

Neural Models and Language Model Prompting for the Multidimensional Evaluation of Open-Ended Conversations

Michelle Elizabeth^{*,1,2}, Alicja Kasicka^{*,1}, Natalia Krawczyk^{*,1},
Magalie Ochs², Gwénoél Lecorvé¹, Justyna Gromada¹, Lina M. Rojas-Barahona¹

¹Orange Research ²Aix-Marseille University

michelle.elizabeth@orange.com, alicja.kasicka@orange.com, natalia1.krawczyk@orange.com,

magalie.ochs@lis-lab.fr, gwenole.lecorve@orange.com, justyna.gromada@orange.com, lina.rojas@orange.com

Abstract

The growing number of generative AI-based dialogue systems has made their evaluation a crucial challenge. This paper presents our contribution to this important problem through the Dialogue System Technology Challenge (DSTC-12, Track 1), where we developed models to predict dialogue-level, dimension-specific scores. Given the constraint of using relatively small models (i.e. fewer than 13 billion parameters) our work follows two main strategies: employing Language Models (LMs) as evaluators through prompting, and training encoder-based classification and regression models. Our results show that while LM prompting achieves only modest correlations with human judgments, it still ranks second on the test set, outperformed only by the baseline. The regression and classification models, with significantly fewer parameters, demonstrate high correlation for some dimensions on the validation set. Although their performance decreases on the test set, it is important to note that the test set contains annotations with significantly different score ranges for some of the dimensions with respect to the train and validation sets.

1 Introduction

Real-life dialogues are unpredictable and dynamic, making them difficult to reproduce in static corpora. Consequently, dialogue systems are typically evaluated with either simulated users or real users (Zhu et al., 2022). However, a significant gap exists between these approaches, leading to unrealistic simulations or subjective human evaluations (Cordier et al., 2023; Elizabeth et al., 2025). Despite its subjectivity, human evaluation is preferred. In the seminal framework PARADISE (Walker et al., 1997), subjective metrics, such as user satisfaction, were estimated based on objective metrics through linear regression. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ibarz et al., 2018) utilizes regression models as reward

models to evaluate the output of Language Models (LMs) (Ouyang et al., 2022) for better alignment to human preferences. This may provide a rationale for high correlation between LM and human judgments (Kazi et al., 2024; Gunasekara et al., 2021). These results suggest that regression models can be a promising approach to conversation evaluation.

Track 1 of DSTC-12 “Dialog System Evaluation: Dimensionality, Language, Culture and Safety” (Mendonça et al., 2025) focuses on automatic evaluation of open-domain dialogues for ten dimensions, at the dialogue level. The challenge incorporates widely-used dimensions such as overall quality, *Relevance*, and *Proactivity*, alongside less conventional ones including *Empathy*, *Trust*, and *Skill*. This provides a valuable opportunity to assess the correlations between human judgments and automatic evaluation techniques across each dimension. For this purpose, we present four distinct approaches for dialogue-level evaluation that were submitted to the challenge by our team *ORALIS*.

This work covers three possible representations of the scores and one combination of these approaches:

- i the most straightforward approach, treating scores as real numbers and predicting them through a *regression* task;
- ii treating scores as classes, since they are integers that correspond to categories of evaluation (such as good, average, poor), which leads to training *classifiers*;
- iii treating scores as tokens among others as handled in autoregressive LMs, and thus using *LM prompting* to generate scores.
- iv a final strategy, referred to as the *hybrid* approach, consists of mixing predictions from diverse approaches for various dimensions.

According to the results, none of our systems outperforms the baseline (a prompted *Llama-3.1-8B-Instruct*¹ LM) in terms of average *absolute* correlation on the test set. However, our approaches outperform the baseline on most individual dimensions. The LM prompting system shows better generalization to the test set than other approaches, performing better on some dimensions than on the validation set. The regression and classification models demonstrate strong *positive* correlations with human scores on the validation set but achieve lower *absolute* correlation scores when applied to unseen examples, suggesting overfitting. The classification approach, while ranking lowest overall alongside the hybrid method, excels on six dimensions including *Empathy*, outperforming all other approaches in terms of number of winning dimensions. The hybrid method, which selects the best-performing approaches on the validation set (combining LM prompting and regression while excluding classification), does not generalize well to the test data. These results can also be explained by the fact that there are inconsistencies between the training-validation sets and the test set, especially regarding the score distribution and score ranges as depicted in Figure 2 and Figure 3.

The paper is organized as follows: Section 2 provides a literature review on dialogue evaluation; Section 3 introduces the datasets and dimensions used in our work, while Section 4 details the four implemented evaluators. Finally, Section 5 reports the results of the validation set (as used to develop the evaluators) as well as on the test set (as used to rank the submitted evaluators in the challenge).

2 Related Work

This section discusses evaluation paradigms, recent advances in automatic and LM-based metrics, current multi-dimensional frameworks and open challenges.

Open-ended conversational AI systems require multi-dimensional assessment due to the complex nature of dialogue, where multiple valid responses exist for any given context. Key dimensions include *coherence* (the contextual appropriateness and logical consistency of responses (Bao et al., 2021)), *engagement* (sustaining user interest (Venkatesh et al., 2018)), *informativeness* (providing relevant content (Bao et al., 2021)), *specificity* (context-tailored

responses (Harrison et al., 2023)), *consistency* (avoiding contradictions (Bao et al., 2021)), and *factual correctness* (minimizing hallucinations and ensuring accurate information (Bao et al., 2021)). Additional dimensions include *fluency*, *personality*, and *context management* (maintaining memory across multiple turns (Wang et al., 2024)).

Traditional reference-based metrics, e.g. BLEU or ROUGE, show weak correlation with human judgments in open-domain dialogue (Liu et al., 2016; Saleh et al., 2020). While human evaluation remains most reliable (Li et al., 2019; Venkatesh et al., 2018), it is costly, time-consuming, and can suffer from inconsistency (Ji et al., 2022; Smith et al., 2022).

Researchers have developed various automatic metrics to overcome evaluation challenges. Embedding-based metrics capture semantic similarity beyond surface-level lexical overlap but struggle with catching conversational nuances. Regression models, inspired by PARADISE (Walker et al., 1997) and RLHF (Christiano et al., 2017), and classification models are constrained by availability and quality of training data. Reference-free metrics like FED (Mehri and Eskénazi, 2020) evaluate responses in context with better human alignment. LM-based evaluation uses LMs as judges, showing stronger correlation with human ratings (Lin and Chen, 2023; Yu et al., 2024), despite challenges including self-preference bias (Chen et al., 2025) and sensitivity to response characteristics and context complexity (Xu et al., 2025).

The adoption of LMs as judges (Gunasekara et al., 2021; Kazi et al., 2024) enables scalable evaluation, although concerns about annotation quality persist. Small LMs offer cost-effective alternatives and have recently shown strong potential as capable judges. Although they may seem less accurate due to their size, recent work (Deshpande et al., 2024) shows that well-aligned LMs with around 3B parameters can achieve the performance of much larger systems.

Evaluation campaigns like DSTC-9 (Gunasekara et al., 2021) and DSTC-11 (Soltau et al., 2023) have advanced the field through interactive and multilingual evaluation tracks. However, the most successful systems still rely heavily on fine-tuned generative models or LMs for data augmentation, and achieving high correlation with human judgments remains a challenge, especially for multi-dimensional conversation aspects.

Modern evaluation frameworks assess conversa-

¹<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

tion quality across several interdependent dimensions like coherence, engagement, and context management (Bao et al., 2021; Harrison et al., 2023; Wang et al., 2024). Multi-dimensional LM-based approaches (e.g. LLM-Eval (Lin and Chen, 2023), MT-Bench (Bai et al., 2024), KIEval (Yu et al., 2024)) offer thorough assessments yet face challenges with bias, generalizability, and scalability.

Despite significant progress in automatic evaluation of conversational systems, current methods still face limitations in robustness, interpretability, and scalability, highlighting the need for improved multi-dimensional approaches that reliably reflect human perceptions of conversational quality across diverse scenarios.

3 Datasets and Dimensions

We use three datasets in this work: DSTC-12 (Mendonça et al., 2025), the official competition dataset; CONTURE (Gunasekara et al., 2021), which was used in the DSTC-9 evaluation campaign and FED (Mehri and Eskenazi, 2020), another dataset published for dialogue evaluation research.

As detailed later in this section, these datasets predominantly contain open-ended human-machine dialogues annotated by humans on dialogue-level for various evaluation metrics, although the FED dataset also includes human-human dialogues.

3.1 DSTC-12 Dataset and Metrics

DSTC-12 (Mendonça et al., 2025) is an official dataset released as part of the competition. It contains 185 open-domain human-machine dialogues in English. Each dialogue covers a wide variety of everyday topics such as personal stories, preferences and recommendations, and fact-based planning queries. Detailed dataset statistics are displayed in Table 1. What is worth noting is the significant difference between the length of utterances in DSTC-12, compared to other datasets. Still, in all datasets, the average number of words in machine turns is significantly greater than in human utterances.

The dialogues were evaluated by human annotators across ten dimensions. According to the organizers, human annotators were either MTurk workers or lab members, thus different dialogues might have been annotated by different annotators. This combination of research staff and online work-

ers might raise concerns about potential inconsistencies in how dimensions were scored across conversations.

Unfortunately, the annotated data is highly imbalanced, as not all dialogues have scores assigned for each evaluation dimension. While each dialogue received scores for at least four dimensions, the coverage varies considerably. Only 54 out of 185 dialogues (29%) are annotated with all ten dimensions, while the majority include annotations for only four or five. At the dimension level, annotation counts are also unevenly distributed: most dimensions are well represented with over 120 annotations, whereas *Overall* appears in less than one-third of the dialogues. These patterns are illustrated in Figure 1. The test set consists of 120 dialogues, with varying coverage of annotations as in the training set.

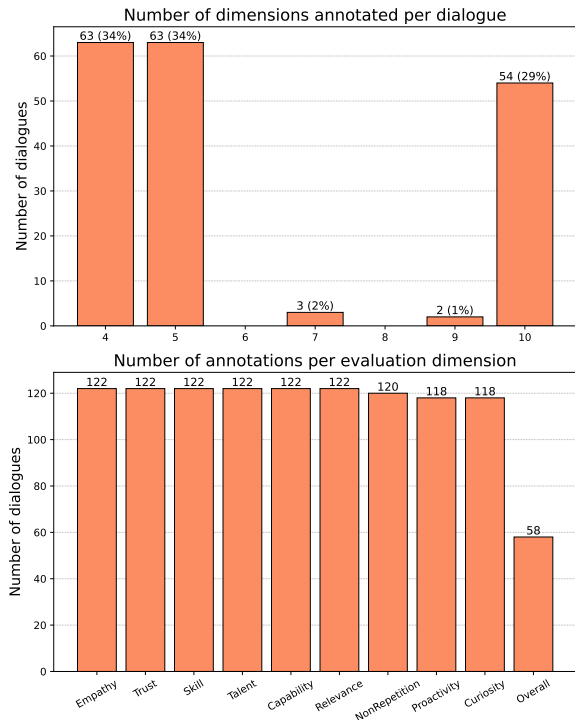


Figure 1: Distribution of the dialogues in the DSTC-12 dataset (train/validation) based on the scores (top), and the dimension (bottom).

In addition, dimensions have different score ranges. Their names, along with their ranges, are: *Empathy* (1-12), *Trust* (0-5), *Curiosity* (0-100), *Proactivity* (0-100), *NonRepetition* (0-100), *Relevance* (0-100), *Overall* (0-100), *Skill* (0-5), *Talent* (0-5), and *Capability* (0-5). For dimensions such as *Relevance*, *NonRepetition*, *Proactivity*, and *Curiosity*, the majority of the human scores are between 6

	DSTC-12		FED	ConTurE
	train	test		
#Dialogues	185	120	125	119
#Ann. per Dialogue	1	1	5	3
Avg. #turns	15	21	6	9
Avg. #words per turn (H)	25	51	6	7
Avg. #words per turn (M)	130	193	12	19

Table 1: Statistics for all the datasets (Ann. stands for annotations, H stands for human and M stands for machine).

and 10, despite the score range being 0-100. The distributions of the scores are uneven, particularly for dimensions with the 0-100 range, see Appendix A, Figure 2. In the test set, the score ranges are between 1-5 for *Skill, Talent, Capability, Trust, and Overall* while the range is 1-10 for *Empathy, Relevance, NonRepetition, Proactivity and Curiosity*, which differs from the score ranges observed in the training set. See Appendix A, Figure 3.

Metrics: The challenge assesses the evaluators based on the mean *absolute* Spearman correlation with human judgments. In our experiments, we also consider the mean *positive* correlation with human judgments.

3.2 FED and ConTurE

The official dataset released for the challenge, i.e. DSTC-12, is rather small, containing only 185 dialogues. Each dialogue was annotated by only one evaluator. To increase data diversity and avoid overfitting when training our regression and classification models, we utilized two other open-domain human-machine dialogue datasets: CONTURE (Gunasekara et al., 2021), with 119 dialogues, which was proposed in DSTC-9 Track 3, and FED (Mehri and Eskenazi, 2020), with 125 dialogues, introduced at SigDial 2020. In all these three datasets, the dialogues predominantly involve interactions between a human and a machine, although the FED dataset also includes human-human interactions with one participant simulating a machine. The conversations cover a variety of everyday topics such as personal preferences, opinions, popular culture, and general knowledge, resembling natural and informal interactions.

In contrast to DSTC-12 dataset, the CONTURE dataset has each dialogue annotated by three different raters, while in FED dataset each dialogue was annotated by five different evaluators. To ensure that every dialogue contributes exactly one score per dimension to our models, just as in the DSTC-

FED & CONTURE		DSTC-12
Inquisitive	→	Curiosity
Avg(Informative, Coherence)	→	Relevance
Topic depth	→	Talent
Flexible	→	Proactivity
Diverse	→	Non-repetition
Likeable	→	Empathy
Consistent	→	Trust
Understanding	→	Capability
Error recovery	→	Skill

Table 2: FED & CONTURE to DSTC-12 mapping.

12 dataset, and to prevent over-representation of multi-rated dialogues, we first averaged the dimension scores in both CONTURE and FED. Then, we mapped the dimension names from the additional datasets to match those in the DSTC-12 dataset. This mapping was based on the heuristics shown in Table 2, where we paired each metric from CONTURE and FED with the DSTC-12 dimension that best reflected its core intent.

In both external datasets, the dimensions are annotated on different scales: for most dimensions, annotations are on a 3-point scale, except for Consistent, which is binary, and Overall quality, which is on a 5-point scale.

After averaging the raw scores per dialogue, we applied a linear rescaling function to fit each dimension into the corresponding dimension range in the DSTC-12 dataset:

$$q = (p - A) \frac{D - C}{B - A} + C \quad (1)$$

where p is the original value in $[A, B]$ and q is the mapped value in $[C, D]$. Finally, we rounded q to the nearest integer to obtain the final score on the target dimension scale in the DSTC-12 dataset. We split the official DSTC-12 dataset equally into train and validation sets for each dimension. The training set was further enhanced by adding the dialogues from the two additional datasets.

4 Evaluators

In this section, we first describe the baseline system (provided by the challenge organizers) and then present the evaluation systems submitted to the challenge. Our first approach is based on the LM-as-a-judge paradigm, namely *LM Prompting*. The next two systems are classic neural models: *regression* and *classification*. Finally, the last system is a hybrid evaluator that combines the *regression* model with the *LM Prompting* system.

4.1 Baseline

This system was proposed by the organizers and it is a fine-tuned *Llama-3.1-8B-Instruct*² pretrained model for content safety classification, prompted for dialogue-level evaluation. For all of the dimensions the same prompt template was used, which included all evaluation dimensions in one prompt.

4.2 LM Prompting

We tested various prompting methods: (i) basic prompt with just the name of the dimension, see example in Appendix C.1; (ii) zero-shot learning with the definition of the dimension, see example in Appendix C.2; (iii) one-shot learning, see example in Appendix C.3; (iv) few-shot learning (with 3 samples), see example in Appendix C.4 and (v) self-consistency prompting, see example in Appendix C.5.

We assigned various roles to the LM (such as "crowd-worker", "expert", or "human evaluator") and provided task descriptions with score ranges at varying levels of detail.

For one-shot and few-shot learning methods, we provided examples with their assigned scores in two formats: either as conversation excerpts or as summarized dialogue. For the few-shot learning, we randomly sampled three dialogues: one with the lowest possible score, one from the median, and one with the highest possible score. For one-shot learning, we randomly sampled a dialogue with a score around the median. In self-consistency prompting, the LM was provided with a short description of the meaning of the scores within the score range, as well as was asked to check the validity of its response and fix it, if needed, before responding.

In every prompting method, the LM was asked to return the score along with a short explanation for the given score. Various versions of the prompts were evaluated on each dimension separately, and different prompts were selected for later study based on these preliminary results (listed in the Appendix B).

We explored several dialogue context strategies to feed the LM: the last 40% of the conversation, the first 40% of the conversation, the first 20% and last 20% of the conversation, a summarized version of the dialogue, or the full dialogue. The summarised version was obtained by summaris-

²<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

ing each utterance using *Llama 3.1 8B Instruct* model, as it performs well across variety of tasks (Grattafiori et al., 2024). We tested two different summarisation prompts for sentences whose length exceeded 200 words, in conversations with more than 3000 words in total. These prompts mainly differed by the maximum number of tokens in their summarised versions: either 50 (*summarisation 1*) or 150 (*summarisation 2*). The exemplary prompt can be found in the Appendix, section B.11.

We considered three state-of-the-art LMs: *Deepseek Llama 8B*³, *Deepseek Qwen 7B*⁴, and *Qwen 2.5 7B Instruct 1M*⁵. These models were selected based on their demonstrated effectiveness on multiple natural language processing tasks (Guo et al., 2025; Yang et al., 2025).

We tested various combinations of prompting on the validation set. We modified the prompt, the dialogue context, and utilized distinct LMs. Then, we selected the best performing combination for each dimension, i.e., achieving the highest *positive* correlation values with human annotations. The chosen configuration for each dimension is shown in Table 3.

Analysis of the standard deviation values for correlation results for different models (average std=0.09) and for different dialogue contexts (average std=0.12) implies that the latter, on average, impacts the final correlation result slightly more than the choice of the model.

Systems' performances on the validation set and test set are presented in Table 4 and Table 5, respectively.

4.3 Regression

We trained a regression model for each dimension. The model's architecture consists of a regression layer on top of a ModernBERT Large encoder (Warner et al., 2024). It is worth noting that ModernBERT has a context limit of 8K tokens allowing for encoding a larger dialogue context, in contrast to BERT-family models (Devlin et al., 2019), which are limited to only 512 tokens. ModernBERT is also notably smaller than LMs, with fewer than 1 billion parameters (395 million). We utilized the score ranges provided in the challenge

³<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B>

⁴<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B>

⁵<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M>

Dimension	Prompting	Dialogue part	Language Model
Empathy	zero-shot	last 40%	Qwen 2.5 7B Instruct
Trust	zero-shot	full conversation	Qwen 2.5 7B Instruct
Skill	zero-shot	first 20% + last 20%	Qwen 2.5 7B Instruct
Talent	zero-shot	summarisation 1	Qwen 2.5 7B Instruct
Capability	zero-shot	summarisation 1	Qwen 2.5 7B Instruct
*Capability	zero-shot	summarisation 2	Deepseek Qwen 7B
Relevance	few-shot	summarisation 1	Deepseek Llama 8B
*Relevance	few-shot	first 20% + last 20%	Qwen 2.5 7B Instruct
NonRepetition	few-shot	first 40%	Deepseek Qwen 7B
*NonRepetition	few-shot	first 20% + last 20%	Qwen 2.5 7B Instruct
Proactivity	zero-shot	first 40%	Deepseek Llama 8B
*Proactivity	zero-shot	first 20% + last 20%	Qwen 2.5 7B Instruct
Curiosity	zero-shot	summarisation 2	Deepseek Qwen 7B
*Curiosity	zero-shot	first 40%	Deepseek Llama 8B
Overall	zero-shot	full conversation	Qwen 2.5 7B Instruct

Table 3: LM prompting approach: Chosen methods and models for each evaluation dimension, * refers to the combination that obtained the highest absolute correlation on the validation dataset.

dataset and the mean-square error as the loss function. To generalize better and avoid overfitting, we utilized CONTURE and FED datasets in addition to the DSTC-12 dataset, using the mapping introduced in Section 3.2. Regression performance on the validation and test sets is presented in Table 4 and Table 5, respectively. A later experiment with varying values of weight decay for training, did not show any considerable improvement in the correlations on the test set.

4.4 Classification

Similar to the regression system, we trained individual classifiers for each dimension on our combined training set. All dialogues were encoded using *Sentence-BERT (SBERT)*⁶. Since we require discrete categories, each integer score was rescaled to an integer range using Equation 1 and rounded to the nearest integer.

Model development followed a two-stage grid search. In the first stage, we explored different class ranges and selected [0, 8] based on validation performance. In the second stage, we tuned *Multi-Layer Perceptron (MLP)* hyperparameters separately for each dimension to maximize *positive* validation Spearman correlation. To reduce overfitting, we specifically optimized regularization and training duration balancing convergence

with generalization. The best hyperparameters for each dimension were used to train the final classifiers and predict dimension scores on the test set. Predicted scores were rescaled back into the original ranges, using the inverse of Equation 1. The validation and test performance of our classifiers trained with SBERT encodings are presented in Table 4 and Table 5, respectively.

To further address potential overfitting, we applied ModernBERT encodings as in the regression models, combined with label smoothing. However, these changes resulted in slightly lower test set correlations, suggesting that increased model capacity was not sufficient to improve performance.

4.5 Hybrid

The hybrid system combines methods that performed well on the validation set for each dimension, as shown in Table 4. When selecting these methods, we limited our choice to only regression and LM prompting approaches, excluding the classification method from consideration. This decision was based on the prioritisation of approaches with stronger contextual understanding and generalization capabilities. In the regression system we utilized ModernBERT that has a larger context window, in comparison to the SBERT model used in the classifier. This allows us to process more dialogue context. We strategically chose LM prompting for several dimensions due to its demon-

⁶<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

Dimension	LM prompting	Regression	Classification	Hybrid
Empathy	0.3	0.23	0.35	0.3
Trust	0.38	-0.02	-0.07	0.38
Skill	0.33	-0.09	-0.06	0.33
Talent	0.26	0.41	0.15	0.41
Capability	0.17	-0.21	0.00	0.17
Relevance	0.19	0.79	0.71	0.79
NonRepetition	0.16	0.75	0.68	0.75
Proactivity	0.01	0.79	0.66	0.79
Curiosity	-0.02	0.68	0.65	0.68
Overall	0.4	0.27	0.49	0.4
Abs. Average	0.22	0.42	0.38	0.5

Table 4: Correlation between the gold labels and system’s outputs on the validation set for each system. **Bold** values indicate the highest *absolute* correlation across all systems.

strated ability to generalize well to unseen data (Wang et al., 2023). Additionally, our classification approach returns discrete integer values, e.g., on the scale 0-8, requiring mapping to, e.g., the 0-100 scale, and potentially introducing approximation errors, while both regression and prompting methods produce continuous values within the desired range without the need for additional mapping. This combination of enhanced contextual processing and a potential for better generalization influenced the choice of methods for our hybrid system.

It is worth noting that all three approaches, i.e. LM prompting, regression, and classification, were submitted to the challenge separately.

The regression system was chosen for the following dimensions: *Talent*, *Relevance*, *NonRepetition*, *Proactivity*, and *Curiosity*. For the remaining dimensions, i.e. *Empathy*, *Trust*, *Skill*, *Overall*, and *Capability*, the LM-prompting was chosen as it obtained the most promising results on the validation set, see Table 4. The results of our hybrid system on the test set are shown in Table 5. This system underperformed on this dataset in comparison to its scores on the validation set.

5 Results

The results of our systems on the test set are presented in Table 5, along with the baseline approach published by the DSTC-12 challenge organizers.

The baseline is based on the LM-as-a-judge approach, similar to one of our systems; however, it uses a different LM and different prompt.

The *absolute* average correlation on the test set for all systems is relatively low, between 0.14 and

0.15, while the baseline achieves 0.17. This represents a significant decrease from the validation set, where the regression and hybrid systems achieved values between 0.4 and 0.5 (see Table 4).

None of our systems achieved a higher average *absolute* score than the baseline; however, our approaches outperform the baseline on most of the individual dimensions. The baseline has higher scores only for the *NonRepetition* and *Overall* dimensions. Nevertheless, the difference on the *NonRepetition* dimension is significant enough to influence the absolute average score for the whole system.

Each of our approaches outperforms the baseline on multiple dimensions in terms of the *absolute* score. The classification approach performs best, in terms of number of winning dimensions, exceeding the baseline on six dimensions, while the LM prompting, regression, and hybrid approaches each outperform on five dimensions. All four of our systems outperform the baseline on *Empathy*, *Capability* and *Proactivity*, and three of them excel on *Talent* as well.

Performance patterns vary across dimensions. The classification approach maintains its strength for *Empathy* from validation to test set in terms of *absolute* correlation, though with reduced values. For *Talent* and *Capability*, the regression system outclasses other approaches across both sets. However, some dimensions show inconsistent results, for example, LM prompting excels on *Trust* on the validation set but its performance drops significantly on the test set. On the test set, the regression system shows the opposite trend for this dimension,

Dimension	LM prompting	Regression	Classification	Hybrid	Baseline
Empathy	-0.08	0.17	-0.17	-0.08	0.06
Trust	0.01	0.2	0.13	0.01	-0.11
Skill	-0.22	0.07	-0.02	-0.22	-0.1
Talent	0.05	0.24	0.22	0.24	0.1
Capability	0.13	0.24	0.12	0.13	0.07
Relevance	0.08	-0.1	-0.28	-0.1	0.23
NonRepetition	0.11	0.14	-0.0	0.14	0.39
Proactivity	-0.15	0.08	0.2	0.08	-0.02
Curiosity	0.37	0.09	0.08	0.09	0.23
Overall	0.31	0.13	-0.17	0.31	0.38
Abs. Average	0.15	0.15	0.14	0.14	0.17

Table 5: Correlation between the gold labels and systems’ outputs on the test set. **Bold** values indicate the highest *absolute* correlation across all systems.

performing better than on the validation set.

We observe significant performance decrease between validation and test sets for several dimensions for regression and classification systems, suggesting potential overfitting. The regression system shows drastic decreases for *Relevance*, *NonRepetition*, *Proactivity*, and *Curiosity*, despite achieving correlations of 0.68-0.79 on the validation set. The classification system demonstrates similar patterns on the same dimensions, with correlations of 0.65-0.71 on the validation set. Nevertheless, it maintains superior performance for *Relevance* on the test set.

Interestingly, LM prompting demonstrates the opposite pattern for some dimensions, performing better on the test set than on the validation set. It achieved the highest *absolute* correlations on test set for *Proactivity*, *Curiosity*, and the *Overall* dimension, despite weaker results on the validation set.

Inspecting both Table 4 and Table 5 raises concerns about why some dimensions show *negative* correlation values. One possible explanation lies in the conceptual mismatch between how LMs and humans interpret evaluation metrics. The inconsistent score ranges between the training and test set also leads us to question the quality of the annotations. Dimensions may have been understood differently by annotators and models, leading to inconsistent judgments that weakened or even inverted expected correlations. Evaluation systems often reflect individual user experiences shaped by emotion and subjectivity, making consistent human assessment especially difficult (Fan and Luo, 2020).

Furthermore, scoring chatbot responses remains a fundamentally subjective and challenging task even for human evaluators, which increases the likelihood of annotation noise in human labels (Yuwono et al., 2019).

6 Conclusions and Future Work

In this paper, we present four distinct dialogue-level evaluators for different dimensions that were submitted to the DSTC-12 challenge. We explored distinct prompting strategies, including varying the dialogue context across different LMs. We also trained very small regression and classification models on the challenge dataset enriched with other evaluation datasets (CONTURE and FED). We also considered a hybrid system that combines the LM prompting and regression approaches. Furthermore, we analyzed the data and found that there are inconsistencies between the training-validation sets and the test set, in terms of the score distribution and score ranges. Although our systems did not outperform the baseline, classical approaches, such as regression and classification, show interesting results, competitive with larger models of 7 and 8 billion parameters used in LM prompting approach.

In terms of future work, we first suggest enhancing the quality of the dataset in a dedicated annotation campaign. Second, we would like to explore domain adaptation techniques for training models on similar but larger datasets from distinct sources (such as the *DSTC-11* dataset) to overcome data scarcity.

7 Limitations

The scaling laws have shown that the impressive capabilities of LLMs are highly influenced by three factors: the size of the model, the size of the dataset, and the amount of computing power used for training (Kaplan et al., 2020). All LMs used in our experiments have fewer than 13B parameters. The regression and classification models have fewer than 1B parameters.

We tuned our systems to maximize the *positive* correlation, however the systems were ranked based on the *absolute* correlation.

Moreover, the dataset provided in the challenge is quite small, making it difficult to use for training regression and classification models. The mapping we made between the annotation of the additional datasets and the DSTC-12 dataset is entirely subjective, which may require in depth investigation to study the impact of various mappings. It would have been beneficial to have the instructions provided to human annotators for a more accurate mapping as well as to define the dimensions more accurately for the LM prompting method.

Finally, there are concerns regarding the consistency of the annotations for certain dimensions, since their score ranges vary significantly between the train-validation and the test sets.

References

- Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *Annual Meeting of the Association for Computational Linguistics*.
- Siqi Bao, H. He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zhengyu Niu. 2021. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *AAACL/IJCNLP*.
- Wei-Lin Chen, Zhepei Wei, Xinyu Zhu, Shi Feng, and Yu Meng. 2025. Do llm evaluators prefer themselves for a reason? *arXiv preprint arXiv:2504.03846*.
- Paul Francis Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). *ArXiv*, abs/1706.03741.
- Thibault Cordier, Tanguy Urvoy, Fabrice Lefèvre, and Lina M. Rojas Barahona. 2023. Few-shot structured policy learning for multi-domain and multi-task dialogues. In *EACL*.
- Darshan Deshpande, Selvan Sunitha Ravi, Sky CH-Wang, Bartosz Mielczarek, Anand Kannappan, and Rebecca Qian. 2024. Glider: Grading llm interactions and decisions using explainable ranking. *arXiv preprint arXiv:2412.14140*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Michelle Elizabeth, Morgan Veyret, Miguel Couceiro, Ondrej Dusek, and Lina M. Rojas-Barahona. 2025. [Exploring react prompting for task-oriented dialogue: Insights and shortcomings](#). In *IWSDS*.
- Yifan Fan and Xudong Luo. 2020. A survey of dialogue system evaluation. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (IC-TAI)*, pages 1202–1209. IEEE.
- 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*.
- Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D’Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, and 1 others. 2021. [Overview of the ninth dialog system technology challenge: Dstc9](#). *Proceedings of the 9th Dialog System Technology Challenge Workshop in AAAI2021*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Vrindavan Harrison, Rishi Rajasekaran, and M. Walker. 2023. A transformer-based response evaluator for open-domain spoken conversation. *arXiv.org*.
- Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. [Reward learning from human preferences and demonstrations in atari](#). *ArXiv*, abs/1811.06521.
- Tianbo Ji, Yvette Graham, Gareth J. F. Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. *Annual Meeting of the Association for Computational Linguistics*.

- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- Taaha Kazi, Ruiliang Lyu, Sizhe Zhou, Dilek Hakkani-Tür, and Gokhan Tur. 2024. Large language models as user-agents for evaluating task-oriented-dialogue systems. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 913–920. IEEE.
- Margaret Li, J. Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv.org*.
- Yen-Ting Lin and Yun-Nung (Vivian) Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *NLP4CONVAI*.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael 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. *Conference on Empirical Methods in Natural Language Processing*.
- Shikib Mehri and Maxine Eskenazi. 2020. [Unsupervised evaluation of interactive dialog with DialogPT](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.
- Shikib Mehri and M. Eskénazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. *SIGDIAL Conferences*.
- John Mendonça, Lining Zhang, Rahul Mallidi, Luis Fernando D’Haro, and João Sedoc. 2025. Overview of dialog system evaluation track: Dimensionality, language, culture and safety at dstc 12. In *DSTC12: The Twelfth Dialog System Technology Challenge*, 26th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Avignon, France.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Abdelrhman Saleh, Tovly Deutsch, Stephen Casper, Yonatan Belinkov, and Stuart M. Shieber. 2020. Probing neural dialog models for conversational understanding. *NLP4CONVAI*.
- Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and J. Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *NLP4CONVAI*.
- Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Wei Han, and Yuan Cao. 2023. [DSTC-11: Speech aware task-oriented dialog modeling track](#). In *Proceedings of the 11th Dialog System Technology Challenge (DSTC-11)*, pages 226–234, Prague, Czech Republic. Association for Computational Linguistics.
- Anu Venkatesh, Chandra Khatri, A. Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, R. Prasad, Ming Cheng, Behnam Hedayatnia, A. Metallinou, Rahul Goel, Shaohua Yang, and A. Raju. 2018. On evaluating and comparing conversational agents. *arXiv.org*.
- Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. [Paradise: a framework for evaluating spoken dialogue agents](#). In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL ’98/EACL ’98*, page 271–280, USA. Association for Computational Linguistics.
- Jun Wang, Jiamu Zhou, Muning Wen, Xiaoyun Mo, Haoyu Zhang, Qiqiang Lin, Cheng Jin, Xihuai Wang, Weinan Zhang, and Qiuying Peng. 2024. Hammer-bench: Fine-grained function-calling evaluation in real mobile device scenarios.
- Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.
- Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. 2025. Does context matter? contextual-judgebench for evaluating llm-based judges in contextual settings. *arXiv preprint arXiv:2503.15620*.
- An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.
- Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. 2024. Kieval: A knowledge-grounded interactive evaluation framework for large language models. *Annual Meeting of the Association for Computational Linguistics*.

Steven Kester Yuwono, Biao Wu, and Luis Fernando D’Haro. 2019. Automated scoring of chatbot responses in conversational dialogue. In *9th International Workshop on Spoken Dialogue System Technology*, pages 357–369. Springer.

Qi Zhu, Christian Geishauer, Hsien chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. 2022. [Convlab-3: A flexible dialogue system toolkit based on a unified data format](#). *arXiv preprint arXiv:2211.17148*.

A Additional figures

We provide the distribution of human annotations by dimension for both datasets provided by the organizers: the training set in Figure 2, and test set in Figure 3.

B Selected Prompts

In this section we present the selected prompts.

B.1 Relevance

You are an expert evaluator tasked with assessing the relevance of chatbot’s answers.

Relevance refers to the system’s ability to provide answers that are related or useful to what is happening or being talked about.

Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot’s answers are often irrelevant, and 100 suggests that the chatbot’s answers are always relevant.

The final output should include the score (0-100) and your explanation for the given score.

Here are the examples of the excerpts of the conversations and the score these conversations received. Chatbot’s and user’s utterances are separated using ”;”.

Excerpt from the example conversation: ”{excerpt1}”

Score for the example conversation: ”{score1}”
(...)

The conversation for evaluation:
{conversation}

B.2 Proactivity

Act like a human evaluator tasked with assessing the proactivity of chatbot queries.

Proactivity refers to the system’s ability to anticipate user’s future problems, needs, and changes. A proactive chatbot often takes

initiative and guides the conversation.

Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot is not proactive at all, and 100 suggests that the chatbot often takes initiative and anticipates the needs of the user.

The final output should include the score (0-100) and your explanation for the given score.

The conversation for evaluation:
{conversation}

B.3 NonRepetition

Act like a human evaluator tasked with assessing the chatbot’s ability to avoid repeating responses within a conversation.

Non-repetition refers to the system’s ability to avoid repeating information or questions the user has already provided. A chatbot with strong non-repetition capabilities ensures a smoother conversation by recognising and adapting to previously shared inputs.

Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot often repeats itself, and 100 suggests that the chatbot has strong non-repetition capabilities.

The final output should include the score (0-100) and your explanation for the given score.

Here are the examples of the summaries of the conversations (you will be evaluating a full conversation, not the summary) and the score these conversations received.

Summary of the example conversation: ”{summary1}”

Score for the example conversation: ”{score1}”
(...)

The conversation for evaluation:
{conversation}

B.4 Trust

You are an expert evaluator tasked with assessing how trustworthy the chatbot seems to the user. Trustworthy chatbot is a chatbot that seems sincere, reliable, and honest, whose responses seem true and not harmful or intended to trick the user.

The final output should include the score (from the range 0-5) and your explanation for the given score.

The conversation for evaluation:



Figure 2: Distribution of human annotations by evaluation dimension in the training set.

{conversation}

B.5 Skill

You are a human evaluator tasked with assessing the *skill* of the chatbot in this dialogue.

Skill means how well the chatbot executes the task or responds to the user's input. Consider how accurate, clear, and appropriate the responses

are.

Give a score between 0 and 5, and provide a short explanation for your score.

Dialogue:

{conversation}



Figure 3: Distribution of human annotations by evaluation dimension in the test set.

B.6 Capability

You are a human evaluator tasked with assessing the capability of responses.

Evaluate only capability (how effectively the chatbot fulfils user needs and achieves the purpose of the conversation). Do not assess any other dimension. Focus only on whether the chatbot meets or exceeds the user's expectations.

Give a score between 0-5 and a brief explanation for your score.

Dialogue to evaluate:

{conversation}

B.7 Empathy

You are an expert evaluator tasked with assessing the level of empathy of the chatbot in the conversation. Chatbot that displays high levels of empathy is the one that shows understanding, awareness, sensitivity to the feelings, thoughts, and experience of the user.

The final output should include the score (from the range 1-12) and your explanation for the given score.

The conversation for evaluation:

{conversation}

B.8 Curiosity

You are an expert evaluator tasked with assessing the curiosity of the chatbot in the conversation. Curiosity refers to how well the chatbot engages the user and shows interest in the responses by asking questions encouraging further interactions.

The final output should include the score (from the range 0-100) and your explanation for the given score.

The conversation for evaluation:

{conversation}

B.9 Talent

You are a crowdworker asked to rate the chatbot's *talent* in this conversation.

Talent means how naturally or intelligently the chatbot handles the conversation.

Was it thoughtful, clever, or showed any spark of conversational ability? Use your instinct- if it felt smart or interesting, that's talent.

Give a score from 0 to 5 and a short reason for your choice.

Dialogue:

{conversation}

B.10 Overall

Evaluate the following conversation between a user and a chatbot. The evaluation should be for the responses generated by the chatbot.

Give an integer score the scale of 0-100 to evaluate the overall impression, where 0 indicates the worst score possible and 100 indicates the best score possible.

The final answer must contain an integer in the range 0-100 and the reason for giving the score.

Here is the conversation to evaluate:

{conversation}

B.11 Summarisation prompt

Prompt:

You are an expert copywriter tasked with shortening a chatbot's utterances from a conversation between a chatbot and a user.

Objective:

Shorten the chatbot's response while preserving its original communication style and all relevant details necessary for later evaluation. Ensure that the short version remains faithful to the chatbot's intent, tone, and structure.

Guidelines:

- Retain all details that could be useful for evaluating the chatbot's performance.
- Encode proper names that are irrelevant to the evaluation (e.g., specific phone models) using placeholders like [model-name1].
- Return the shortened dialogue as a string.
- The summary must not exceed 50 words.

Chatbot's utterance to shorten:

{conversation}

Output: A concise yet comprehensive concise version of the chatbot's response (max 50 words).

C LM Prompts examples

In this section we present some outputs of the distinct prompt strategies.

C.1 Basic prompt example

Act like a human evaluator tasked with assessing the relevance of chatbot's answers. Assess only the chatbot, not the user. The final output should include the score (from the range 0-100) and your explanation for the given score.

The conversation for evaluation:

{conversation}

C.2 0-shot learning example

Act like a human evaluator tasked with assessing the relevance of chatbot's answers.

Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.

Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot's answers are often irrelevant, and 100 suggests that the chatbot's answers are always relevant.

The final output should include the score (0-100) and your explanation for the given score.

The conversation for evaluation:

```
{conversation}
```

C.3 1-shot learning example

You are an expert evaluator tasked with assessing the relevance of chatbot's answers.

Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.

Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot's answers are often irrelevant, and 100 suggests that the chatbot's answers are always relevant.

The final output should include the score (0-100) and your explanation for the given score.

Here is an example excerpt of the conversation and the score this conversation received. Chatbot's and user's utterances are separated using ";"

Excerpt from the example conversation: ""{excerpt}""

Score for the example conversation: ""{score}""

The conversation for evaluation:

```
{conversation}
```

C.4 Few-shots learning example

Act like a human evaluator tasked with assessing the relevance of chatbot's answers.

Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.

Please, evaluate queries of the chatbot in the following conversation by assigning it a score from the scale 0-100, where 0 means that the chatbot's answers are often irrelevant, and 100 suggests that the chatbot's answers are always relevant.

The final output should include the score (0-100) and your explanation for the given score.

Here are the examples of the excerpts of the conversations and the score these conversations received. Chatbot's and user's utterances are separated using ";"

Excerpt from the example conversation: ""{excerpt1}""

Score for the example conversation: ""{score1}""

Excerpt from the second example conversation: ""{excerpt2}""

Score for the second example conversation: ""{score2}""

Excerpt from the third example conversation: ""{excerpt3}""

Score for the third example conversation: ""{score3}""

The conversation for evaluation:

```
{conversation}
```

C.5 Self-consistency prompting example

Act like a human evaluator tasked with assessing the relevance of chatbot's answers. Assess only the chatbot, not the user.

Relevance refers to the system's ability to provide answers that are related or useful to what is happening or being talked about.

Rate the chatbot's relevance on a scale from 0 to 100, where:

- 0-20: Very low relevance - The chatbot's responses are mostly irrelevant or off-topic. Users may find the answers confusing or unhelpful.
- 21-40: Low relevance - The chatbot provides some relevant information, but many responses are not aligned with the user's queries. Users may struggle to find useful insights.
- 41-60: Moderate relevance - The chatbot's answers are somewhat relevant, with a mix of useful and irrelevant information. Users may find some value but will likely encounter inconsistencies.
- 61-80: High relevance - The chatbot generally provides relevant and useful answers. Most responses align well with user queries, though occasional irrelevant information may still appear.
- 81-100: Very high relevance - The chatbot consistently delivers highly relevant and useful responses. Users can rely on the answers to be directly related to their queries, enhancing their experience significantly.

Return the score (0-100) along with a concise explanation of why the chatbot received that score.

Think like a domain expert and check the validity of your score. Fix the score if needed.

Dialogue for Evaluation:

{conversation}