SODA-EVAL: Open-Domain Dialogue Evaluation in the age of LLMs

John Mendonça[1,2], Isabel Trancoso[1,2] and **Alon Lavie[3,4]**

[1] INESC-ID, Lisbon
[2] Instituto Superior Técnico, University of Lisbon
[3] Carnegie Mellon University, Pittsburgh
[4] Phrase, Pittsburgh

{john.mendonca, isabel.trancoso}@inesc-id.pt, alavie@cs.cmu.edu

Abstract

Although human evaluation remains the gold standard for open-domain dialogue evaluation, the growing popularity of automated evaluation using Large Language Models (LLMs) has also extended to dialogue. However, most frameworks leverage benchmarks that assess older chatbots on aspects such as fluency and relevance, which are not reflective of the challenges associated with contemporary models. In fact, a qualitative analysis on SODA (Kim et al., 2023), a GPT-3.5 generated dialogue dataset, suggests that current chatbots may exhibit several recurring issues related to coherence and commonsense knowledge, but generally produce highly fluent and relevant responses.

Noting the aforementioned limitations, this paper introduces SODA-EVAL[1], an annotated dataset based on SODA that covers over 120K turn-level assessments across 10K dialogues, where the annotations were generated by GPT-4. Using SODA-EVAL as a benchmark, we then study the performance of several open-access instruction-tuned LLMs, finding that dialogue evaluation remains challenging. Fine-tuning these models improves performance over few-shot inferences, both in terms of correlation and explanation.

1 Introduction

Current evaluation practices of dialogue systems are still highly dependent on human evaluation. Whilst evaluation using strong LLMs has increased in popularity, human evaluation remains the gold standard since it offers a direct view towards human preferences and satisfaction in the interaction, which is the major goal of open domain dialogue. Leaderboard platforms such as ChatBot Arena (Chiang et al., 2024) rank models by calculating an Elo rating obtained from pairwise comparisons of chatbot responses. However, DA (direct

---

> **Yasmine:** I've been feeling really anxious lately about my job. I'm not sure if I can find a new job that pays as much as my current job.
>
> **Wife:** Why are you feeling anxious about your job?
>
> **Yasmine:** I'm just not sure if I have the skillset to find a new job that pays as much as my current one. My current job is pretty secure, but it doesn't pay that well. And with the economy the way it is, I'm just not sure if I can find a new job that pays as much as my current one.

Issues:

- Coherence: Yasmine contradicts herself by first saying she's not sure she can find a job that pays as much, then saying her current job doesn't pay well.

- Repetition: The response includes unnecessary repetition about finding a new job that pays as much as the current one.

Overall Assessment: 2. The response contains contradictions and repetition, affecting coherence.

Table 1: Example dialogue extracted from the LLM-generated SODA dataset, with corresponding issue detection using our taxonomy and quality annotation provided by GPT-4 (§4). This annotated dataset is used to train our dialogue response evaluators (§5).

---

assessments) of responses provide a more granular evaluation of response quality that pairwise comparisons lack, especially when comparing models that differ only slightly in quality but are otherwise similar (Smith et al., 2022).

Despite the potential benefits of direct assessments for open-domain dialogues, the evaluation community is constrained to using a limited number of benchmark datasets, many of which have become outdated (Mendonça et al., 2024). For instance, FED (Mehri and Eskenazi, 2020a), a typically used benchmark in dialogue evaluation, annotates responses generated by arguably obsolete chatbots, and targets quality aspects such as fluency or relevance. While the annotation of these quality aspects may have been of interest at the time, it is not clear if contemporary chatbots still suffer from

---

[1]github.com/johndmendonca/soda_eval.

these issues.

In this work we conduct a qualitative analysis of the dialogues that constitute the SODA dataset (Kim et al., 2023). SODA contains dialogues distilled from GPT-3.5 (Ouyang et al., 2022), which allows us to better understand the limitations of dialogue generation for this model. Our findings suggest that most issues pertain to a lack of coherence, commonsense knowledge, and repetitions. In contrast, generation is almost always fluent and relevant to the prior context, thus confirming newer models have mostly achieved human level fluency and relevance.

Several authors have proposed automated evaluation frameworks to scale such an analysis and/or complement human evaluation. With the introduction of LLMs for this task, many studies have surfaced, ranging from direct assessment of responses and dialogues (Liu et al., 2023; Lin and Chen, 2023) to a more comprehensive analysis (Finch et al., 2023b). However, we point out two limitations within current work: firstly, given the complexity of the task, most studies leverage GPT-4 (OpenAI, 2024), which is known to perform as well as human annotators in many tasks (He et al., 2024). However, such models have downsides with respect to accessibility. Secondly, since the development of frameworks that use these models are mostly informed by quality aspects used in older benchmarks, their performance when evaluating contemporary chatbots remains an open question.

Given these limitations, we conduct a large scale dialogue quality annotation based on the SODA dataset. Our annotations, which we call SODA-EVAL, include over 120 thousand turn level assessments covering 10 thousand dialogues. These annotations are conducted by GPT-4, and target a diverse range of quality aspects, as illustrated in Table 1. Human validation and annotation tasks confirm the quality of our automated annotation, both in terms of issue detection and overall assessment. Additionally, we confirm many of the trends found in our qualitative analysis, namely that the majority of responses are fluent, but some contain coherence and commonsense issues that degrade the quality of the interaction.

With SODA-EVAL as a benchmark, we conduct a study on the performance of several open-access instruction-tuned LLMs as dialogue evaluators, and show that the evaluation of stronger chatbots is a challenging task. Utilizing SODA-EVAL, we also experiment with finetuning these models, and

demonstrate an improvement in performance, both in terms of their correlation with GPT-4 assessments and of the validity of their explanations. Furthermore, we also assess the impact of finetuning the models by evaluating the resulting models on out-of-domain datasets, where we also observe improved correlation performance. This indicates the models' adaptability to different evaluation guidelines and a diverse set of dialogue responses.

Overall, our contributions are as follows:

- We conduct a qualitative analysis of responses in SODA which highlights consistent issues w.r.t coherence and commonsense knowledge, but are generally fluent and relevant.

- We curate a novel dialogue evaluation benchmark called SODA-EVAL, containing over 120k turn-level assessments obtained by GPT-4, targeting various quality aspects and validated by human annotators.

- We evaluate the performance of several open-access instruction-tuned LLMs as dialogue evaluators using SODA-EVAL, demonstrating that finetuning these models improves their performance.

2 Related Work

Evaluation Taxonomies Higashinaka et al. (2021) developed an integrated taxonomy of errors (combining both theory- and data-driven taxonomies). These 17 errors cover surface level, contextual level, and society level errors with responses. Finch et al. (2023a) identified a set of 16 binary behaviour labels, which were refined down to 10 labels after conducting an evaluation study of four (at the time state-of-the-art) chatbots. Our work takes inspiration from these studies and, informed by our own qualitative analysis and pilot studies with GPT-4, proposes a refined set of these labels tailored specifically for SODA.

Automatic Dataset Generation and Annotation There are several studies that propose augmentation and synthetic data generation approaches to scale dataset sizes that target commonsense reasoning (Ye et al., 2022; Bhagavatula et al., 2023; Wang et al., 2023), summarisation (Jung et al., 2024), dialogues (Chen et al., 2023; Kim et al., 2023), and evaluation (Perez et al., 2022; Hartvigsen et al., 2022; Sorensen et al., 2024). The majority of these studies take inspiration from Symbolic Knowledge

Distillation (West et al., 2022), which shows that it is possible to distil knowledge from the textual outputs of large models. With the introduction of LLMs and their reported performance parity with crowdsourcing in many NLP tasks (Veselovsky et al., 2023; Jiao et al., 2023; Cegin et al., 2023), this paradigm has only increased in popularity.

Automatic Dialogue Evaluation Metrics such as BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) remain a popular choice for dialogue evaluation, even though their correlation with human judgements is very low (Liu et al., 2016). Their popularity can be attributed to their ease of use, especially when compared to learned metrics (Mehri and Eskenazi, 2020b; Phy et al., 2020; Sai et al., 2020; Mendonca et al., 2022), which require substantial effort to employ.

With the widespread introduction of LLMs, this limitation has been mostly circumvented. Most recent contributions typically leverage GPT-3.5-Turbo or GPT-4 for dialogue quality assessments (Liu et al., 2023; Lin and Chen, 2023; Mendonça et al., 2023). Of note, Finch et al. (2023b) investigated the ability of GPT-3.5-Turbo to perform evaluation of real human-bot dialogues using the ABC evaluation framework (Finch et al., 2023a). There have also been some efforts to divest from closed-source LLMs, with XDIAL-EVAL (Zhang et al., 2023) probing the evaluation capabilities of several open source LLMs against GPT-3.5-Turbo (Ouyang et al., 2022) and Mendonca et al. (2024) proposing an explainable evaluator of coherence by synthetically generating positive and adversarial negative responses using a closed-source LLM to finetune a smaller, open-source one. However, to the best of our knowledge, our work is the first that conducts direct knowledge distillation of multiple dialogue quality aspects.

3 Qualitative Analysis

3.1 Dialogue Dataset

For the qualitative evaluation of LLM-based generation in the context of dialogues, we focused on two key aspects: firstly, we wanted annotations that were conducted on dialogues where responses are generated by contemporary chatbots. This would allow us to better understand current limitations in chatbots, and, at the same time, ensure our dataset is relevant and not limited to annotating deprecated models. Secondly, the dialogues under study should target open domain scenarios, and not other

Issue	SODA	Finch et al.
None	71%	47%
Coherence	14%	15%
Commonsense	12%	28%
Repetition	8%	7%
Relevance	0%	48%
Fluency	0%	1%
Other	6%	13%

Table 2: Proportion of identified issues resulting from our analysis of SODA vs Finch et al. (2023a). "Other" denotes subjective aspects like engagement and empathy. Additional details on this analysis are given in Appendix A, and a formal definition is given in Table 3.

kinds of interactions conversational agents are typically recruited for, such as conversational QA or task oriented dialogue (Zheng et al., 2024; Zhao et al., 2024).

With these aspects in mind, we opted with selecting SODA (Kim et al., 2023), a large scale dataset with over 1.5 million dialogues distilled from GPT-3.5 (Ouyang et al., 2022). Commonsense knowledge is obtained from triplets (head, tail, relation) from Atomic10x (West et al., 2022), which are used to generate a narrative with GPT-3.5 that informs the final dialogue generation. Human evaluation conducted on SODA shows that its dialogues are more consistent, specific, and natural than DailyDialog (Li et al., 2017), a popular dialog dataset used for the development of evaluation metrics (Yeh et al., 2021).

3.2 Findings

We restrict our analysis to 100 dialogues from the test set of SODA, an effort per annotator in line with other works (Higashinaka et al., 2021; Finch et al., 2023a). We summarise these findings below.

Responses are fluent and take into account prior context. Extensive anecdotal and experimental accounts support the notion that current LLM-based generation is highly fluent. We also identify a similar behaviour in SODA, since we were able to fully understand the vast majority of the responses. In fact, we only found a minor typo in all of the dialogues studied. Equally infrequent were instances where the response under evaluation lacked relevance – the only example of this in our analysis pertained to a hallucination.

Issue	Definition	Higashinaka et al.
Coherence	Contradicts or ignores prior information in the dialogue.	I5-7, I11-14
Commonsense	Lacks common knowledge and logic.	I4, I9, I17
Assumption	Infers information not available in the dialogue context.	-
Repetition	Repeats prior information in the dialogue.	I15
Engagement	Lacks a behaviour or emotion expected from the situation.	I8, I10
Antisocial	Contains unsafe or inappropriate behaviour.	I16
Fluency	Contains typos or other grammatical errors.	I1-3
Gender Pronoun	Goes against normative pronoun.	-
Non-textual	Includes narrative elements or references unexpected inside a turn of a dyadic interaction.	-
Other	Any other issue that affects the quality of the response.	-

Table 3: Proposed taxonomy for SODA-EVAL. **Bold** correspond to the initial set of issues post analysis (§3). We include the error types from Higashinaka et al. (2021) that our taxonomy covers.

Contradictions and lack of commonsense are frequent. Interestingly, the majority of identified errors consist of responses that contradict prior contextual information. In particular, we found an equal amount of self-contradictions and partner-contradictions[2] (e.g., moving out of the country and then saying they would still be close together; example in Table 1). Additionally, we found instances where the model showcases a lack of commonsense knowledge (e.g., spoiling a surprise party).

Comparison with older chatbots. Overall, and as demonstrated in Table 2, the majority of the responses within our analysis (71%) are of good quality when compared to Finch et al. (2023a) (47%), which analysed chatbots dating prior to 2022. Additionally, we find most errors in our analysis relate to contradictions and lack of commonsense knowledge (19%, of which 6% have both issues), whereas Finch et al. (2023a) reports the most frequent issue being relevance (48%). All in all, while the underlying taxonomy may be still applicable to current generation capabilities, the amount and types of errors we encounter are vastly different.

4 SODA-EVAL

Having identified typically found issues in dialogue generation, we move to the curation of a large scale evaluation benchmark dataset based on SODA (CC-BY 4.0). This dataset, which we call SODA-EVAL, contains annotations by GPT-4 (§4.1) and covers over 120 thousand responses of diverse quality (§4.3). Human annotations confirm the validity of this annotation (§4.2). Additional details regarding the development of SODA-EVAL, including

preprocessing, data selection and an in-depth statistical analysis of the annotations are available in Appendix C.

4.1 Generation of Evaluations

For the response assessment, we ask GPT-4 to identify any issues in the response, and then provide an overall assessment of the response. This chain of thought reasoning (explain then rate) allows for a better evaluation of the response, as confirmed by initial experiments on a held-out subset. Our initial set of issues were selected taking into account the analysis in Section 3, together with prior taxonomies. The full prompt used for the generation of evaluations is presented in Table 11.

Non-conforming issues During preliminary experiments, we found that the model included other types of detected "issues" non confirming to our initial taxonomy. One frequent "issue" pertains to a reported mismatch between the name of the participant and the pronoun used during the interaction. This stems from a bias reduction step in the original dataset, which randomly replaced names (likely frequent names with established pronouns) with alternatives from a diverse name set, to increase diversity. Since the pronouns remained unchanged, the model flags them as an issue. However, we note that the preferred pronouns are part of someone's gender expression, and people can have multiple sets of pronouns for themselves.

Another frequently detected "issue" includes what GPT-4 considers to be an unsupported assumption (e.g., the response assumes a romantic relationship not identified before in the dialogue) and non-textual references in the conversation (e.g., the response contains the narration of actions). In

---

[2]Response contradicts or misremembers something the other participant said earlier in the dialogue.

order to ensure our original set of issues are correctly identified, we include additional categories to our taxonomy (as seen in Table 3) which can be then filtered out if deemed necessary.

Post-processing After our large scale generation, we still detected some instances where the identified issues were non-conforming and which we mapped to our taxonomy (Table 3). Additionally, we conducted a check on the issues identified as "Other" and found a large portion of them to be related to antisocial behaviour (e.g., offensive or inappropriate). We initially did not include this issue in our taxonomy since SODA conducted a comprehensive safety filtering and we did not find such cases in our qualitative analysis (§3)[3]. Consequently, we automate the mapping from Other to a new taxonomy class we call "Antisocial" by prompting GPT-4 to classify the explanation as pertaining to antisocial behaviour. With respect to the remaining issues identified as "Other", given the low numbers (142) we manually checked all of them, and mapped them back to our taxonomy if possible. We removed 35% of the issues since they did not, in fact, present any issue, and kept 38% of identified issues in the "Other" category, since they referred to malformed dialogues (e.g., single or 3-speaker dialogues, or role reversal).

Cost The average cost to generate annotations was around $0.13 USD per dialogue or ¢0.94 per response, an amount substantially lower than a human annotator at $0.30 per response, assuming a minimum wage of $15[4].

4.2 Human Validation

In order to confirm the quality of the generations provided by GPT-4, we conducted two human annotation tasks to determine whether the issues detected are correct, and if the overall assessment is in line with human judgements. Additional information regarding these annotations, including guidelines, is available in Appendix D.

4.2.1 Issue Detection

We sampled 200 examples[5] from the test set (each annotator validated 100 examples, and we recruited 3 annotators per example) for validation. To reduce

---

[3]Finch et al. (2023a) also removed "Antisocial" from their final taxonomy.

[4]Estimated workload at 2h per 100 responses evaluated.

[5]This validation sample size is in line with other works (Zeng et al., 2024; Wang et al., 2024).

Aspect	Initial	Revised
IAA	0.7455	0.8601
GPT-4	0.7303	0.8248

Table 4: Inter annotator agreement and correlation results between annotator aggregate score and GPT-4 assessment, before and after GPT-4 assisted revision. All correlations $p < 0.01$.

annotator confusion, we ensured that the examples that contain issues belong to the first flagged responses of the dialogue. This is because we found human annotators had difficulties annotating a response when earlier responses had issues.

We calculate Cohen Kappa scores between annotators and report and average score of (Cohen, 1960) of 0.359, which is a fair to moderate agreement. On average, annotators consider 86% of GPT-4 issue detection as correct, with individual annotators' reported validity ranging between 77% and 91% of responses.

4.2.2 Overall Assessment

For overall assessment, we again sampled 100 examples from the test set and ask annotators to rate the response in a 1-5 Likert scale. We recruited 5 annotators for this task. However, we conducted a two-stage annotation in order to determine the usefulness of having automated assistance during the annotation. In a first step, we ask the annotators to rate the response given only the prior context, which is the same setup as most turn level assessments. In the second step, we provide the same example and the list of GPT-4 detected issues as guidance for annotation. Note that the overall assessment score provided by GPT-4 remains hidden. With this information, annotators are able to revise their rating, or keep it unchanged.

Following Mehri and Eskenazi (2020a,b), we report IAA (Inter Annotator Agreement) results in Table 4, corresponding to the average Spearman Correlation between each individual annotation and the mean of the annotations and the GPT-4 assessment, before and after GPT-4 assisted revision. Here, we note an agreement between annotators in line with other works (within the 0.6 - 0.8 range). Additionally, the agreement between the annotators is similar to the agreement with GPT-4, even before the revised annotation with assistance.

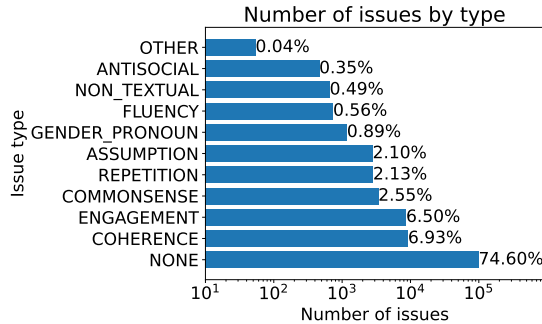With respect to aided revision, we note a large increase in agreement, which suggests that our au-

Figure 1: Number (and percentage) of issues resulting from the annotation for SODA-EVAL.

Figure 2: Turn level score distribution for SODA-EVAL.

tomatic annotation framework could assist in improving issue recall in a crowdsourcing environment. In particular, we found that the majority of revisions were in the direction of a lower score, as all revisions occurred when GPT-4 detected issues. Overall, this resulted in the reduction of distance between the annotation and the GPT-4 assessment.

4.3 Statistics

In total, 10,000 dialogues were automatically annotated, 2/3 of which contain more than 6 turns (with the remainder being 6-turn dialogues). This resulted in 122,648 turn level annotations, which is significantly larger than other evaluation datasets (as reported in Table 5). Furthermore, all previous datasets employ models that no longer accurately reflect current generation capabilities. Additionally, SODA-EVAL is the only dataset that includes natural language explanations for issue detection and the overall assessment of the response.

Dataset	# Examples	Level	Explanation
FED (2020a)	500	Both	✗
USR (2020b)	540	Turn	✗
DSTC10 (2021)	13,944	Both	✗
DSTC11 (2023)	5,116	Both	✗
SODA-EVAL	132,648	Turn	✓

Table 5: Comparison between SODA-EVAL and other typically used evaluation datasets. A detailed description of each dataset is given in Appendix B.

We present the distribution of identified issues in Figure 1. As expected, the majority (74.60%) of responses do not contain any issues. The most frequently identified issues are Coherence (6.93%) and Engagement (6.50%), whereas the least observed ones were Antisocial, Other and Non-textual (<0.5%). For Engagement in particular, the 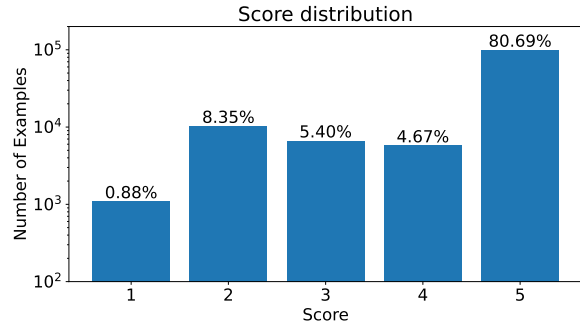vast majority of reported issues pertained to unhelpful and/or non-specific responses (e.g. "ok", or "sure"). Additionally, the turn level score distribution is shown in Figure 2. Similar to the number of issues, the vast majority of responses were rated as high quality (80.69%). The second most frequent score was 2 (8.35%), with scores 3 and 4 having roughly the same amount of responses (around 5%). Overall, our statistical observations share many similarities with the analysis conducted in Section 3, namely in terms of proportion of issue-free responses, and the low amounts of fluency error when compared to coherence and commonsense.

5 Training and Benchmarking of Evaluation Models

As identified in prior sections, the typically employed benchmark datasets focus on the evaluation of responses that do not accurately reflect current generation capabilities. As a result, most evaluator development has been focused on maximising predictive performance of no longer relevant quality aspects, instead of focusing on more complex issues. SODA-EVAL positions itself as both a more realistic training dataset and as an improved evaluation benchmark for open-domain dialogue evaluation. In this section, we examine the performance of several open-access LLMs for the task of overall assessment of responses with explanations.

5.1 Preliminaries

Our reference-free evaluation setup consists of the assessment of a response hypothesis r given a dialogue history (frequently denoted as context) c of varying amount of turns. The goal is to learn a scoring function that assigns a score $f(c, r) \rightarrow s \in [1, 5]$, where 1 indicates minimum quality and 5 maximum quality, and an explanation for the score, which identifies, in natural language, the issues (or lack thereof) in the response.

5.2 Experimental Setup

Training We split SODA-EVAL by dialogues with approximate proportions of 70/20/10 for train, validation and test splits, respectively. The splitting process was conducted such that there is a similar distribution of scores and issues in all subsets. In the end, the dataset split resulted in 85,876/24,535/12,237 responses, per set.

We experimented with several small open-access LLMs since, ideally, we want an evaluator that is lightweight in order to maximise accessibility. As such, we opted with using Flan-T5 (Chung et al., 2022), Qwen1.5-Chat (Bai et al., 2023); Phi-3 instruct (Abdin et al., 2024); LLama-3 instruct (AI@Meta, 2024), the latter two of which were reported to achieve similar performance to that of large LLMs in several benchmarks. Additional training details are given in Appendix E.

Baselines We compare our approach against several models. Firstly, we compare against UNIEVAL (Zhong et al., 2022), a multi-dimensional evaluator that uses T5 as base model and supports reference free evaluation. Additionally, we conduct zero and few shot inference using the respective instruction tuned models we used for training, and GPT-3.5-Turbo (Ouyang et al., 2022), which is also a typically used LLM for evaluation.

Evaluation We split the evaluation suit into two distinct objectives. Firstly, we follow the evaluation literature and evaluate the performance of the models when predicting the judgement for overall quality. For this, we employ Pearson and Spearman correlations. With respect to the quality of the explanations, we calculate the BLEU-4 score of the response compared against the GPT-4-generated explanation, used as a reference.

Additionally, we complement this evaluation with a manual validation of the explanations, where we manually check if the explanation is fluent and acknowledges all the issues in the response. Since this is a human effort, we restrict this assessment to the subset of validated GPT-4 issue detection (§4.2), and report the proportion of correct explanations when taking into account (1) all issues of the taxonomy; and (2) excluding Engagement (which is mostly subjective), both on the full subset or the smaller one with only detected issues. Additional details regarding this annotation are available in Appendix D.

Model	ρ	r	BLEU-4
UNIEVAL (2022)	.1295	.1448	-
Instruction-tuned			
Flan-T5-xl	.0932	.1118	0.00 [a]
Qwen1.5-0.5B	.1002	.1074	7.05
Qwen1.5-1.8B	.1194	.1332	2.91
Qwen1.5-4.B	.2278	.2376	6.30
Phi-3-mini-4k	.2965	.3344	21.56
LLama-3-8B	.3046	.3335	9.79
GPT-3.5-Turbo	.2978	.3418	2.13
SODA-EVAL finetuned			
Flan-T5-xl	.2800	.2800	27.42
Qwen1.5-0.5B	.3636	.3870	29.31
Qwen1.5-1.8B	.4131	.4359	29.90
Qwen1.5-4.B	.4867	.5150	31.10
Phi-3-mini-4k	**.5938**	**.6289**	36.14
LLama-3-8B	.5240	.5628	**40.41**

[a] This model consistently failed to produce explanations.

Table 6: Reported results on SODA-EVAL-TEST. ρ denotes Spearman, r Pearson. All correlation results are $p < 0.01$. **Bold** denotes best overall model. For the instruction-tuned models, we report the best few-shot (0 up to 5) performance.

5.3 In-domain Results

Correlation We report correlation performance on SODA-EVAL in Table 6. When looking at the performance of UNIEVAL and the instruction-tuned models, we see that their assessment is weakly correlated with GPT-4 judgements. This underlines the challenge of evaluating responses from LLM-based chatbots, since most of their issues require some level of reasoning in order to be correctly identified. We also note that Phi-3 and LLama-3 perform about the same as GPT-3.5-Turbo, which is evidence of these model's capabilities despite their smaller size. When finetuning the models on dialogue evaluation data from SODA-EVAL, we note a large increase in correlation, with the best instruction-tuned model, i.e. Phi-3-mini, increasing Pearson correlation from .2844 to .5938. Furthermore, we observe a significant gap in performance in both the instruction tuned and finetuned models when comparing this model to Qwen-4B or Flan-T5-xl, which have about the same number of parameters. This indicates Phi-3 is better equipped to evaluate conversational dynamics, likely due to higher quality instruction data.

Explanation Validity As expected, finetuning yields higher BLEU scores when compared to few-

Model	Full	w/o Engagement
Instruction-tuned		
Phi-3-mini-4k	27% / 49%	62% / 73%
Qwen1.5-4B	28% / 42%	62% / 66%
LLama-3-8B	29% / 51%	64% / 75%
GPT-3.5-Turbo	36% / 49%	68% / 73%
SODA-EVAL finetuned		
Phi-3-mini-4k	**49% / 64%**	**77% / 84%**
Qwen1.5-4B	41% / 59%	68% / 77%
LLama-3-8B	40% / 57%	69% / 78%

Table 7: Explanation validation results for the full taxonomy (**Full**) and without Engagement. For each entry, we present results considering only responses with issues, or all responses (with issues/all).

Model	FED-Turn		DSTC10-TC	
	ρ	r	ρ	r
UNIEVAL	.3229	.2521	.3302	.3249
Instruction-tuned				
Phi-3-mini-4k	.4489	.5276	.3306	.3513
Qwen1.5-4.B	.3189	.3697	.2081	.2242
Llama-3-8B	.5042	.5438	.3077	.3279
GPT-3.5-Turbo	.5599	.5842	.3320	.3481
GPT-4-Turbo	**.5861**	**.6408**	**.4058**	**.4145**
Finetuned				
Phi-3-mini-4k	.4913	.5135	.3550	.3630
Qwen1.5-4.B	.3762	.3874	.2731	.2906
Llama-3-8B	.4496	.4598	.3480	.3597

Table 8: Turn-level correlations of different metrics on Topical-Chat and FED for Overall Quality. ρ denotes Spearman, r Pearson. All correlation results are $p < 0.01$. **Bold** denotes best overall model.

shot prompting, with the weakest model achieving much better scores than all of the instruction tuned models. However, a high BLEU score does not ensure the validity of the explanation. Consequently, we present explanation validation results in Table 7. If we exclude from analysis Engagement as an issue we see that all models produce valid explanations over 60% of the times, even when excluding issue-free responses from consideration. Additionally, we observe that the finetuned models produce more valid explanations than their corresponding instruction-tuned models, across all scenarios. For instance, finetuning Phi-3-mini yields a 15% absolute improvement in the detection of major issues.

When including Engagement in the analysis, we note that the models struggle with identifying engagement issues (as reported by GPT-4), with all open-access models achieving under 30% validity. The exception to this trend is GPT-3.5-Turbo – this can be explained by the fact it belongs to the same family as GPT-4, and as such is likely to be trained using similar data. When finetuning the models, we observe an even higher improvement when compared to the response set without engagement issues (e.g., we report an absolute improvement of 22% validity with Phi-3-mini), which indicates the finetuning step has led the models to better align themselves with the subjective assessments of the teacher model.

In general, all tested models are able to correctly rate good responses (i.e, high recall – which is also evidenced by the increase in performance when including these in the evaluation). Additionally, they are mostly able to correctly identify fluency and non-textual related issues. With respect to the

other issues, however, all models struggle with recall, despite being precise in their identifications. In particular, all models struggled with correctly identifying engagement, coherence and commonsense issues, which highlights the challenge of their identification in open-domain dialogues.

5.4 Out-of-domain Results

To better understand if finetuning on SODA-EVAL data helps improve correlation on out of domain datasets and guidelines, we evaluate our best finetuned models and their corresponding instruction tuned models (0-shot) on FED (Mehri and Eskenazi, 2020a) and DSTC10-TC (Zhang et al., 2021), two typically employed benchmarks for dialogue evaluation. As shown in Table 8, the best performing models are GPT-4 and GPT-3.5-Turbo, followed by Phi-3-mini (finetuned on SODA-EVAL). For our finetuned models in particular, we report consistent increases in correlation when compared to their base models (the exception being LLama-3 on FED). However, the performance gap between all evaluated models is much lower than for SODA-EVAL. We believe this is partially due to the differences between these benchmarks and SODA-EVAL, both in terms of the responses being evaluated (these datasets leverage old dialogue response generators), and the guidelines for evaluation themselves (which paired with guidelines targeting fluency and surface level relevance, may overestimate the quality of the responses).

6 Conclusions

This paper presents SODA-EVAL, a large-scale open-domain dialogue quality annotation dataset targeting the responses provided by a LLM, encompassing over 120 thousand turn-level assessments. This curation was motivated by the pitfalls of current dialogue evaluation, which is limited to assessing responses generated by outdated chatbots and quality aspects. As highlighted in our qualitative analysis of SODA, newer models achieve human-level fluency and relevance, but fall short in areas like coherence and commonsense reasoning. With SODA-EVAL, we conducted a comprehensive study on the performance of open-access instruction-tuned LLMs when evaluating dialogue responses. Our findings show that finetuning these models on SODA-EVAL improves their performance, both in terms of correlation and explanation quality.

7 Limitations

Dialogue Dataset We acknowledge that the selection of SODA as our base dialogue dataset for analysis and annotation comes with some downsides. Firstly, given the dataset is generated from a single source (social commonsense distilled with GPT-3.5), its dialogues may not fully capture the diversity of human-chatbot interactions. Secondly, since the whole interaction is synthetically generated in a single forward pass, it may not accurately represent dyadic conversational dynamics. For instance, we expect a human would point out a major issue in the response directly, instead of continuing the conversation unimpeded (Petrak et al., 2023). However, since we conduct a turn level assessment, this issue is mostly circumvented, at least for the initial turn the issue is presented.

Taxonomy Our developed taxonomy was tailored for SODA. This may limit its extensibility to other dialogue generators, or even dialogue datasets. For instance, one particularity of SODA is that some dialogues simulate physical conversations, which introduces new dimensions to the conversation, such as breaks in time between turns, or the narration of actions required to contextualise the dialogue. For the former, most instances are mostly picked up by GPT-4 with the "non-textual" class, and can therefore be removed from the dataset. Additionally, the dataset is generated by a single LLM, i.e. GPT-3.5. As chatbots improve, some of these issues' representations may be reduced, or even eliminated. Nevertheless, we took steps to mitigate this bias by including other taxonomies in the decision making process. However, we acknowledge some issues may no longer be relevant in the future (a good example would be fluency, which already has only a minor contribution in evaluating quality in dialogues) or new ones may surface. An important future direction would be to develop a framework that is agnostic to model capabilities at the time of development, thus remaining useful.

LLM as Annotator Similar to other identified limitations, the use of a single LLM, i.e. GPT-4, as a replacement for humans, opens up the possibility for several limitations. For starters, we selected this closed-source model based on its state-of-the-art capability. However, we acknowledge, and empirically observed that this model does not garner unanimous agreement with human annotators in terms of the validity of its annotations. We expect newer models to improve in this direction, and remain confident our framework could be adapted to newer, better models. Secondly, these models typically exhibit several evaluation biases, and mostly prefer responses that are helpful and verbose (Wu and Aji, 2023). We mostly observe this behaviour in the engagement detection, where some of the issues could be considered acceptable by humans.

Cultural Diversity Dialogue quality is a diverse, culturally informed concept. For instance, high context cultures (Hall, 1959) privilege non-verbal methods of communication, which is typically not transcribed into text (Nishimura et al., 2008). However, SODA-EVAL is constituted by only English dialogues. As such, it is not clear if the conclusions regarding issue prevalence in contemporary chatbots, nor the evaluators obtained from it, can be extended to other languages and cultures. For low-resource languages, we also expect that chatbots have more frequent issues. We leave this analysis for future work.

8 Ethical Considerations

Safety Safety and harmfulness assessment in dialogue evaluation has mostly been considered a separate topic (in its own right) from other quality aspects (Sun et al., 2022; Hartvigsen et al., 2022). Nevertheless, our taxonomy includes antisocial behaviour as one of its issues, since GPT-4 correctly

identifies it as a critical issue in dialogues. However, we acknowledge that we did not conduct a comprehensive assessment of safety detection – users of our framework are encouraged to complement their response assessment with dedicated safety evaluators.

Annotations Our annotation effort was supported by volunteers from our research lab. All annotators are graduate students or professionals in the field of Linguistics or Language Technologies, most of which non-native but fluent speakers of English. We acknowledge there is possible bias in the assessment of response quality, since all annotators share (to some extent) similar cultural and educational backgrounds. Furthermore, their exposure to generative models is much higher than other focus groups, which may induce confirmation bias.

Acknowledgments

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

AI@Meta. 2024. Llama 3 model card.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2023. I2D2: Inductive knowledge distillation with NeuroLogic and self-imitation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9614–9630, Toronto, Canada. Association for Computational Linguistics.

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

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. Chatgpt to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. *Preprint*, arXiv:2305.12947.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys*, 51(1):1–40.

Sarah E. Finch, James D. Finch, Ali Ahmadvand, Ingyu, Choi, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, and Jinho D. Choi. 2020. Emora: An inquisitive social chatbot who cares for you. *Preprint*, arXiv:2009.04617.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023a. Don't forget your ABC's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071, Toronto, Canada. Association for Computational Linguistics.

Sarah E. Finch, Ellie S. Paek, and Jinho D. Choi. 2023b. Leveraging large language models for automated dialogue analysis. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 202–215, Prague, Czechia. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019*.

Edward T. Hall. 1959. *The silent language*. Doubleday, Garden City, N. Y.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2024. Impossible distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing. *Preprint*, arXiv:2305.16635.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

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

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. 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 Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. Association for Computational Linguistics.

John Mendonça, Alon Lavie, and Isabel Trancoso. 2024. On the benchmarking of LLMs for open-domain dialogue evaluation. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 1–12, Bangkok, Thailand. Association for Computational Linguistics.

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. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 133–143, Prague, Czech Republic. Association for Computational Linguistics.

John Mendonca, Isabel Trancoso, and Alon Lavie. 2024. ECoh: Turn-level coherence evaluation for multilingual dialogues. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 516–532, Kyoto, Japan. Association for Computational Linguistics.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Shoji Nishimura, Anne Nevgi, and Seppo Tella. 2008. Communication style and cultural features in high-/low context communication cultures: A case study of finland, japan and india.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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. *Preprint*, arXiv:2203.02155.

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

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver

Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

Dominic Petrak, Nafise Moosavi, Ye Tian, Nikolai Rozanov, and Iryna Gurevych. 2023. Learning from free-text human feedback – collect new datasets or extend existing ones? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16259–16279, Singapore. Association for Computational Linguistics.

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

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

Mario Rodríguez-Cantelar, Chen Zhang, Chengguang Tang, Ke Shi, Sarik Ghazarian, João Sedoc, Luis Fernando D'Haro, and Alexander I. Rudnicky. 2023. Overview of robust and multilingual automatic evaluation metricsfor open-domain dialogue systems at DSTC 11 track 4. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 260–273, Prague, Czech Republic. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3776–3783. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, W.K.F. Ngan, Spencer Poff, Naman Goyal, Arthur D. Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *ArXiv*, abs/2208.03188.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *Preprint*, arXiv:2306.07899.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *Preprint*, arXiv:2307.03025.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations (ICLR)*.

Chen Zhang, Luis D'Haro, Chengguang Tang, Ke Shi, Guohua Tang, and Haizhou Li. 2023. xDial-eval: A multilingual open-domain dialogue evaluation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5579–5601, Singapore. Association for Computational Linguistics.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *Preprint*, arXiv:2111.02110.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

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

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Dialogue analysis

Finch et al. (2023a) conducted a comprehensive evaluation of four open-domain chatbots using the

ABC-Eval 16 behavior labels. These chatbots are as follows: (1) BlenderBot-DECODE (Nie et al., 2021); (2) BlenderBot2 [6]; (3) BART-FiD-RAG (Shuster et al., 2021); and (4) Emora (Finch et al., 2020).

For the comparison with our analysis on SODA, we map the 16 behavioural labels as follows:

- **Coherence:** Partner Contradiction , Self Contradiction

- **Commonsense:** Incorrect Fact, Commonsense

- **Repetition:** Redundant

- **Relevance:** Topic Switch, Ignore, Irrelevant

- **Fluency:** Uninterpretable

- **Other:** Empathy, Antisocial

B Benchmark Datasets

This section presents a brief survey of datasets that have been used as a benchmark for LLM-based open-domain dialogue evaluation metrics.

The FED dataset (Mehri and Eskenazi, 2020a) consists of turn level and dialogue level annotations of conversations conducted between a Human (40 dialogues) and two chatbot engines (40 dialogues from **Meena** (Adiwardana et al., 2020) and 44 from **Mitsuku**[7]), targeting eighteen quality aspects. Each conversation received one annotation at the dialog level and three annotations at the turn level, randomly selected from the conversation. In total, the FED dataset comprises 3,348 turn-level and 1,364 dialog-level data points, amounting to 4,712 annotations.

In the case of the USR dataset (Mehri and Eskenazi, 2020b), annotations were collected for models trained on the TopicalChat (Gopalakrishnan et al., 2019) and PersonaChat (Zhang et al., 2018) dialogue datasets. Generated responses were obtained from **Transformer** (Vaswani et al., 2017), **RNN Seq2Seq** (Shang et al., 2015), **LSTM** (Hochreiter and Schmidhuber, 1997), and **KV-MemNN** (Miller et al., 2016) models. For each dialog context, an additional human response was also collected. Human annotation was then carried out on sixty dialog contexts, with six responses per context for Topical-Chat (four transformer outputs with different decoding strategies, one newly-annotated human output,

and the original ground-truth response) and five for PersonaChat (Seq2Seq, LSTM, KV-MemNN, one newly-annotated human output, and the original ground-truth response).

The DSTC10 test set (Zhang et al., 2021) was proposed in the context of the "Automatic Evaluation and Moderation of Open-domain Dialogue Systems" shared task, which offered a competitive venue for participants to design robust automatic dialogue evaluation metrics that correlate well with human judgements across multiple dialogue domains, as well as across different quality aspects. For testing, 3 sources of data were used: (1) CHANEL-JSALT2020, (2) ChatEval (Sedoc et al., 2019) and (3) an additional annotation conducted on TopicalChat (Gopalakrishnan et al., 2019) and PersonaChat (Zhang et al., 2018). Eight systems, a human baseline, and a random utterance were used as response generators. Specifically, the eight systems are based on **LSTM Seq2Seq**, **Attention-based LSTM Seq2Seq** (Sutskever et al., 2014), **HRED** (Serban et al., 2016), **VHRED**, **BlenderBot (400M-Distill)** (Roller et al., 2021), **DialoGPT-medium** (Zhang et al., 2020), **T5-base** (Raffel et al., 2020), and **GPT-3** (Brown et al., 2020).

Similar to DSTC10, the the DSTC11's "Robust and Multilingual Automatic Evaluation Metrics for Open-Domain Dialogue Systems" track is split into development and test sets (Rodríguez-Cantelar et al., 2023). For the development set, the organisers provide data from two clusters of datasets from DSTC10 and 4,470 dialogues (approximately 130k turns) open-domain human-human dialogues which are originally in Chinese. Since the goal of the shared task was to evaluate mulitlinguality and robustness of metrics, development data is translated into English, Chinese, Spanish, and back-translated. For testing, the organisers combine a portion of the DSTC10 test set, and include new Human-Chatbot dialogues generated by SotA chatbots. These are: **ChatGPT** (a platform powered by GPT-3.5-Turbo), **GPT-3.5** (Ouyang et al., 2022) and **BlenderBot3** (Shuster et al., 2022). Similar to the development set, the test set was also translated. In total, 4,839 turn level and 277 dialogue level annotations were conducted.

C SODA-EVAL

Tables 9 and 10 present additional examples from SODA-EVAL.

---

[6]BlenderBot 2 model card
[7]Mitsuku blogpost

Ani: Hi! I'm looking for a complete change. I need a new haircut, new clothes, and a new job.

Stylist: Well, you've come to the right place! We can definitely help you with all of that. Let's start with your hair. What kind of style are you looking for?

Ani: I'm not really sure. Something completely different, I guess.

Stylist: Alright, we can definitely do that. How about we start with a shorter style? We can always go longer if you change your mind.

Issues:

- Commonsense: The stylist suggests starting with a shorter style and going longer if the client changes their mind, which is not possible with haircuts.

Overall Assessment: 2. The response contains a logical error regarding haircuts.

Table 9: Example from SODA-EVAL

C.1 Filtering and Selection

Entity Filtering Despite being one of the steps of the SODA post-processing effort, we still found many instances where one or both of the interlocutors are non-human (e.g. animals or objects). As a complementary filtering step, we leverage Word-Net (Miller, 1994) and check if the speaker is a common hypernyms of "person". We note that this step excludes some valid speakers, as Wordnet does not take into account lemmatization nor compound words, and is missing entities such as "agent" or "driver". Nevertheless, this filtering resulted in the exclusion of 4% of dialogues, leaving sufficient dialogues for annotation.

Selection We assume the larger the dialogue, the more likely it may contain issues pertaining to coherence. As such, we focus on annotating the largest dialogues from SODA test. Within dialogues of the same size, we select those that contain the largest amount of words. Additionally, we only conduct the evaluation of responses that contain at least one prior turn of context. The reason behind this is that we found GPT-4 failed (even after explicitly including this instruction in the prompt) to only focus on the detection of surface level errors, which resulted in the severe underestimation of the quality of the response.

Notnamed: Hey, Meriah. Give me the baseball.

Meriah: What? No way! I just got this thing. It's mine!

Notnamed: C'mon, man. I really need it. Just for a little while. Please?

Meriah: No, I don't think so. You can't just take things from people without asking first. That's not cool.

Issues:

- Coherence: Meriah's response is inconsistent because Notnamed did ask for the baseball before Meriah accused them of not asking.

Overall Assessment: 3. The response is mostly coherent, but there's a inconsistency regarding the request for the baseball.

Table 10: Example from SODA-EVAL

C.2 Generation Details

Prompt The prompt used to generate SODA-EVAL is presented in Table 11. In detail, we used the `gpt-4-1106-preview` model which was accessed late March 2024 using the OpenAI API. The decoding temperature was set to 0.3, top_p to 1, and generation was capped to 300 tokens in order to keep explanations succint.
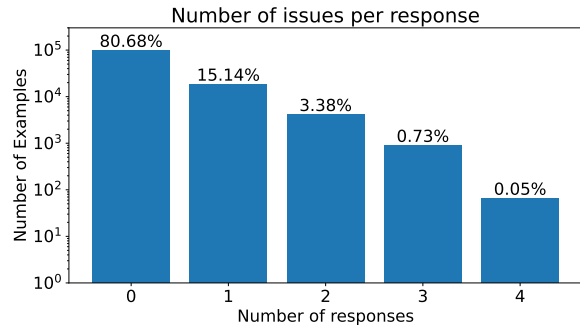

Figure 3: Number (and percentage) of issues per response.

C.3 Additional statistics

In this section, we present additional statistics for SODA-EVAL.

Number of issues per response (Figure 3) Since the majority of responses are of good quality, most responses have 0 issues. As expected, the more frequent the number of issues, the less frequent is such a response. Responses with up to 3 issues can be expected, since responses can have more

Table 11: Dialogue response evaluation instruction template.

Table 12: Inference instruction template.
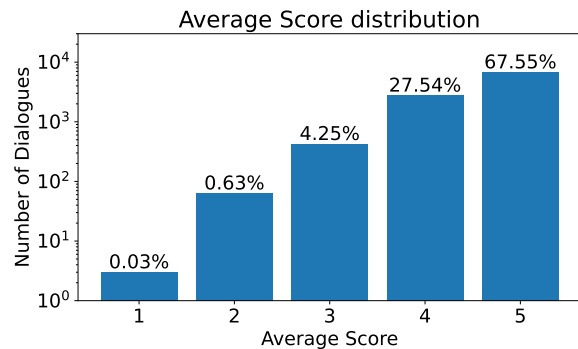
Figure 4: Dialogue level score distribution (turn level average).

than one type of issue (especially engagement plus any other type of issue). However, some responses have more than 3 issues, which is the limit of what we consider to be reasonable. After a quick check, we find these examples belong to two distinct categories: (1) malformed dialogues that contain hallucinations; (2) non-specific responses given on the first few turns of the dialogue.

Number of issues per score (Figure 8) As expected, the lower the score, the higher probability of the response containing what we consider a critical issue. For instance, Score 4 is dominated by Engagement and Assumption issues, which are considered minor, whereas for Scores 1 and 2 the majority of issues relate to Coherence and Commonsense (with engagement also present simultaneously). For Score 1, in particular, we see a large presence of Antisocial issues.

Dialogue level scores (Figures 4, 5) We present the overall quality score distributions when calculating the average and minimum of turn-level scores. While the average of turn-level scores is typically applied by turn-level metrics to output a dialogue-level assessment, we find the minimum to be more representative of the true quality of the dialogue, since it is unreasonable to expect a dialogue possessing a critical error to be of good quality. However, when applying this approach we note that the majority of dialogues are assessed with a Score of 2. Despite not being the focus of this paper, future work should aim to better model dialogue level assessments.

D Human Annotations

For the annotation efforts of this work, we recruited 10 volunteers from our research lab. All participants are graduate professionals with NLP and/or Linguistics background, most of which non-native but fluent speakers of English. Workload for each annotator was limited to 100 examples.

D.1 Issue Detection Validation

For the issue detection validation task (guidelines are presented in Figure 9), we randomly sample 200 examples from the test set. The distribution of issues in this subset is presented in Table 13. If we consider the majority vote to indicate the gold label, GPT-4 performance between issues varies
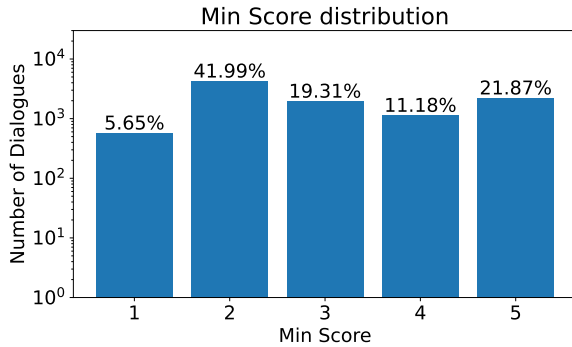
Figure 5: Dialogue level score distribution (turn level minimum).

Issue	Examples	Valid (Maj)	Valid (Abs)
None	30.5 %	91.80 %	98.36 %
Coherence	21.5 %	76.74 %	88.37 %
Commonsense	8.0 %	73.33 %	86.67 %
Assumption	9.0 %	-	-
Repetition	9.0 %	61.11 %	83.33 %
Engagement	23.5 %	-	-
Antisocial	2.0 %	100.00 %	100.00 %
Fluency	2.5 %	100.00 %	100.00 %
Gender Pronoun	0.5 %	-	-
Non-textual	1.5 %	100.00 %	100.00 %
Other	0 %	-	-

Table 13: Distribution of issues in the subset used for human validation of GPT-4 issue detection, together with the validation results when considering majority vote (Maj) and absolute agreement for non-validity (Abs).

significantly, being as low as 61.11 % for Repetition or as high as perfect detection for Antisocial, Fluency and Non-textual. When only considering instances where all annotators agree that GPT-4 was incorrect (13), validation percentages increase significantly. In any case, it is important to note that the vast majority of occurrences correspond to GPT-4 identifying issues where none were present. In the context of issues detection in dialogues, we argue recall is preferable to precision.

D.2 Overall Assessment Annotations

Quality control in crowdsourcing is the subject of significant research in the literature (Daniel et al., 2018). While human evaluators bring subjective insight in to the assessment, they may also be prone to overlooking subtle problems or inconsistencies within the conversation, especially when they are under significant cognitive load. As such, we conducted an experiment that attempts to understand if integrating LLMs as an assistant in the evaluation process can help annotators accurately rate dialogue responses.

For this task, we asked annotators to first provide an assessment having only access to the dialogue history and the response to evaluate. Immediately after this annotation, the annotators are then presented with the issues (or lack of) detected by GPT-4, and without any other information they were then allowed to revise their annotation if deemed appropriate. The guidelines for this task are presented in Figure 10.
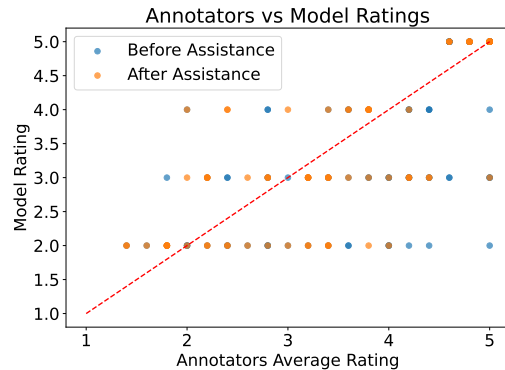

Figure 6: Scatter plot of average ratings vs. model ratings.

A paired samples t-test confirms that the difference between the original assessment and the revised assessment is significant, with $p < 0.01$. We present the scatter plot of average ratings vs. model ratings in Figure 6. Here, we note that the automated issue detection has helped the annotators to converge their ratings towards the model's ratings, improving overall agreement.

To complement Figure 6, we also plot the heatmap of error reductions per annotator and example in Figure 7. This heatmap suggests that the impact of the model's assistance varies significantly among annotators. Annotators 2 and 4 seem to have benefited the most from the model's assistance. This could indicate that annotator low recall, due to high cognitive load, can be mitigated with automated assistance. However, annotator 0 experienced the least impact, with mostly minimal changes in error. Additionally, we observe larger error reductions with responses that contain coherence issues, which is to be expected since this issue requires additional cognitive load, especially when detecting global coherence issues.

With respect to the few instances where there was in increase in the difference between the annotator revision and GPT-4 assessment, both are a result of the annotator changing their assessment from 3 to 1, where the GPT-4 assessment was 2.
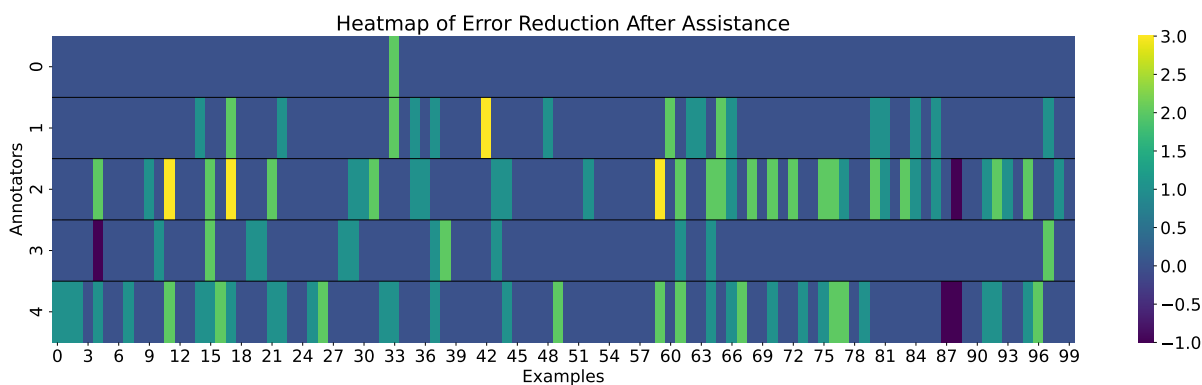
Figure 7: Heatmap of error reductions per annotator.

D.3 Explanation Validation

For this annotation, we manually determine if the explanation is valid. In detail, we provide a binary judgement regarding the validity of the explanation when taking into account the detected issues provided by GPT-4 (which were validated by majority vote by other annotators). An explanation is considered valid if it is fluent and explicitly identifies (if any) the issues reflected in the response under evaluation.

E Implementation Details

We train our models with a language generation objective. Provided with a dialogue history c and a candidate response r, they are tasked to output, in natural language, an overall assessment of the response and a corresponding score $s \in [1, 5]$. We finetune our models on a single RTX A6000 48GB GPU or A100 80GB GPU using Huggingface Transfomers with TRL Supervised Fine-tuning Trainer (SFT) [8] and PEFT[9] for 3 epochs. We conduct a single finetuning run from the base instruction models (full precision) using LoRA (Hu et al., 2021), with $r = 8$, $\alpha = 32$ and dropout set to 0.1. Gradient accumulation steps is set to 4 with a learning rate of $1e - 4$. Batch size was set to maximise VRAM consumption, ranging from 4 up to 64 per device.

For inference using the instruction models, we employ a shared prompt (Table 12) which may be complemented with examples at the end, firstly drawn from the examples used for GPT-4 generation, and then from the training set. For all of our experiments, we employ greedy decoding.

---

[8] huggingface.co/docs/trl/sft_trainer
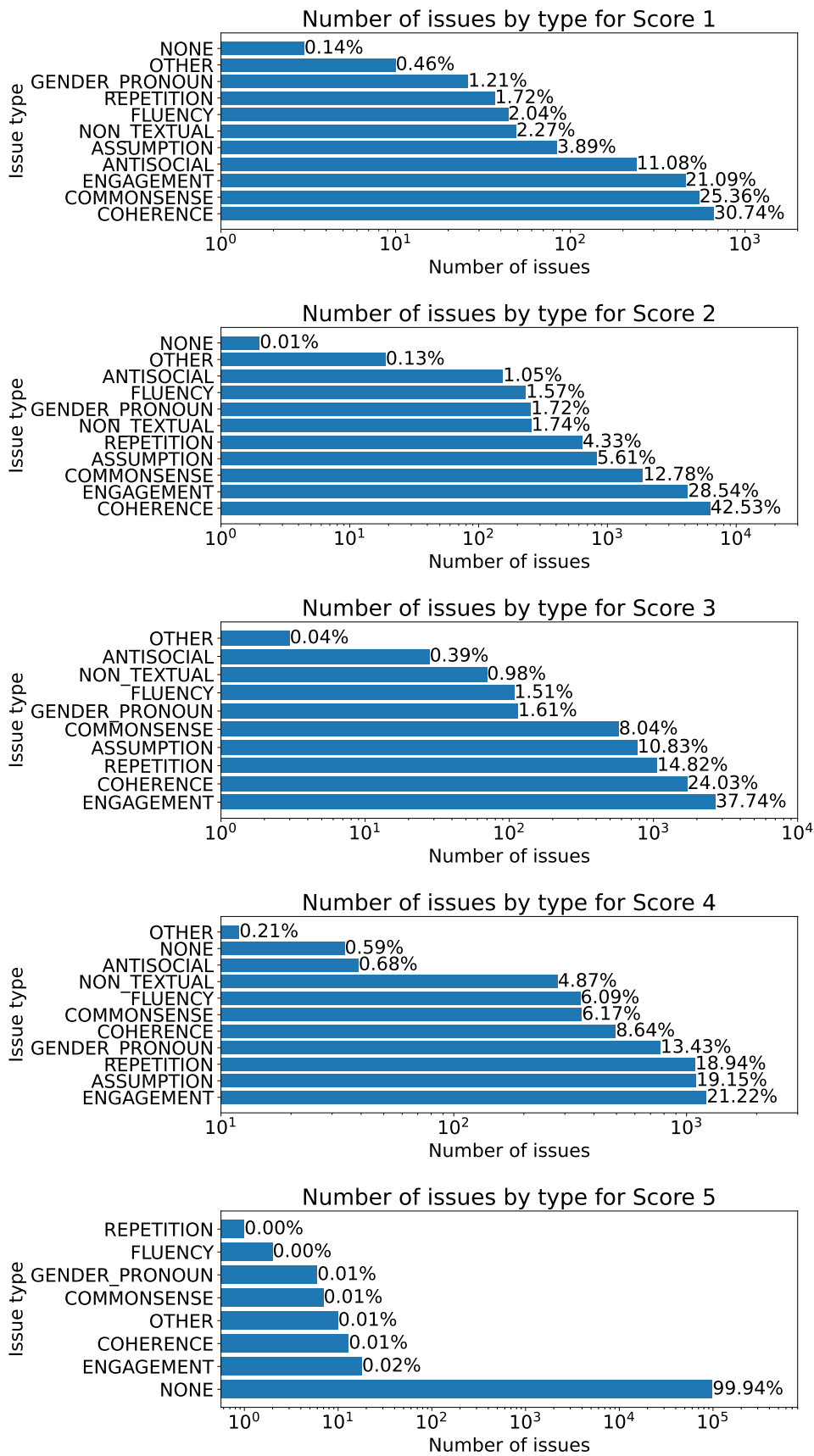[9] huggingface.co/docs/peft

Figure 8: Number (and percentage) of issues per score.

Our work is focused on predicting quality in dialogues. To this end, we asked GPT-4 to detect issues in dialogue responses. **Your task is to determine if the GPT-4 detections are correct. Please note that there might be more than one issue per response.**

We developed a categorization that suits our goals. These categories are:

Objective:
COHERENCE: Contradicts or ignores prior information in the dialogue;
COMMONSENSE: Lacks common knowledge and logic;
REPETITION: Infers information not available in the dialogue context;
FLUENCY: Repeats prior information in the dialogue;
ANTISOCIAL: Contains unsafe or inappropriate behaviour.
NON-TEXTUAL: Includes narrative elements or references unexpected inside a turn of a dyadic interaction.

Subjective:
ENGAGEMENT: Lacks a behaviour or emotion expected from the situation;
ASSUMPTION: Infers information not available in the dialogue context;
GENDER_PRONOUN: Goes against normative pronouns;
OTHER: any other issues.

Examples of each issue are provided at the end.

We are solely focused on the validation of **objective** issues in the response. The remaining categories are included for completeness.

Please provide a ternary answer (0, 1 or 2) to the following question:

Are the objective issue(s) detected by GPT-4 correct (or lack thereof)?

Your annotation(0-2):0 bad indicates that the issue identified is not present in the response, or it detected no objective issues when it was clear at least one objective issue was present;

Your annotation(0-2):1 fair indicates it detected 1 objective issue but missed detecting an additional issue; or the issue detected is present but not correctly described;

Your annotation(0-2):2 good indicates that the issues identified are present in the response; or that there are indeed no issues.

Figure 9: Issue detection validation guidelines.

Our work is focused on predicting quality in dialogues. **Your task is to provide an overall assessment for the response (1-5), given the prior context.**

The annotation is to be conducted in 2 steps. In the first step, you will rate the response when provided with the dialogue alone. After rating the response, you will then (**and only then**) move to the next step, where you will rate the same response, using the same annotation schema, but provided with automated guidance (in the form of issues detection).

You may disagree with the automated guidance, or find that it does not change your initial assessment. If so, you can input the same score as before. If not, only change the assessment in the second step. **Always keep the initial assessment unchanged.**

Your annotation(1-5):1 awful   The response contains several major issues (e.g contradicts itself or lacks common sense) that severely affect the interaction. The user would be hard pressed to continue such a conversation.

Your annotation(1-5):2 bad   The response contains major issues that affect the conversation.

Your annotation(1-5):3 fair   The response contains some issues that moderately reduces the quality of the interaction (e.g. unexpected/ non-engaging response, or minor contradiction).

Your annotation(1-5):4 good   The response may contain a minor issue (e.g a small typo) that does not affect the quality of the response.

Your annotation(1-5):5 excellent   Perfect response, without any issues.

Figure 10: Overal assessment guidelines.