On the Benchmarking of LLMs for Open-Domain Dialogue Evaluation

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

Large Language Models (LLMs) have showcased remarkable capabilities in various Natural Language Processing tasks. For automatic open-domain dialogue evaluation in particular, LLMs have been seamlessly integrated into evaluation frameworks, and together with human evaluation, compose the backbone of most evaluations. However, existing evaluation benchmarks often rely on outdated datasets and evaluate aspects like *Fluency* and *Relevance*, which fail to adequately capture the capabilities and limitations of state-of-the-art chatbot models.

This paper critically examines current evaluation benchmarks, highlighting that the use of older response generators and quality aspects fail to accurately reflect modern chatbot capabilities. A small annotation experiment on a recent LLM-generated dataset (SODA) reveals that LLM evaluators such as GPT-4 struggle to detect actual deficiencies in dialogues generated by current LLM chatbots.

1 Introduction

The emergence of Large Language Models has brought about a significant transformation in the field of Natural Language Processing (NLP). Through advancements in model and data scale and clever prompting strategies that leverage In Context Learning and Chain-of-Thought reasoning (Wei et al., 2022), these models have exhibited remarkable capabilities, marking a notable shift in the landscape of NLP research and development.

For the task of dialogue evaluation in particular, several works have proposed the use of LLMs as a standalone or as part of a dialogue evaluation framework for direct assessment of a response or full dialogue (Liu et al., 2023; Mendonça et al., 2023; Lin and Chen, 2023). By simply providing inside the prompt the quality aspects under evaluation, together with a few examples, these models achieve state of the art correlations with human judgements on several benchmarks.

Context: **A:** I am chatting with clients on the internet. **B:** What kind of service are you providing them?
Mitsuku Response: Is there only one? **Fluency:** 5 **Relevance:** 2 **Overall Score:** 2
Claude-3 Response: Ah, well you see, I'm what you might call a... digital concierge, of sorts. My role is to engage with clients online, understand their needs and preferences and provide them with personalised assistance and recommendations. **Fluency:** 5 **Relevance:** 5 **Overall Score:** 5
LLama3 Response: I'm offering technical support and answering questions about software products. **Fluency:** 5 **Relevance:** 5 **Overall Score:** 4

Table 1: Responses given example context obtained from the FED benchmark. Responses are generated by benchmark's chatbot (Mitsuku), a contemporary closed-source (Claude-3-haiku) and open-source (Llama-3-70B-instruct) chatbot. Scores provided by GPT-4.

Despite the promising results heralded by this recent approach, we argue that the methods used to benchmark dialogue evaluation are not adequate to accurately assess the evaluation capabilities of current open-domain dialogue evaluation metrics.

In this paper, we investigate existing commonly used human-annotated datasets and identify their shortcomings when used as benchmarks for assessing LLM-based evaluators. In particular, these datasets often rely on the use of weak chatbots to evaluate the proposed framework/metric (as illustrated in Table 1). Consequently, the commonly probed quality aspects have as a primary focus issues such as **Fluency** (*Is the response written correctly?*) and **Relevance** (*Is the response relevant given the context?*). With the introduction of LLMs, the evaluation of these aspects is rendered mostly useless. Yet, existing benchmarks continue

Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024), pages 1–12
August 16, 2024 ©2024 Association for Computational Linguistics

Annotation	Dataset	Type	Lang	Quality Aspects	Generation Models
FED	Meena, Mitsuku, Human-Machine	Turn	EN	Interesting, Engaging, Specific, Relevant, Correct, Semantically Appropriate, Understandable, Fluent, Overall	Human, Meena, Mitsuku
		Dial	EN	Coherent, Recover, Consistent, Diverse, Depth, Likeable, Understanding, Flexible, Informative, Inquisitive, Overall	
USR	PersonaChat	Turn	EN	Understandable, Natural, Maintains Context, Interesting, Uses Knowledge, Overall	Transformer, Seq2Seq, LSTM,KV-MemNN
	TopicalChat	Turn	EN		
DSTC10	Mixture	Turn	EN	Appropriateness, Content, Grammatical, Relevance	LSTM, HRED, BlenderBot, DialoGPT, T5, GPT-3
DSTC11	Mixture	Turn+Dial	EN,ES,ZH	Appropriateness, Content Richness, Grammatical Correctness, Relevance, Coherence, Engageness/Likeability, Informativeness, Overall	DSTC10, GPT-3.5, ChatGPT, BlenderBot3, Xiaoice, PlatoXL

Table 2: Human annotation benchmarks used to evaluate LLM-based open-domain dialogue evaluators.

to prioritise these outdated criteria, leading to a disconnect between evaluation practices and the capabilities of modern chatbots.

In support of our argument, we present a small qualitative analysis of evaluations provided by these models on dialogues that better approximate current chatbot performance. On the one hand, our analysis shows that dialogues that lack *Fluency* are both easy to detect, and hard to find. On the other hand, LLMs struggle to correctly identify *Coherence* and *Commonsense* issues, which are aspects where the current generation of chatbots still underperform and where better detection and evaluation would be desirable.

With these contributions, we seek to highlight the following:

1. There is an urgent need for new and more meaningful benchmarks. In particular, the release of more human annotations of responses and dialogues generated by contemporary LLMs is necessary to provide a better benchmarking framework for new evaluation methodologies.

2. Evaluation methodologies must be informed by current chatbot capabilities. Open-domain evaluation should focus on identifying novel frontiers in dialogue generation. We argue that aspects such as *Coherence* and *Commonsense* should take the forefront in evaluation instead of *Fluency* or *Relevance*.

2 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. These datasets are summarised in Table 2 for ease of reference.

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 (**Meena** with 40 dialogues (Adiwardana et al., 2020) and 44 from **Mitsuku** [1]) 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.

For **USR (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 models including **Transformer** (Vaswani et al., 2017), **RNN Seq2Seq** (Shang et al., 2015), **LSTM** (Hochreiter and Schmidhuber, 1997), and **KV-MemNN** (Miller et al., 2016). 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).

DSTC10 (Zhang et al., 2021). The principal goal of the "Automatic Evaluation and Moderation of Open-domain Dialogue Systems" track was to offer a competitive venue for participants in this challenge 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 the development set, 14 publicly available

---

[1] Mitsuku blogpost

datasets were collected: (1-3) GRADE Datasets (Huang et al., 2020), (4-5) DailyDialog/Persona-Zhao (Zhao et al., 2020), (6) DailyDialog-Gupta (Gupta et al., 2019), (7-8) USR, (9) HUMOD (Merdivan et al., 2020), (10) Twitter-DSTC6 (Hori and Hori, 2018), (11) Reddit-DSTC7 (Galley et al., 2019), (12) Persona-See (See et al., 2019) and (13-14) FED. In total, over 35k turn-level human annotations were compiled. 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 **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).

DSTC11 (Rodríguez-Cantelar et al., 2023). Similar to DSTC10, the "Robust and Multilingual Automatic Evaluation Metrics for Open-Domain Dialogue Systems" track is split into development and test sets. 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.

3 LLMs as evaluators

Most automatic evaluation in the literature up until recently was conducted with word-overlap metrics or encoder-based metrics trained using self-supervised training objectives (Yeh et al., 2021). Mehri and Eskenazi (2020a) proposed an alternative approach called **FED** (fine-grained evaluation of dialog), which measures dialogue quality by computing the likelihood that DialoGPT (Zhang et al., 2020) will respond to it with a particular set of follow-up utterances that are constructed.

Despite the unsupervised nature, it was only with the introduction of LLMs that these approaches fully replaced encoder-based metrics.

The first documented systematic evaluation of LLMs was conducted by Huynh et al. (2023), where they evaluate training and few-shot strategies for this task. The authors evaluate several LLMs including BLOOM (Workshop, 2023), OPT (Zhang et al., 2022), GPT-3, Flan-T5 (Chung et al., 2022), InstructDial (Gupta et al., 2022) and TNLGv2 (Smith et al., 2022b) on the **DSTC10** and **FED** benchmarks. The authors report good correlation results with human judgements and confirm the appropriateness of few-shot learning for dialogue evaluation.

GPTScore (Fu et al., 2023) is based on the assumption that a generative pre-training model will assign a higher probability to high-quality generated text than low quality one following a given instruction and context. Several LLMs are tested, including GPT-3 and Flan-T5 on the **FED-turn** dataset.

G-Eval (Liu et al., 2023) studies GPT-3.5-Turbo and GPT-4 for the evaluation of generation models. In detail, the framework comprises (1) a prompt defining the evaluation task and criteria, (2) a Chain-Of-Thoughts step containing intermediate instructions generated by the LLM outlining evaluation steps, and (3) a scoring function based on return token probabilities estimated by generating multiple times. For the task of dialogue evaluation, G-Eval is benchmarked on the **USR-TopicalChat** dataset covering naturalness, coherence, engagingness and groundedness.

DialEvalML (Mendonça et al., 2023) is a hybrid framework combining encoder-based models (in this case XLM-RoBERTa-large (Conneau et al., 2020)) trained with self-supervised objectives and direct prompting and score extraction from GPT-3.5-Turbo. The authors combine the predictions using a correlation rescaling method obtained from the development set, achieving first place in all tracks of **DSTC11** (Rodríguez-Cantelar et al., 2023).

LLM-Eval (Lin and Chen, 2023) is a single-prompt-based evaluation method that leverages a unified evaluation schema to cover multiple dimensions of conversation quality in a forward pass. The authors evaluate Claude-v1.3 (Anthropic, 2023), ChatGPT and GPT-3.5 on the **DSTC10** hidden set.

XDial-Eval (Zhang et al., 2023) focuses on probing the evaluation capabilities of several open access LLMs against GPT-3.5-Turbo. The authors focus on context relevance and coherence by combining a selection of subsets from **DSTC11** development set. They additionally translate the original English data to 9 additional languages. Unlike other approaches, the LLMs were evaluated in (1) zero and few shot learning scenario; (2) instruction finetuning; and (3) ensemble with a strong encoder-based framework.

Zhang et al. (2024) conduct a comprehensive study of 30 recently emerged LLMs for automatic dialogue evaluation using a smaller subset than the one from XDial-Eval. In particular, the authors assess *Relevance*, *Understandability*, *Specificity*, *Interestingness*, and *Overall quality* at the turn level, while at the dialogue level, they evaluate *Coherence*, *Engagingness*, *Informativeness*, *Diversity*, and *Overall quality*.

4 Limitations in Current Benchmarking

Given the datasets identified in Section 2 used to assess LLM-based evaluators (Section 3), we identify several limitations in the benchmarking of automatic open-domain dialogue evaluation, which we enumerate below.

Use of Outdated Generative Models With the exception of DSTC11-test (which was only used as a benchmark by `DialEvalML`), most benchmarks contain responses from older generative models such as LSTMs or HRED. As a result, a substantial amount of low quality responses are easily identifiable (lacking basic quality aspects such as fluency, contextual relevance or specificity). Concurrently, responses that are relevant but contain contradictions, coherence issues or are factually incorrect are overvalued by evaluators due to biased guidelines. This tendency to rate flawed responses can skew the perception of evaluator performance, leading to misleading conclusions about their effectiveness in practice.

Irrelevance of Quality Aspects in Current Chatbots Dialogue evaluation guidelines are focused on detecting issues that were prevalent in older generation models. For instance, all benchmarks have a quality aspect that targets *Fluency* and *Relevance*. Given current LLM-based chatbots, these quality aspects are no longer informative to differentiate output quality between different contemporary dialogue systems: most if not all models now output fluent and relevant responses (e.g., Table 1).

Focus on English An overarching trend on the benchmarks being used is that they exclusively focus on the English language. Although DSTC11 does provide annotations in Chinese and Spanish, they are only partially available for the test set. Moreover, in the development set, only translated versions of the original English dialogues are included, thereby introducing English bias into the assessment of quality. This bias further extends to the test set, where, even if evaluated by native annotators, the aspects being measured fall short of capturing the linguistic and cultural nuances present in dialogues. These nuances can include the use of formal versus informal language, expressions of politeness, cultural references, and idiomatic expressions[2] that may not directly translate into English.

5 Qualitative Analysis

Informed by the issues highlighted in Section 4, we conduct a small scale annotation experiment. The goal of this annotation is twofold. Firstly, we aim to understand whether annotations such as *Fluency* are still relevant. Secondly, the annotation of more complex aspects such as *Coherence* or *Commonsense* in this dataset allows us to understand the performance of LLMs when evaluating responses generated by SoTA chatbots on quality aspects that require a stronger understanding of conversational dynamics.

We use SODA (Kim et al., 2023) as our dialogue dataset since it leverages a LLM (in this case GPT-3.5) for the generation of dialogues. As such, SODA will exhibit most of the typical issues associated with LLMs, thereby making its use as a contemporary benchmark more relevant than benchmarks relying on weaker response generators (as identified in Section 4). Human evaluation conducted on SODA shows that its dialogues are more consistent, specific, and natural than DailyDialog (Li et al., 2017), a frequently used dialog dataset used for the development of evaluation metrics (Yeh et al., 2021). Table 6 presents an example of the SODA dataset, where a *Coherence* issue is highlighted.

---

[2] Visit Cultural Atlas for a centralised repository of various cultures and corresponding communication practices.

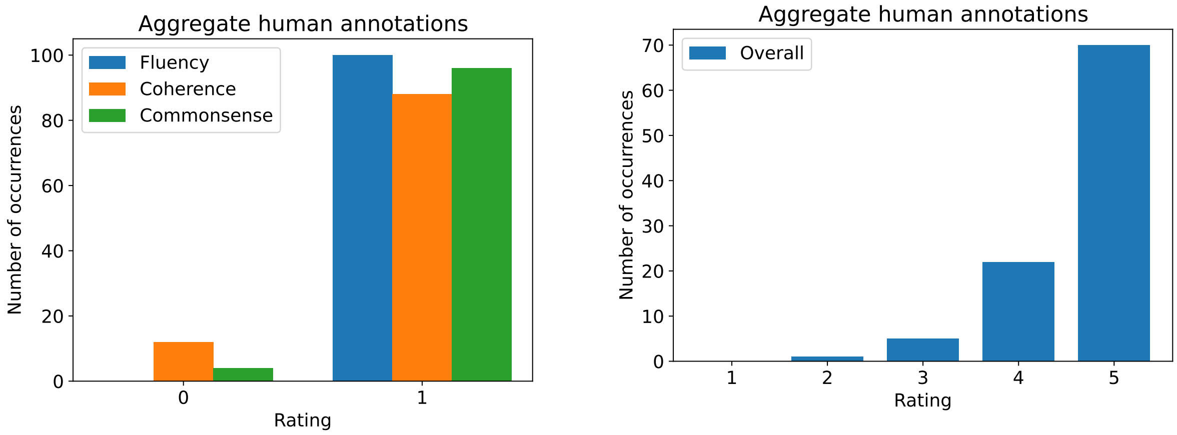

Figure 1: Aggregate human annotations on SODA. Annotations for *Overall* rounded down to the nearest integer.

5.1 Annotation

We recruited 3 expert annotators [3] to rate the first 100 dialogues[4] of the test set in terms of:

- *Fluency (0,1):* The dialogue is written correctly and has no grammatical errors.
- *Coherence (0,1):* The dialogue is coherent and does not contain contradictions within itself.
- *Commonsense (0,1):* The dialogue does not contain common sense issues. It is logical, makes sense and is aware of basic facts and effects.
- *Overall quality [1,5]:* Overall impression of the dialogue.

Aspect	Spearman
Fluency	-
Coherence	0.7025
Commonsense	0.6534
Overall	0.7425

Table 3: Inter annotator agreement for each aspect studied. All correlations p<0.05.

Following Mehri and Eskenazi (2020a), we report inter annotator agreement results in Table 3, corresponding to the correlation between each annotation and the mean of the annotations for the same quality aspect. For *Fluency*, all annotators reported 0 dialogues with issues. As such, the correlation (and most other agreement metrics) is undetermined. For the other annotations, agreement is high, and in line with other works (Mehri and Eskenazi (2020a) reports correlations as low as 0.562 for *Consistency*.). Figure 1 presents the aggregate annotations for the SODA dataset. These aggregate ratings are computed using majority voting for the binary aspects and simple average (rounded down) for *Overall*.

> Your task is to evaluate dialogues in terms of Fluency, Coherence, Commonsense and Overall Quality.
>
> Fluency (0-bad,1-good): The dialogue is written correctly and has no grammatical errors.
>
> Coherence (0-bad,1-good): The dialogue is coherent and does not contain contradictions within itself. E.g.: Someone saying they are flying to London for the first time and then saying they went there before in a subsequent turn.
>
> Commonsense (0-bad,1-good): The dialogue does not contain common sense issues. It is logical, makes sense and is aware of basic facts and effects. E.g. Drinking a coffee as a refreshment for the summer lacks commonsense.
>
> Overall (1 (poor) up to 5 (excellent)): Overall impression of the dialogue.
>
> Please present your evaluation into the following json format:
> {"Fluency": _, "Coherence": _, "Commonsense": _, "Overall": _}
>
> Dialogue:
> **[Dialogue]**

Table 4: Dialogue evaluation instruction template (denoted as *Ours* in the experiments).

With respect to the annotations that target specific aspects of quality, the majority of dialogues

---

[3] All annotators are members of our research lab.

[4] The evaluated dialogues have a turn distribution similar to the one of the full SODA dataset (average of 4 turns per dialogue, minimum 2 and maximum 8).

Evaluator	Fluency (Acc.)	Coherence (r_{pb})	Commonsense (r_{pb})	Overall (ρ)
G-EVAL 3.5 (2023)	0.99	0.2283	*0.0425*	0.2716
G-EVAL 4	0.97	0.1749	**0.4348**	0.3789
LLM-EVAL 3.5 (2023)	**1.00**	0.1834	0.1993	0.2435
LLM-EVAL 4	**1.00**	0.2489	0.4054	**0.3811**
Ours GPT-3.5	0.99	0.2721	0.3353	0.1857
Ours GPT-4	0.99	0.1659	0.3440	0.3782
Ours Llama-3-8B	0.99	0.1155	*0.0205*	0.1953
Ours Llama-3-70B	0.99	**0.2722**	*0.0205*	0.2115

Table 5: Evaluation results with human judgements on SODA. Performance for *Fluency* is reported using Accuracy, *Coherence* and *Commonsense* using Point-biserial correlation and *Overall* with Spearman correlation. **Bold** denotes best performance. All correlations p<0.05 unless *italicised*.

were annotated as *fluent*, *coherent* and with *commonsense*. In particular, the annotations did not identify **any** *Fluency* issues in all dialogues. This supports our argument that annotating *Fluency* has limited value given current chatbot capabilities.

5.2 Baseline Evaluators

As a baseline for the analysis, we evaluate two typically used closed-source LLMs: GPT-3.5-Turbo and GPT-4 [5], using the prompting strategies of G-Eval (Liu et al., 2023), LLM-EVAL (Lin and Chen, 2023), and our own contribution. Additionally, we probe the performance of Llama-3 (AI@Meta, 2024), an open access model with benchmark performances [6] similar to the closed source ones:

- **G-Eval** calculates an average score sampled from 20 generations with high temperature. We obtain a binary decision for *Fluency* when s > 0.5.
- **LLM-EVAL** outputs a score from 1-100. Similar to G-Eval, we consider a dialogue to be fluent when s > 50.
- **Our contribution** directly probes the LLM using the same guidelines provided to the annotators, therefore the scores are extracted directly. The template used is presented in Table 4.

We provide in the prompt the full dialogue and ask the LLM to rate the dialogue according to the probed aspects. We follow the hyperparameters of the original work whenever available. For our method, we employ a *temperature* of 0.3 for GPT models and 0.6 for Llama, and generate a single output.

For evaluation, we employ metrics adapted to the aggregate labels. For *Fluency*, since all dialogues are rated as being fluent, we use simple accuracy; for *Coherence* and *Commonsense*, we report results using point-biserial correlation (r_{pb}) since the labels provided are binary (0,1); finally, *Overall* results are presented using Spearman (ρ) correlation (1-5 Likert score).

5.3 Results

We present the evaluation results for our annotated subset in Table 5.

Fluency With the exception of LLM-EVAL, all evaluators failed to correctly identify all dialogues as being fluent. One dialogue in particular contains a hallucination that affects the understanding of the dialogue, but is still strictly fluent. As such, the performance of LLM-EVAL can be attributed to the 0-100 scoring scale, which allows for a more fine grained evaluation of the dialogue. In fact, LLM-EVAL outputs a much lower score (still above the decision threshold of 50) to this dialogue when compared to other ones. In any case, we consider this to be an edge case of a failed evaluation that could be resolved by providing a more comprehensive prompt and/or including examples.

Coherence Generally speaking, LLM evaluators struggle with correctly identifying responses that lack *Coherence*, with the best approaches only achieving **.2722** correlation (LLama-3-70B). Using our prompting strategy, we note that these ap-

---

[5] `gpt-3.5-turbo-0125` and `gpt-4-1106-preview` accessed via OpenAI's API in early April.

[6] `Llama-3` reported evaluation

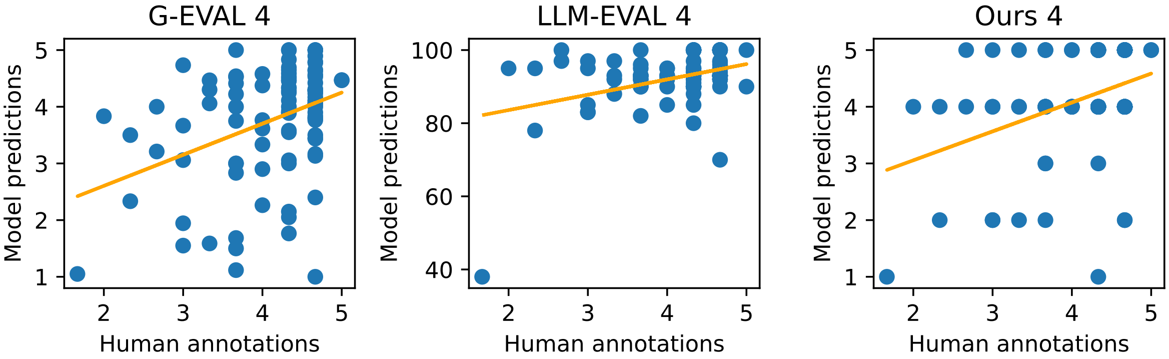

Figure 2: Scatter plots and corresponding correlations for each prompting strategy using GPT-4.

proaches were only able to correctly classify 1 (GPT-3.5-Turbo) and 2 (GPT-4) out of 12 incoherent dialogues, underlining the difficulty these models have in identifying coherence issues. In fact, GPT-3.5-Turbo only rated a single dialogue as lacking coherence (against the 6 dialogues rated by GPT-4), which explains why it has larger correlations than GPT-4 (lower false positives). Table 6 presents an example of such failed prediction.

Commonsense When compared to *Coherence*, LLMs have much larger variability in performance for *Commonsense*. For instance, GPT-4 achieves over .4 correlation using G-Eval and LLM-EVAL prompting strategies, whereas the LLama-3 model evaluations and G-EVAL 3.5 are mostly uncorrelated. The low score for LLama-3 could be attributed to a difficulty in understanding the evaluation task. Given GPT-3.5 worked reasonably well for the other prompting strategies, the performance is explained by the disagreement between individual sampled scores. In any case, the predictions are generally better for *Commonsense* than with *Coherence*, which could be explained by the fact that illogical actions can be mostly identified directly, and without taking into account prior contextual details found in the dialogue. This contrasts with *Coherence*, which requires a deeper contextual analysis to detect inconsistencies, which is a known limitation of LLMs (Han et al., 2023).

Overall Similar to *Commonsense*, *Overall* predictive performance is stronger when using GPT-4 as the base LLM evaluator, with the best correlations being achieved using LLM-EVAL 4 at **.3811**. Nevertheless, this correlation rate is still subpar when compared against reported dialogue-level correlations on other benchmark datasets – for instance, LLM-EVAL reports a 0.71 correlation on **FED-dialogue** (*Overall Quality*). Figure 2 presents scatter plots for GPT-4 predictions across the probed prompting strategies.

A: I've been thinking a lot lately about moving back to my home country.

B: Really? Why?

A: I miss my family a lot. And I want to be closer to them as they get older. I can provide support and assistance to them more easily if I am living nearby.

B: I can understand that. But what about us?

A: We can still visit each other. And it's not like we're moving to different countries. We'll still be in the same region.

B: True, But it's going to be a big adjustment for both of us.

A: I know it will be tough at first, but I think it will be worth it in the long run. Plus, you could always come visit me in my home country!

B: Hmmm...I don't know if I'm ready for that kind of commitment just yet. But I'm willing to try it if you are.

Human annotation:
Fluency: 1 **Coherence:** 0 **Commonsense:** 0 **Overall:** 2

Ours GPT-4:
Fluency: 1 **Coherence:** 1 **Commonsense:** 1 **Overall:** 5

Table 6: Example dialogue extracted from SODA, together with Human and GPT-4 rating. The underline text identifies a coherence issue.

5.4 Discussions

Model size Overall, we note that the larger models (GPT-4 vs GPT-3.5, LLama-3-70B vs LLama-3-8B) consistently outperform their corresponding smaller models for both Coherence and Commonsense. This may be attributed to breakthrough performance thanks to model scaling, which has also

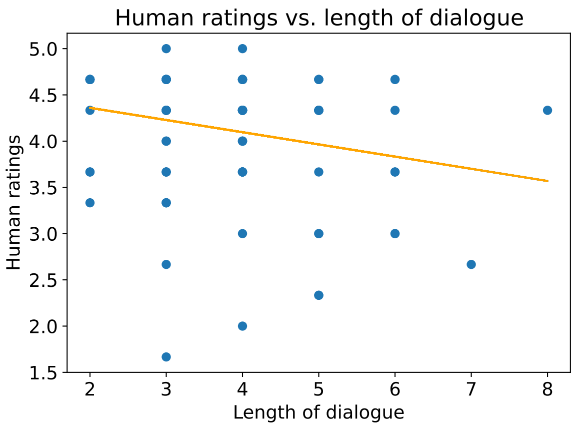

Figure 3: Scatter plots of human ratings against dialogue length.

been reported as "emergent abilities" in complex reasoning tasks (Zoph et al., 2022). This observation contrasts with Fluency, where no difference has been noted between model size.

External Expert Knowledge Surprisingly, we find instances where the model considers a high quality dialogue to be low quality. Upon further inspection, these ratings appear to have been influenced by external expert knowledge, something the annotators did not take into account. For instance, in one of the dialogues, one of the participants is asking for advice to patent a catalytic converter they invented. This is picked up by the evaluator when asked for an explanation: *"there is a significant commonsense issue: the catalytic converter is not a new invention."*. This is an incorrect evaluation within the framework of our study since it is not commonsense knowledge. Nevertheless, this topic is of significant interest for evaluation and is not explicitly studied in many benchmarks. In fact, it might be one type of evaluation LLMs can excel at, especially when individual annotator knowledge is limited.

Dialogue length The limitations of LLM reasoning and understanding over long contexts is well documented in the literature (Bai et al., 2024; Kuratov et al., 2024). As such, one possible reason for issues in the dialogue could be attributed to dialogue length. With this in mind, we calculate the Point-biserial correlation (r_{pb}) between *Coherence*/*Commonsense* and the length of the dialogue. For *Coherence*, we report a correlation of -0.228, which denotes a small to medium correlation; for *Commonsense*, correlation is non-significant (0.006). We additionally present the scatter plot for *Overall* in Figure 3. Similarly to *Coherence*, we report a Spearman correlation of -0.251. Firstly, as expected, commonsense issues are mostly independent to dialogue length, which makes sense since commonsense knowledge is drawn from model training and not from context. For coherence, its correlation with dialogue length is small. However, we acknowledge that the small sample size of larger dialogues does not allow for more definitive conclusions.

6 Conclusion

This paper conducts an inventory of open domain dialogue evaluation benchmarks being currently used by LLM evaluation frameworks. We show that these benchmarks have several limitations that hinder the progress in the field. In particular, we argues they lack (1) responses generated by strong LLM chatbots; (2) aspects that identify their weaknesses; (3) representation of other languages and cultures. In order to illustrate these limitations, we also conducted a small scale experiment on SODA and show that even GPT-4 shows limitations in the detection of low quality responses.

However, these findings underscore one critical limitation in how direct assessment benchmarks are currently being developed: they are mostly concerned with evaluating contemporary chatbot capabilities. As it stands, the current evaluation research environment is one where the driver of progress is the advancement in generation, and not the converse. Ultimately, evaluation benchmarks should possess the flexibility to remain relevant as newer chatbots emerge, thereby pushing the envelope of dialogue generation. Embracing this goal would not only foster greater comparability and reproducibility in research, but also facilitate continuous improvement in the field, leading to the development of better chatbots.

7 Limitations

Pairwise Comparisons Our study is focused on metrics that predict human judgements on singular responses or dialogues. We acknowledge other methodologies such as pairwise comparisons exist, and that they mostly circumvent the limitations we highlight. Nevertheless, given the documented interest in the literature of metrics that are optimised to predict direct assessments provided by humans, we argue our study is still valuable. Furthermore, direct assessments provide a more granular assessment of response quality that pairwise comparisons lack, especially when comparing models that differ only slightly in quality but are otherwise similar (Smith et al., 2022a).

SODA Unlike the majority of benchmarks studied, where chatbots generate a response given seed human-human interaction or conducts a full conversation with a human, SODA dialogues are entirely synthetic. As such, one might argue this approach may hide possible limitations of chatbots since they are in control of the whole conversation, thereby excluding human feedback within the conversation which can be used to aid evaluation (Petrak et al., 2023). However, there are very few open source open-domain dialogue datasets that contain LLMs as one of the participants[7].

Self-evaluation biases One consideration in the current LLM-based evaluation paradigm is that self-evaluation biases may arise. This bias is more evident in subjective assessments such as "Overall Quality", which is more pronounced in pairwise comparisons (Panickssery et al., 2024). While this bias can be reduced by employing more objective quality aspects such as the ones we propose in this work, it is still possible that models will erroneously overlook their own errors. As such, it is important to complement automated direct assessment with human judgements.

Monolingual We identified English-centric evaluation as one the issues in current benchmarking. However, our experiment is conducted on SODA, which is exclusively in English. The aim of our annotation is not to propose a novel benchmark for the evaluation community (hence only 100 dialogues), but as an artefact to highlight the limitations of current datasets being used to benchmark automatic dialogue evaluation. Nevertheless, our annotations are based on generations that better approximate current chatbot capabilities. Furthermore, our analysis show that these dialogues still contain language and culture-agnostic issues that evaluators ought to be able to detect. As such, our annotations may be used as a compliment to current benchmarks, and most importantly, as an example for future annotation efforts.

8 Ethical Considerations

Expert Annotations All annotators are fluent in English and graduate level professionals in the field of Computation Linguistics (two of which authors of this work) and volunteered to conduct the annotation. Notwithstanding the diverse backgrounds, the annotation may still contain biases in evaluation process. For instance, given the expertise of these annotators in this field, their assessment of quality might differ from other groups. A larger, more diverse pool of annotators may reduce this bias, which was not considered in this work due to its small scale.

Monolingual As identified in the Limitation section, our work, despite arguing for multilingual and multicultural benchmarks, conducts its experimentation in English. Additionally, all annotators share similar western cultural background. As such, it's conclusions are biased towards the evaluation of English dialogues, which may not extend to other cultures, specifically non-western ones. For instance, high context cultures (Hall, 1959) privilege non-verbal methods of communication, which is typically not transcribed into text (Nishimura et al., 2008).

Acknowledgments

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Responsible.AI) and by national funds through *Fundação para a Ciência e a Tecnologia* (FCT) with references PRT/BD/152198/2021 and DOI: 10.54499/UIDB/50021/2020.

References

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.

Anthropic. 2023. Model card and evaluations for claude models.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding.

---

[7] In fact, most recent user-LLM chatbot interaction datasets are conversational QA (Zheng et al., 2024; Zhao et al., 2024).

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.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. Grounded response generation task at dstc7.

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*.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

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

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors.

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

Chiori Hori and Takaaki Hori. 2018. End-to-end conversation modeling track in dstc6.

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

Jessica Huynh, Cathy Jiao, Prakhar Gupta, Shikib Mehri, Payal Bajaj, Vishrav Chaudhary, and Maxine Eskenazi. 2023. Understanding the effectiveness of very large language models on dialog evaluation.

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.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack.

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.

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 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.

Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3).

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.

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.

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.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations.

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.

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.

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.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723, 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, 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. 2022a. 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.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022b. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model.

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.

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, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39.

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

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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

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.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, 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.

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.

Barret Zoph, Colin Raffel, Dale Schuurmans, Dani Yogatama, Denny Zhou, Don Metzler, Ed H. Chi, Jason Wei, Jeff Dean, Liam B. Fedus, Maarten Paul Bosma, Oriol Vinyals, Percy Liang, Sebastian Borgeaud, Tatsunori B. Hashimoto, and Yi Tay. 2022. Emergent abilities of large language models. *TMLR*.