforms LLM-ft, reinforcing the dominant influence of PoE predictions in the PoE + LLM regression model.

6 Conclusion and Future lines

This paper presents various approaches for developing an automatic rating prediction system for open-domain dialogues between real users and con-

Table 2: Average Pearson Correlation obtained for the best model of each experiment on the test sets and the eval set

Model	Test Sets	Eval Set
PoE-ft-5-turns	0.309 ± 0.009	0.325 ± 0.006
LLM-zero shot (prompt engineering)	0.212 ± 0.028	0.225 ± 0.000
LLM-Fine-tuned	0.257 ± 0.026	0.264 ± 0.015
Dialogue Features	0.245 ± 0.013	0.308 ± 0.000
Dialogue Features + PoE-ft-5-turns predictions	0.339 ± 0.009	0.387 ± 0.006
Dialogue Features + LLM-ft predictions	0.304 ± 0.015	0.347 ± 0.006
Dialogue Features + PoE-ft-5-turns + LLM-ft predictions	**0.346 ± 0.015**	**0.404 ± 0.010**

versational systems.

Among the individual models, the best-performing approach is an encoder-based evaluation system using the Panel of Experts (PoE) model. Its superior performance stems from its specialized design for multi-domain and multi-dimensional dialogue evaluation, as well as its training data, which closely aligns with the evaluated dataset. Contrary to expectations, fine-tuning an LLM did not improve results, likely due to its strong alignment with higher-quality, natural dialogues. However, we would like to highlight our methodology of using the LLM to automatically analyze the motivations for users to provide a given rating and then incorporating these motivations into the prompt.

The highest overall performance was achieved by combining all proposed methods using classical machine learning models (XGBoost). This integration enhanced performance by leveraging diverse methodologies and complementary perspectives on the data.

Future research directions to further advance this work include: (a) Investigating alternative fusion strategies, such as incorporating model outputs into an LLM with a refined prompt to generate final predictions, (b) Exploring silver labeling techniques to assign reliable labels to many unlabeled conversations not considered in the experiments. Expanding the dataset with these labeled instances could improve model fine-tuning and performance, and (c) Extending experiments to LLMs with larger parameters and reasoning capabilities, alongside the necessary computational resources, to assess whether larger models offer a deeper understanding of dialogue characteristics.

Limitations

Despite the promising results reported in this paper, several limitations must be acknowledged. First, the performance of the evaluation model depends heavily on the quality and representativeness of the collected dialogues. Because the training data consists solely of conversations from a single chatbot, the extent to which the proposed methods generalize to other dialogue systems is uncertain. In addition, although fine-tuning and incorporating dialogue features improved performance, the resulting correlation values suggest that automatic dialogue evaluation remains challenging—especially in real-world settings where human judgments can be noisy and may differ substantially from traditional benchmark ratings. Finally, the use of relatively small LLMs limits the capacity to follow complex instructions and to assess subtle aspects of dialogue quality.

From an ethical standpoint, we adhered to the Alexa competition's guidelines, utilizing only anonymized automatic spoken dialogue transcriptions, and non-proprietary LLMs were never employed (which could potentially improve results). Although our results are favorable when compared with a similar model trained on 10 times more in-domain data (Le et al., 2023), a direct comparison is not possible due to different datasets, dialogue systems, and rating conditions. Therefore, future work should consider new annotated datasets and the development of normalization techniques for cross-dataset comparisons, should those datasets remain relevant.

Acknowledgments

This work was supported by Amazon Alexa Prize Sponsorship, the European Commission through Project ASTOUND (101071191 — HORIZON EIC-2021- PATHFINDERCHALLENGES-01), and by projects BEWORD (PID2021-126061OB-C43) and BRAINS (PID2024-155948OB-C52) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union".

References

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Kyle Dai, Maurice Burger, Roman Engeler, Max Bartolo, Clémentine Fourrier, Toby Drane, Mathias Leys, and Jackson Golden. 2024. Judge arena: Benchmarking llms as evaluators. *HuggingFace Blogs*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *Preprint*, arXiv:2009.06978.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Daniil Huryn, William M. Hutsell, and Jinho D. Choi. 2022. Automatic generation of large-scale multi-turn dialogues from Reddit. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3360–3373, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Takumi Ito, Kees van Deemter, and Jun Suzuki. 2025. Reference-free evaluation metrics for text generation: A survey. *arXiv preprint arXiv:2501.12011*.

Michael Johnston, Cris Flagg, Anna Gottardi, Sattvik Sahai, Yao Lu, Samyuth Sagi, Luke Dai, Prasoon Goyal, Behnam Hedayatnia, Lucy Hu, and 1 others. 2023. Advancing open domain dialog: The fifth alexa prize socialbot grand challenge.

Cat P Le, Luke Dai, Michael Johnston, Yang Liu, Marilyn Walker, and Reza Ghanadan. 2023. Improving open-domain dialogue evaluation with a causal inference model. *arXiv preprint arXiv:2301.13372*.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. 2025. Realtalk: A 21-day real-world dataset for long-term conversation. *arXiv preprint arXiv:2502.13270*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Varvara Logacheva, Valentin Malykh, Aleksey Litinsky, and Mikhail Burtsev. 2019. Convai2 dataset of non-goal-oriented human-to-bot dialogues. In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 277–294. Springer.

Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017a. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017b. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, and 1 others. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *arXiv preprint arXiv:2203.10012*.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

John Mendonça, Alon Lavie, and Isabel Trancoso. 2024. On the benchmarking of llms for open-domain dialogue evaluation. *arXiv preprint arXiv:2407.03841*.

John Mendonça, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple llm prompting is state-of-the-art for robust and multilingual dialogue evaluation. *arXiv preprint arXiv:2308.16797*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024a. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19515–19524.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024b. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19515–19524.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022a. FineD-eval: Fine-grained automatic dialogue-level evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3336–3355, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2023. Poe: A panel of experts for generalized automatic dialogue assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1234–1250.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Thomas Friedrichs, and Haizhou Li. 2022b. Investigating the impact of pre-trained language models on dialog evaluation. In *Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, pages 291–306. Springer.

Kun Zhao, Bohao Yang, Chen Tang, Chenghua Lin, and Liang Zhan. 2024. Slide: A framework integrating small and large language models for open-domain dialogues evaluation. *arXiv preprint arXiv:2405.15924*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Appendixes

A.1 Cyclic Context Data Structure for Panel of Experts

To clarify the cyclic context structure used for fine-tuning and evaluating the Panel of Experts (PoE) model, Table 3 illustrates the structure for a context of 3 turns, Table 4 for 4 turns, and Table 5 for 5 turns, all based on a 10-turn dialogue. Human interactions are denoted by "H" and chatbot interactions by "C," followed by the respective turn number. The ground truth dialogue score, provided by the real user, is assigned to all turns within a given dialogue. During evaluation, the predicted score across all turns is averaged and then used for correlation purposes.

Context	Response
H 1 + C 1 + H 2	C 2
H 2 + C 2 + H 3	C 3
H 3 + C 3 + H 4	C 4
...	..
H 9 + C 9 + H 10	C 10

Table 3: Data distribution for a context of 3 turns

Context	Response
C 1 + H 2 + C 2 + H 3	C 3
C 2 + H 3 + C 3 + H 4	C 4
C 3 + H 4 + C 4 + H 5	C 5
...	..
C 8 + H 9 + C 9 + H 10	C 10

Table 4: Data distribution for a context of 4 turns

Context	Response
H 1 + C 1 + H 2 + C 2 + H 3	C 3
H 2 + C 2 + H 3 + C 3 + H 4	C 4
H 3 + C 3 + H 4 + C 4 + H 5	C 5
...	..
H 8 + C 8 + H 9 + C 9 + H 10	C 10

Table 5: Data distribution for a context of 5 turns

A.2 Final prompt for section 5.2.1

“You are a human score annotator for dialogues. You are going to be given a dialogue between a human and a chatbot. Your task is to evaluate the overall quality of the dialogue with a score between 1 and 5.

1. A score of 1 means there is a lack of coherence, engagement, and relevance, with robotic responses and disjointed flow making interactions frustrating.

2. A score of 2 means there’s poor flow, repetitive questions, and lack of focus, with some minor engagement attempts but overall monotonous and irrelevant responses.

3. A score of 3 means there’s basic coherence but struggle with abrupt topic shifts, moderate engagement, and occasional relevance, leading to an average experience.

4. A score of 4 means it’s coherent and engaging with a friendly tone and relevant topics, but some minor repetitiveness and occasional abrupt shifts are noted.

5. A score of 5 means it’s smooth, engaging, and with a friendly tone and natural flow, though some errors like repetition or excessive detail.

Do not be afraid to assign a score of 5 if the conversation goes well. You should provide your output with two sections: REASONING with the motives you assign the score that you do and SCORE with the assigned score only."

A.3 LLM finetuning results 5.2.2

Learning rate	$5E-06$
Epochs	4
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.1
Average Pearson Correlation	**0.257 ± 0.026**

Table 6: Best parameters obtained for LoRA finetuning and Average Pearson Correlation

A.4 Final prompts for flag extraction for section 5.2.3

The prompts used for each flag can be seen in Table 7.

A.5 Individual regression experiments

Model	Hyperparameters	Average Pearson Correlation
SVM	’kernel’:rbf ’epsilon’: 0.1 ’C’: 10’	0.243 ±0.013
XGBoost	’colsample_bytree’: 0.8 ’gamma’: 0.3 ’learning_rate’: 0.2 ’max_depth’: 2 ’min_child_weight’: 0.3 ’n_estimators’: 100	**0.245 ± 0.013**

Table 8: Average Pearson Correlation for models SVM and XGBoost and the best hyper-parameters obtained using Flag features

Model	Hyperparameters	Average Pearson Correlation
SVM	’kernel’: poly ’degree’: 2 ’epsilon’: 0.001 ’C’: 10	0.334 ± 0.012
XGBoost	colsample_bytree’: 0.8 ’gamma’: 0.2 ’learning_rate’: 0.2 ’max_depth’: 2 ’min_child_weight’: 0.3 ’n_estimators’: 50	**0.339 ± 0.009**

Table 9: Average Pearson Correlation for models SVM and XGBoost and the best hyper-parameters obtained using PoE predictions and Flag features.

Flag	Prompt	Occurrence
confusion	[...]Your task is to check if the human is aware that is talking to a chatbot and understands the situation. If the human is aware, you should assign a 0 and if the human is not aware you should assign a 1. [...]	29.89 %
angry	[...] Your task is to check if the human is angry during the conversation, giving harsh and insulting responses to the chatbot. If the human is angry, you should assign a 1 and if the human is not angry you should assign a 0. [...]	39.36 %
engaging	[...] Your task is to check if the human is engaging in the conversation, answering with long and meaningful responses to the chatbot's interactions or if it's being non-engaging by giving short and bland responses when the chatbot is providing meaningful answers. If the human is engaging, you should assign a 1 and if the human is not engaging you should assign a 0. [...]	30.67 %
loop	[...] Your task is to check if the chatbot stays on the same topic when the human is asking to change it or repeating the same responses during many many turns. If the chatbot is looping, you should assign a 1 and if the chatbot is not looping you should assign a 0. [...]	60.20 %
toxic	[...] Your task is to check if the human is being toxic, saying racist, sexually explicit or homophobic comments. If the human is being toxic, you should assign a 1 and if the human is not being toxic you should assign a 0. [...]	15.66 %
correction	[...] Your task is to check if the human is repeatedly correcting the chatbot along many turns by indicating that it didn't understand the human's response or that the information provided is not true. If the human is correcting, you should assign a 1 and if the human is not correcting you should assign a 0. [...]	61.31 %

Table 7: Prompts used to extract flags from the conversations

Model	Hyperparameters	Average Pearson Correlation
SVM	’kernel’: linear ’epsilon’: 0.1 ’C’: 2	0.301 ± 0.025
XGBoost	colsample_bytree’: 0.8 ’gamma’: 0.2 ’learning_rate’: 0.1 ’max_depth’: 2 ’min_child_weight’: 0.3 ’n_estimators’: 50	**0.304 ± 0.021**

Table 10: Average Pearson Correlation for models SVM and XGBoost and the best hyper-parameters obtained using LLM finetuned predictions and Flag features.

Model	Hyperparameters	Average Pearson Correlation
SVM	’kernel’: linear ’epsilon’: 0.1 ’C’: 500	0.343 ± 0.014
XGBoost	colsample_bytree’: 1.0 ’gamma’: 0 ’learning_rate’: 0.1 ’max_depth’: 2 ’min_child_weight’: 5 ’n_estimators’: 100	**0.346 ± 0.015**

Table 11: Average Pearson Correlation for models SVM and XGBoost and the best hyper-parameters obtained using PoE predictions, LLM finetuned predictions and Flag features.