A Comprehensive Analysis of the Effectiveness of Large Language Models as Automatic Dialogue Evaluators

Chen Zhang[1], Luis Fernando D'Haro[2], Yiming Chen[1], Malu Zhang[3*], Haizhou Li[1,4]

[1] National University of Singapore
[2] Speech Technology Group - Universidad Politécnica de Madrid, Spain
[3] University of Electronic Science and Technology of China
[4] The Chinese University of Hong Kong (Shenzhen), China
chen_zhang@u.nus.edu, maluzhang@uestc.edu.cn

Abstract

Automatic evaluation is an integral aspect of dialogue system research. The traditional reference-based NLG metrics are generally found to be unsuitable for dialogue assessment. Consequently, recent studies have suggested various unique, reference-free neural metrics that better align with human evaluations. Notably among them, large language models (LLMs), particularly the instruction-tuned variants like ChatGPT, are shown to be promising substitutes for human judges. Yet, existing works on utilizing LLMs for automatic dialogue evaluation are limited in their scope in terms of the number of meta-evaluation datasets, mode of evaluation, coverage of LLMs, etc. Hence, it remains inconclusive how effective these LLMs are. To this end, we conduct a comprehensive study on the application of LLMs for automatic dialogue evaluation. Specifically, we analyze the multi-dimensional evaluation capability of 30 recently emerged LLMs at both turn and dialogue levels, using a comprehensive set of 12 meta-evaluation datasets. Additionally, we probe the robustness of the LLMs in handling various adversarial perturbations at both turn and dialogue levels. Finally, we explore how model-level and dimension-level ensembles impact the evaluation performance. All resources are available at https://github.com/e0397123/comp-analysis.

Introduction

Evaluation remains a persistent challenge in dialogue system research (Mehri et al. 2022a). At present, human evaluation is regarded as the most reliable method for comprehensively assessing the quality of dialogue. However, because of the considerable expense and lack of reproducibility associated with human evaluations, automatic measures have been proposed to complement human evaluation. The automatic measures can be categorized into two main groups: reference-based and reference-free. Due to the poor alignment of reference-based metrics, such as BLEU (Papineni et al. 2002), with human evaluation (Liu et al. 2016), existing studies mainly focus on developing neural-based reference-free evaluators (Yeh, Eskenazi, and Mehri 2021). Although such reference-free evaluators have demonstrated improved correlations with human evaluation over reference-based metrics, they are still far from being the perfect proxy of human judges. Additionally, they exhibit poor generalization to dialogue data that are different from what they are trained on (Zhang et al. 2022a).

The recent advancement in large language models (Brown et al. 2020; Chowdhery et al. 2022; Touvron et al. 2023a) coupled with refined alignment techniques (Wei et al. 2022a; Ouyang and et al. 2022) lead to an extensive array of general-purpose AI assistants that are capable of tackling a broad spectrum of NLP tasks. Harnessing their strong language understanding capability, LLMs, especially the family of instruction-tuned models, offer great promise as effective and generalized automatic dialogue evaluators. Several recent works (Huynh et al. 2023; Chen et al. 2023a; Liu et al. 2023; Lin and Chen 2023; Fu et al. 2023) report strong correlations of LLMs with human evaluation. Yet, the scope of their assessment is limited: (1) The coverage of LLMs is restricted, with a primary emphasis on proprietary models, such as OpenAI ChatGPT/GPT-4. Lately, there has been an exponential growth of open-source foundation models (Almazrouei et al. 2023; Touvron et al. 2023a; Scao et al. 2023) and ChatGPT-like LLMs (Taori et al. 2023; Chiang et al. 2023; Chen et al. 2023b), a timely survey is neccessary to examine their effectiveness as automatic dialogue evaluators. (2) The mode of evaluation primarily concentrates on correlation analysis with a limited number of meta-evaluation datasets. We not only conduct a comprehensive correlation analysis on 12 meta-evaluation datasets[1] along different evaluation dimensions[2], but also probe the robustness of LLMs against adversarial perturbations at both turn and dialogue levels.

Our work serves as a useful guide for future research on applying LLMs to automatic dialogue evaluation and the contributions are listed as follows:

- We conduct a comprehensive analysis of the multi-dimensional evaluation ability of 30 recent LLMs at both

[1] 6 turn-level and 6 dialogue-level respectively.

[2] Dimension refers to the quality aspect that is usually assessed in the human evaluation process, such as relevance at the turn level and dialogue coherence at the dialogue level.

*corresponding author.

turn and dialogue levels. Specifically, we evaluate coherence, engagingness, diversity, informativeness, and overall quality at the dialogue level. For turn-level evaluation, we assess context relevance, understandability, interestingness, specificity, and overall quality.

- Such a comprehensive assessment is impossible without the availability of large-scale meta-evaluation datasets with annotations. Hence, we complement 12 existing meta-evaluation datasets by providing annotations that were previously unavailable. The datasets together with the full annotations will be made publicly available for researchers and practitioners to benchmark their new evaluation metrics.
- Besides correlation analysis, we introduce a series of adversarial perturbation strategies to reduce the response or dialogue quality along various dimensions. In this way, we can probe the robustness of the LLMs, which has not been explored in existing works.
- Lastly, we study the impact of different ensemble strategies on the dialogue evaluation performance, including dimension-level and model-level ensembles.

Preliminaries

Meta-Evaluation

Datasets We adopt 12 meta-evaluation datasets in our analysis comprising 6 at the turn level and another 6 at the dialogue level. Table 1 summarizes the dataset details. For both turn-level and dialogue-level analysis, we evaluate five different quality aspects/dimensions respectively. At the turn level, we assess context relevance (rel), understandability (und), specificity (spe), interestingness (int), and overall response quality (ovr) while at the dialogue level, we evaluate coherence (coh), engagingness (eng), informativeness (inf), diversity (div), and overall dialogue quality (ovr).

Fill Up Missing Annotations With GPT-4 To save costs and speed up the annotation process, we perform the necessary annotations with GPT-4 instead of using crowdworkers. The motivation is that existing works show that GPT-4 can achieve human-level judgment while being more scalable and less expensive (Zheng et al. 2023; Gilardi, Alizadeh, and Kubli 2023; Liu et al. 2023). It serves as a supplementary tool to human evaluation, especially when considering the high costs and the reproducibility issues of human annotations in open-ended generation tasks (Karpinska, Akoury, and Iyyer 2021). Note that GPT-4 annotations may contain biases and future works can explore automatic and manual ways to mitigate the biases.

Figure 1 illustrates how we prompt GPT-4 to perform the annotation task. When calling the GPT-4 API, we set the temperature and top_p to 0.7 and 0.95 respectively. The annotation process is repeated five times to mimic the scenario of having multiple crowd workers annotate each data instance. The inter-annotator or inter-round agreement for GPT-4 is derived by averaging the pairwise Pearson correlations between the annotation scores from any two annotation rounds. In general, a good inter-annotator or inter-round agreement exceeding 0.65 is observed when annotating the missing dimensions of different datasets.

```
### Dialogues:
[Here is the input dialogue for annotation]

## Instruction:
Rate the coherence, engagingness, diversity, informativeness, and overall quality of the input dialogue on a scale of 1 to 5 and just output the corresponding ratings.

### Output Format:
coherence - x
engagingness - x
diversity - x
informativeness - x
overall - x

### Your Response:
[Here is GPT-4's output]
```

```
### Context:
[Here is the dialogue context]

### Response:
[Here is the input response for annotation]

### Instruction:
Rate the context relevance, specificity, interestingness, understandability, and overall quality of the response on a scale of 1 to 5 and just output the corresponding ratings.

### Output Format:
relevance - x
specificity - x
interestingness - x
understandability - x
overall - x

### Your Response:
[Here is GPT-4's output]
```

Figure 1: The instruction template for prompting GPT-4 to annotate both dialogue-level (top) and turn-level (bottom) data. For our meta-evaluation of the proprietary models including ChatGPT and Palm-2 Bison, we also use this instruction template.

Meta-Evaluation Metrics The reliability of LLMs as automatic evaluators is assessed by computing how well their evaluation scores (s_{llm}^{dim}) correlate with the corresponding human (s_{human}^{dim}) or GPT-4 judgment (s_{gpt4}^{dim}) with a correlation function g for a specific dimension. s_{human}^{dim} is derived by averaging the annotations of multiple annotators while s_{gpt4}^{dim} is obtained by averaging the annotations from the multiple annotation rounds. We adopt the commonly-used Pear-

Turn-Level Datasets	#Data	#Utt	Doc Len	IAA Range	Reused Annotations	Missing Annotations
Persona-USR (2020b)	300	9.3	98.4 / 12.0	0.3 ~ 0.7	rel, int, und, ovr	spe
Persona-Zhao (2020)	900	5.1	48.8 / 11.5	> 0.7	ovr	rel, int, und, spe
DailyDialog-Zhao (2020)	900	4.7	47.5 / 11.0	> 0.7	rel, ovr	int, und, spe
Topical-USR (2020b)	360	11.2	236.3 / 22.4	0.5 ~ 0.7	rel, int, und, ovr	spe
FED-Turn (2020a)	375	10.4	87.3 / 13.3	0.5 ~ 0.8	rel, int, spe, und, ovr	-
ConTurE-Turn (2022)	1066	3.8	21.7 / 11.0	~ 0.3	ovr	rel, int, und, spe
Dialogue-Level Datasets	**#Data**	**#Utt**	**Doc Len**	**IAA Range**	**Reused Annotations**	**Missing Annotations**
IEval-Dial (2022)	500	6.0	74.4	-	-	coh, eng, inf, div, ovr
Persona-See (2019)	500	12.0	91.2	-	-	coh, eng, inf, div, ovr
Reliable-Dial (2022)	500	21.2	178.1	-	-	coh, eng, inf, div, ovr
ConTurE-Dial (2022b)	119	17.9	153.9	-	-	coh, eng, inf, div, ovr
FED-Dial (2020a)	125	12.7	116.8	0.7 ~ 0.8	coh, eng, inf, div, ovr	-
Human-Eval (2022)	286	12.0	139.2	-	-	coh, eng, inf, div, ovr

Table 1: Details of the meta-evaluation datasets. The "Reused" columns indicate the dimensions with available human-annotated scores. The "Missing" column denotes the dimensions that need to be annotated by us. The doc length is the average #words per context/response for turn-level datasets and dialogue for dialogue-level datasets. The IAA range shows the range of inter-annotator agreements of available human annotations. #Utt refers to the number of context utterances and dialogue utterances for the turn-level and dialogue-level datasets respectively.

son (ρ) measure as g.

Large Language Models

30 LLMs comprising 28 open-source and 2 proprietary LLMs are examined. The proprietary LLMs are OpenAI ChatGPT[3] (gpt-3.5-turbo) and Google Palm-2 Bison (text-bison-001). The performance of OpenAI GPT-4 on human-annotated data is also reported. The 28 open-source LLMs can be grouped into two categories, the vanilla foundation models and the instruction-tuned models. The foundation models include different variants of Meta LLaMA (Touvron et al. 2023a,b), SalesForce XGen[4], TII-UAE Falcon (Almazrouei et al. 2023), MosaicML MPT[5], OpenLLaMA[6], Pythia (Biderman et al. 2023), and BLOOM (Scao et al. 2023). The instruction-tuned models are mainly derivatives of the aforementioned vanilla foundation models, such as Alpaca (Taori et al. 2023), Vicuna (Chiang et al. 2023), Tulu (Wang et al. 2023), and Chimera (Chen et al. 2023b). They are finetuned to mimic the abilities of proprietary LLMs, such as ChatGPT and GPT-4. Alignment techniques, such as instruction-based supervised finetuning (SFT) and reinforcement learning from human feedback (RLHF), are applied to align these models with humans' general task-solving abilities. We refer interested readers to their respective technical reports for the details of these LLMs.

Dialogue Evaluation with LLMs

As we cannot obtain the output probabilities of the proprietary models at the time of paper preparation, we adopt the explicit scoring procedure to directly prompt them to produce multi-dimensional ratings of dialogues or responses. ChatGPT, Palm-2 Bison, and GPT-4 share the same instruction template outlined in Figure 1. Due to their strong instruction-following abilities, we can easily extract the scores with matching heuristics. The rare erroneous cases are manually fixed. We repeat the scoring process of each proprietary model five times and apply the average score of the five runs as the corresponding s_{llm}^{dim}.

We do not apply the same approach for obtaining ratings from other open-source LLMs because their instruction-following abilities are weaker than the proprietary models. Sometimes, their generation becomes intractable. Instead, we follow Gupta et al. (2022) by applying an implicit scoring procedure. More specifically, given an instruction prompt input, we concentrate on the output probabili-

```
### Context:
[Here is the dialogue context]

### Response:
[Here is the input response for evaluation]

### Instruction:
Above is a dialogue context and the
corresponding response.

Question: Is the response relevant to the context?

### Your Answer:
[Here is LLM's output in terms of "Yes" or "No"]
```

Figure 2: An example for prompting open-source LLMs to evaluate the contextual relevance of the input response.

[3]We use the 2023-03-15-preview version on Azure.

[4]https://blog.salesforceairesearch.com/xgen/

[5]https://huggingface.co/mosaicml/mpt-7b

[6]https://github.com/openlm-research/open_llama

ties related to the label words "Yes" and "No" as generated by the LLM. Then, we normalize the probability of "Yes" as $P(Yes) = P(Yes)/(P(Yes) + P(No))$ and $P(Yes)$ serves as the corresponding s_{llm}^{dim}. Greedy decoding is applied such that the generation process of the LLMs is deterministic. Figure 2 showcases an example to prompt open-source LLMs to evaluate response relevance[7]. For different LLMs, we adapt the instruction template to match that used in their instruction-tuning process.

Multi-Dimensional Correlation Analysis

We report dimension-wise Pearson correlation scores averaged across either the turn-level or the dialogue-level datasets[8] in Table 2.

Proprietary vs Open-Source Models It can be observed that ChatGPT and Palm-2 Bison are among the top 5 models across all the dimensions at both turn and dialogue levels. On average, they ranked first and second respectively. Out of the 28 open-source models available, Tulu-13B (Wang et al. 2023), Chimera-inst-chat-13B (Chen et al. 2023b), and Baize-v2-13B (Xu et al. 2023a) are distinguished as the top three performers at the turn-level. When it comes to dialogue-level performance, the top three models are Chimera-inst-chat-13B, Tulu-13B, and WizardLM-13B-V1.2 (Xu et al. 2023b). A significant gap can be observed between the proprietary models and the best open-source model. For example, at the turn level, ChatGPT outperforms Tulu-13B by 11% on average while at the dialogue level, ChatGPT outperforms Chimera-inst-chat-13B by 7.7%. The observations suggest the importance of the model scale and the quality of instruction data. The proprietary models are much larger than other open-source LLMs. They are trained on more sophisticated human-annotated instruction data while the open-source models are mainly instruction-tuned on data distilled from proprietary models.

Instruction-Tuned vs Vanilla Models We can also observe that instruction-tuned variants generally outperform their corresponding vanilla backbone models. For instance, Alpaca-7B surpasses LLaMA-7B by a significant 25.1% on average at the turn level and 36.4% on average at the dialogue level. Similarly, Tulu-13B beats LLaMA-13B by a notable 9.1% and 18.2% on average. The observations showcase that alignment techniques, such as instruction-based supervised finetuning, can greatly enhance the dialogue understanding capabilities of LLMs, thereby making them more useful for automatic dialogue evaluation.

LLaMA vs Other Open-Source Families Among 7B vanilla models, LLaMA-2-7B tops the list at the turn level with an average of 0.200, while XGen-8K-7B leads at the dialogue level with an average of 0.404. For the 13B vanilla LLMs, LLaMA-13B stands out at the turn level with a 0.353 score on average, and OpenLLaMA-13 takes the lead at the dialogue level with an average score of 0.397. We can also observe a large performance variation among the vanilla models. For example, the gap between LLaMA-2-13B and BLOOM-7B is more than 25% at the turn level. This can be attributed to the differences in their model sizes, pretraining corpora, and training strategies.

In the 7B instruction-tuned category, Alpaca-7B and Chimera-inst-chat-7B stand out as the top two performers across both turn and dialogue levels, both of which are derivatives of LLaMA-7B. Given that LLaMA-13B serves as a strong foundational model for automated dialogue evaluation, it's understandable why its instruction-tuned variants, including Tulu-13B, Chimera-inst-chat-13B, Baize-v2-13B, and WizardLM-13B-V1.2, are among the best open-source models. With the observations, we can conclude that by far, models in the LLaMA family are stronger automatic dialogue evaluators than other open-source LLMs.

Impact of Instruction Data Even among the models with the same number of parameters and the same backbone, the performance varies greatly. For instance, Tulu-13B outperforms Vicuna-13B by 0.148 and 0.149 at turn and dialogue levels respectively in terms of the average Pearson correlations. The common attribute shared by Tulu-13B and Chimera-inst-chat-13B is that they use a more diverse set of instruction data than other models. Their instruction data come from different sources. For instance, the training data of Tulu consist of collections of NLP datasets with human-written instructions like FLAN V2 (Chung et al. 2022) and CoT (Wei et al. 2022b), collections of human-written instruction data from scratch, such as Dolly[9], and data mixture that is distilled from proprietary models like text-Davinci-003, ChatGPT, and GPT-4. Thus, it's evident that the diversity and quality of instruction data can significantly influence the model performance, emphasizing the importance of sourcing diverse datasets that better align with the task when finetuning our LLM-based evaluators.

Performance Across Dimensions Generally, the majority of models excel in areas like relevance, coherence, engagement, and overall quality. However, their performance diminishes slightly when assessing interestingness and understandability at the turn level, and diversity at the dialogue level. Future explorations into leveraging LLMs for automatic dialogue evaluation might benefit from designing objectives that target these specific dimensions.

Performance of GPT-4 The turn-level and dialogue-level evaluation abilities of GPT-4 are compared to the respective top-5 LLMs. The results are presented in Table 3. Note that Table 3 cannot be directly matched to Table 2 because we only use the human-annotated data for correlation computation here while in the previous section, we conduct correlation analysis on the full data, which contain both human annotations and GPT-4 annotations. From Table 3, we can see that GPT-4 achieves the best correlations with human

[7] All dimension-specific questions for prompting the open-source LLMs can be found in the Appendix.

[8] For example, a model's performance on a dimension at the turn level is derived by averaging the correlations obtained across the six turn-level meta-evaluation datasets w.r.t. that dimension.

[9] https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

	Turn-Level						Dialogue-Level					
Models	**Rel**	**Spe**	**Int**	**Und**	**Ovr**	**Avg**	**Coh**	**Eng**	**Div**	**Inf**	**Ovr**	**Avg**
LLaMA-7B	0.114	0.056	0.073	0.113	0.018	0.075	0.341	0.142	0.049	0.166	0.109	0.161
LLaMA-13B	0.442	0.240	0.333	0.348	0.402	0.353	0.522	0.425	0.237	0.324	0.404	0.382
LLaMA-2-7B	0.241	0.187	0.178	0.134	0.259	0.200	0.225	0.227	0.044	0.261	0.292	0.210
LLaMA-2-13B	0.303	0.122	0.161	0.269	0.225	0.216	0.318	0.362	-0.068	0.343	0.417	0.274
XGen-8K-7B	0.214	0.152	0.173	0.119	0.218	0.175	0.419	0.431	0.325	0.387	0.456	0.404
Falcon-7B	0.133	0.061	0.174	0.101	0.104	0.114	0.366	0.476	0.329	0.341	0.413	0.385
MPT-8K-7B	0.209	0.048	0.123	0.197	0.172	0.150	0.126	0.221	0.112	0.187	0.088	0.147
OpenLLaMA-7B	0.158	0.008	0.105	0.154	0.184	0.122	0.402	0.452	0.270	0.336	0.429	0.378
OpenLLaMA-13B	0.231	0.112	0.175	0.184	0.226	0.186	0.425	0.506	0.239	0.342	0.471	0.397
Pythia-7B	0.145	0.097	0.083	0.106	0.146	0.115	0.079	0.204	0.026	0.094	0.220	0.124
BLOOM-7B	0.082	0.021	0.124	0.099	0.122	0.089	0.224	0.271	0.137	0.217	0.277	0.225
LLaMA-2-Chat-7B	0.434	0.191	0.243	0.251	0.240	0.272	0.534	0.485	0.375	0.447	0.498	0.468
LLaMA-2-Chat-13B	0.559	0.353	0.207	0.320	0.380	0.364	0.644	0.610	0.228	0.514	0.613	0.522
Alpaca-7B	0.430	0.293	0.253	0.267	0.386	0.326	0.586	0.624	0.372	0.477	0.566	0.525
Vicuna-7B	0.368	0.096	0.221	0.100	0.219	0.201	0.490	0.468	0.279	0.488	0.482	0.441
Vicuna-13B	0.400	0.318	0.229	0.224	0.309	0.296	0.515	0.420	0.306	0.419	0.417	0.415
Falcon-Ins-7B	0.272	0.152	0.293	0.179	0.246	0.228	0.504	0.513	0.375	0.448	0.500	0.468
Tulu-13B	0.585	0.427	0.369	0.350	0.488	0.444	0.659	0.661	0.326	0.518	0.657	0.564
Chimera-7B	0.489	0.276	0.373	0.309	0.368	0.363	0.563	0.599	0.439	0.525	0.607	0.547
Chimera-13B	0.547	0.449	0.377	0.366	0.404	0.428	0.582	0.671	0.432	0.585	0.563	0.567
Phoenix-7B	0.314	0.101	0.291	0.258	0.234	0.240	0.480	0.493	0.146	0.334	0.416	0.374
Oasst-sft-Pythia-12B	0.144	0.028	0.203	0.132	0.110	0.123	0.386	0.358	0.236	0.346	0.423	0.350
Baize-v2-13B	0.477	0.350	0.333	0.353	0.337	0.370	0.568	0.544	0.397	0.482	0.469	0.492
Dolly-v2-12B	0.030	-0.009	0.061	-0.004	0.020	0.020	0.182	0.238	0.071	0.139	0.105	0.147
MPT-8K-7B-Instruct	0.139	0.092	0.176	0.095	0.099	0.120	0.321	0.352	0.316	0.308	0.315	0.322
XGen-8K-7B-Inst	0.272	0.293	0.267	0.145	0.265	0.248	0.502	0.515	0.308	0.417	0.506	0.450
ChatGLM-v2-6B	0.368	0.267	0.191	0.181	0.184	0.238	0.214	0.236	0.262	0.248	0.359	0.264
WizardLM-13B-V1.2	0.463	0.422	0.390	0.245	0.314	0.367	0.572	0.580	0.455	0.513	0.632	0.550
Palm-2 (text-bison-001)	0.666	0.563	0.454	0.422	0.601	0.541	0.649	0.674	0.473	0.557	0.674	0.605
ChatGPT (gpt-3.5-turbo)	0.595	0.578	0.518	0.536	0.542	0.554	0.724	0.705	0.516	0.568	0.707	0.644

Table 2: Dimension-specific Pearson correlation scores of different LLMs on fully-annotated data. The top-five scores in each dimension are underlined. 7B refers to 7 billion number of parameters.

evaluation in almost all the dimensions, except for specificity at the turn level and diversity at the dialogue level. It performs exceptionally well for relevance, coherence, and overall quality. On average, GPT-4 outperforms the second-best LLM (ChatGPT) by a large absolute margin of 0.085 and 0.042 at turn and dialogue levels respectively. The observations justify our using GPT-4 to complete the missing annotations in the datasets. However, we should note that even the powerful GPT-4 model cannot reach perfect correlations (> 0.8) on average suggesting that automatic dialogue evaluation remains an open problem. Future research should continue enhancing the conversation understanding capabilities of the language models.

Deviation Between GPT-4 and Human Preferences We analyze how much the assessment of the LLMs based on GPT-4 annotations deviates from that based on human annotations. Specifically, we compare the ranking results of the 32 LLMs evaluated by GPT-4 vs those assessed by human annotations. The deviation is quantified by the Spearman correlation between the two ranking lists. A greater Spearman correlation indicates smaller deviations in the model rankings. We perform the analysis on the FED dataset (Mehri and Eskenazi 2020a), which contains human annotations for all the dimensions. As shown in Table 4, the results reveal minimal deviation of GPT-4's evaluation from human evaluation (>0.85 agreements) in all dimensions except response specificity.

Ensemble Analysis

In this section, we delve into two straightforward ensemble strategies. The first approach involves averaging the scores of each LLM assigned to different dimensions[10] to see if this average provides a stronger correlation with overall human evaluations than directly prompting the LLM to assess the overall quality. We refer to this method as the "dimension-wise ensemble". The second strategy entails averaging scores from multiple LLMs for a given dimension, allowing us to determine if this method can match the per-

[10]At turn level, the relevance, specificity, understandability, and interestingness scores are averaged to derive an overall score while at the dialogue level, the coherence, engagingness, diversity, and informativeness scores are averaged.

Turn-Level						
	Rel	**Spe**	**Int**	**Und**	**Ovr**	**Avg**
Baize	0.449	0.147	0.302	0.290	0.337	0.305
Tulu	0.544	0.193	0.254	0.324	0.488	0.361
Chimera	0.507	0.234	0.312	0.316	0.404	0.354
Palm-2	0.616	0.317	0.343	0.384	0.601	0.452
ChatGPT	0.576	**0.408**	0.446	0.424	0.542	0.479
GPT-4	**0.704**	0.342	**0.538**	**0.558**	**0.677**	**0.564**
Dialogue-Level						
	Coh	**Eng**	**Div**	**Inf**	**Ovr**	**Avg**
Tulu	0.668	0.629	0.414	0.584	0.681	0.595
Chimera	0.595	0.628	0.507	0.609	0.525	0.573
WizardLM	0.536	0.522	0.477	0.540	0.605	0.536
Palm-2	0.584	0.633	0.550	0.604	0.614	0.597
ChatGPT	0.650	0.647	**0.551**	0.570	0.715	0.627
GPT-4	**0.760**	**0.689**	0.534	**0.620**	**0.744**	**0.669**

Table 3: Dimension-specific Pearson correlation scores of GPT-4 vs other strong LLMs on human-annotated data. The best score in each dimension is highlighted in bold. All the open-source LLMs are the 13B variants.

	Rel	**Spe**	**Int**	**Und**	**Ovr**
FED-Turn	0.950	0.551	0.887	0.869	0.971
	Coh	**Eng**	**Div**	**Inf**	**Ovr**
FED-Dial	0.861	0.900	0.883	0.869	0.945

Table 4: Spearman of model rankings evaluated by GPT-4 ratings vs evaluated by human ratings.

formance of the stronger proprietary LLMs. We call this approach the "model-wise ensemble".

	Turn Level		**Dialogue Level**	
Model	**Ensemble**	**Direct**	**Ensemble**	**Direct**
Tulu	0.492	0.488	0.656	0.657
Chimera	0.476	0.404	0.680	0.563
Baize	0.417	0.337	0.578	0.469
WizardLM	0.408	0.314	0.598	0.632
Palm-2	0.602	0.601	0.693	0.674
ChatGPT	0.564	0.542	0.709	0.707

Table 5: Dimension-wise ensemble results of difference models for the overall quality evaluation. All the open-source LLMs are the 13B variants.

Dimension-Wise Ensemble We limit the analysis to Tulu-13B, Chimera-inst-chat-13B, Baize-v2-13B, WizardLM-13B-V1.2, Palm-2 Bison, and ChatGPT, which have strong multi-dimensional evaluation performance. Table 5 presents performance comparisons between the dimension-wise ensemble and the direct prompting approaches for evaluating the overall response or dialogue quality of each model. We

Turn Level			
Dimensions	**Palm-2 Bison**	**ChatGPT**	**Ensemble**
Rel	0.666	0.595	0.632
Int	0.454	0.518	0.465
Und	0.422	0.536	0.407
Spe	0.563	0.578	0.487
Ovr	0.601	0.542	0.491
Average	0.541	0.554	0.496
Dialogue Level			
Coh	0.649	0.724	0.700
Eng	0.674	0.705	0.725
Div	0.473	0.516	0.492
Inf	0.557	0.568	0.610
Ovr	0.674	0.707	0.686
Average	0.605	0.644	0.643

Table 6: Performance of the ensemble vs proprietary models for each dimension at the turn and dialogue levels.

can observe that the ensemble approach yields strong correlations with overall human judgments in general. Especially for Chimera-inst-chat-13B and Baize-v2-13B at the dialogue level, the ensemble approach provides gains of more than 10% than direct prompting. We also observe that the scores assigned by Tulu-13B, ChatGPT, and Palm-2 Bison to different dimensions are highly similar while Chimera-inst-chat-13B and Baize-v2-13B provide more diverse scores when evaluating different dimensions. This may explain why the ensemble of different dimension-specific scores of Chimera-inst-chat-13B and Baize-v2-13B leads to more significant improvements than other LLMs.

Model-Wise Ensemble For the model-wise ensemble, we average the scores of the top 3 open-source models for each dimension as indicated in Table 2. We can observe that the ensemble achieves comparable performance to ChatGPT and outperforms Palm-2 Bison at the dialogue level. At the turn level, the ensemble's performance is worse than that of both ChatGPT and Palm-2 Bison for dimensions other than relevance and interestingness. The simple ensemble showcases the potential benefits of combining multiple models to boost evaluation performance. Future research might delve deeper into optimal ways of ensembling, such as how to best combine models, which models to include in the ensemble, and how to weigh individual model outputs.

Robustness of the LLM Evaluators

Motivated by prior works on applying perturbation strategies for metric robustness probing (Sai et al. 2021; Khalid and Lee 2022), we analyze the robustness of LLM evaluators with a series of adversarial strategies and table 7 presents the data sources and statistics of our perturbation test suit[11].

[11]The detailed definitions of each perturbation strategy are outlined in the Appendix.

	Level	Source	#Data	Dims	θ
RR	Turn	DD & PC	1894	Rel	0.3
RP	Turn	DD & PC	2919	Rel	0.2
RNE	Turn	DD & PC	712	Rel	0.2
Con	Turn	DD & PC	1094	Rel	0.2
Rep	Turn	DD & PC	1993	Und	0.2
UP	Turn	DD & PC	369	Und	0.2
Dul	Turn	DD & PC	2811	Int/Spe	0.2
OS	Dial	FED-Dial	200	Coh	0.2
UR	Dial	FED-Dial	200	Coh	0.1
SC	Dial	DECODE	200	Eng	0.1
UD	Dial	FED-Dial	200	Eng	0.1
RO	Dial	FED-Dial	200	Eng	0.1
GU	Dial	FED-Dial	200	Eng/Inf	0.2
CR	Dial	FED-Dial	200	Inf	0.2

Table 7: Adversarial Perturbation Data Statistics. DD & PC refer to DailyDialog & PersonaChat Respectively. RR, RP, RNE, Con, Rep, UP, Dul, OS, UR, SC, UD, RO, GU, and CR refer to Random Response, Replace Pronoun, Replace Named Entity, Contradiction, Repetition, Unnatural Paraphrase, Dullness, Order Shuffle, Utterance Replacement, Self-Contradiction, Utterance Duplication, Repeating Others, Generic Utterance, and Content Reduction respectively.

We focus on negative adversarial perturbations that diminish the quality of the original response or dialogue. Formally, let q_{dim} represent the score assigned by LLMs for a high-quality dialogue/response specific to a certain dimension. Conversely, p_{dim} denotes the score given by the LLMs when a particular negative perturbation, targeting that dimension, is applied to the response or dialogue. The LLMs' robustness against that particular perturbation can be determined by $\mathrm{R} = \frac{1}{N}\sum y$. N is the number of data instances generated with that particular perturbation and y is calculated as

$$y = \begin{cases} 1 & \text{if } q_{dim} - p_{dim} > \theta \\ 0 & \text{otherwise} \end{cases}$$

where θ is a positive threshold value determining the extent of quality reduction introduced by the perturbation. It's worth noting that some perturbation strategies result in greater quality degradation than others. A larger R signifies greater robustness. As that robustness analysis is only meaningful when applied to strong automatic metrics, we limit the analysis to ChatGPT, Palm-2 Bison, Tulu-13B, Chimera-inst-chat-13B, Baize-v2-13B, WizardLM-13B-V1.2, and LLaMA-2-Chat-13B.

Impact of Different Strategies For the "random response" perturbation at the turn level, Palm-2 Bison, the top-performing LLM, scores the original response over 0.3 higher than the perturbed response in 62.3% of cases. The other LLMs manage this less than half the time. Notably, Baize-v2-13B and WizardLM-13B-V1.2 frequently fail to recognize the "random response" perturbation. The "replace pronoun" and "replace named entity" perturbations are more challenging than "random response" as the robustness ratio, R, of all the LLMs drops to below 40%. These two strategies require a more fine-grained understanding of the semantics of the dialogue context and the response. For the "contradiction" perturbation, Palm-2 Bison and three other open-source LLMs, Tulu-13B, Chimera-13B, and LLaMA-2-Chat-13B achieve a robustness ratio of more than 50%, significantly outperforming ChatGPT.

In general, all the LLMs perform poorly on "repetition", "unnatural paraphrase", and "dullness" perturbations. The observation is in line with their weaker correlations in the interestingness and understandability dimensions than in the relevance dimension as shown in Table 2. Notably, ChatGPT performs much better than other LLMs in handling the "unnatural paraphrase" perturbation.

The proprietary models are more adept at handling the perturbations targeting dialogue-level coherence than the open-source ones. For example, ChatGPT achieves a robustness ratio of 0.535 and 0.730 for "Order Shuffle" and "Utterance Replacement" respectively. Most LLMs struggle with perturbations targeting dialogue-level engagingness and informativeness, such as "Utterance Duplication", "Repeating Others", and "Generic Utterance", except for LLaMA-2-Chat-13B, suggesting that future research on LLMs for automatic dialogue evaluation should prioritize a deeper comprehension of multi-turn dialogues, such as the depth of topic engagement, speaker sentiments, interactivity & proactivity among the speakers, etc., which goes beyond mere surface-level coherence assessments.

Proprietary vs Open-Source LLMs As illustrated in Table 8, Palm-2 Bison exhibits superior robustness at the turn level, whereas LLaMA-2-Chat-13B performs the best at the dialogue level. Palm-2 Bison can handle most of the adversarial perturbations. Notably, it excels in identifying declines in response relevance, interestingness, specificity, and dialogue coherence. ChatGPT is capable of dealing with adversarial perturbations targeting response understandability and dialogue coherence. Surprisingly, it performs poorly in handling other types of perturbations. Among the open-source LLMs, Tulu-13B and LLaMA-2-Chat-13B perform similarly on average at the turn level. They are better than the other three open-source models. At the dialogue level, LLaMA-2-Chat-13B performs exceptionally well and outperforms Palm-2 Bison by 3.7% and ChatGPT by 16.4% on average. It demonstrates consistent strength in dealing with all the perturbations except those targeting response understandability and dialogue coherence. In contrast, both Baize-13B and WizardLM-13B struggle to handle negative perturbations at both the turn and dialogue levels.

In general, we can observe that none of the LLMs are robust against all the adversarial perturbations. The variance in performances underscores the inherent complexity of dialogues and the challenges in creating a universally robust automatic dialogue evaluator. Future work should prioritize building on these findings to improve the robustness and adaptability of LLMs across diverse perturbations.

Turn Level							
Perturbations	**Palm-2**	**ChatGPT**	**Tulu**	**Chimera**	**Baize**	**WizardLM**	**L2-Chat**
Random Response (↓Rel)	**0.623**	0.391	0.419	0.307	0.043	0.125	0.429
Replace Pronoun (↓Rel)	**0.388**	0.189	0.156	0.249	0.041	0.083	0.272
Replace Named Entity (↓Rel)	**0.391**	0.160	0.274	0.301	0.097	0.156	0.337
Contradiction (↓Rel)	0.529	0.102	0.523	**0.535**	0.214	0.282	0.529
Repetition (↓Und)	0.127	**0.139**	0.060	0.068	0.003	0.002	0.016
Unnatural Paraphrase (↓Und)	0.307	**0.451**	0.182	0.252	0.014	0.008	0.065
Dullness (↓Int)	**0.433**	0.107	0.260	0.122	0.000	0.128	0.245
Dullness (↓Spe)	**0.411**	0.264	0.191	0.092	0.000	0.023	0.166
Average	**0.401**	0.225	0.258	0.241	0.052	0.101	0.257
Dialogue Level							
Order Shuffle (↓Coh)	0.440	**0.535**	0.135	0.000	0.000	0.000	0.155
Utterance Replacement (↓Coh)	0.490	**0.730**	0.130	0.000	0.000	0.000	0.075
Generic Utterance (↓Eng)	0.295	0.145	0.110	0.085	0.000	0.050	**0.620**
Self-Contradiction (↓Eng)	0.365	0.075	**0.645**	0.455	0.050	0.155	0.535
Utterance Duplication (↓Eng)	0.230	0.095	0.235	0.150	0.000	0.030	**0.665**
Repeating Others (↓Eng)	0.285	0.090	0.120	0.255	0.025	0.060	**0.445**
Content Reduction (↓Inf)	**0.355**	0.040	0.235	0.170	0.000	0.055	0.230
Generic Utterance (↓Inf)	0.330	0.065	0.210	0.155	0.000	0.050	**0.365**
Average	0.349	0.222	0.228	0.159	0.009	0.050	**0.386**

Table 8: Percentage of cases when the LLMs successfully detect a perturbation (Robustness Ratio R). The best ratio for each perturbation is highlighted in bold. All the open-source LLMs are the 13B variants and L2-Chat refers to LLaMA-2-Chat-13B.

Related Work

Huynh et al. (2023) conduct a comprehensive analysis of the dialogue evaluation capability of LLMs with varying model types, sizes, choices of training data, etc. Their study is limited to correlation analysis on a few LLMs, which are mainly vanilla models without instruction-tuning, such as BLOOM (Scao et al. 2023) and OPT (Zhang et al. 2022b). With the increasing popularity of API-based proprietary instruction-following LLMs, such as OpenAI's ChatGPT and Anthropic Claude (Bai et al. 2022). Several recent works (Chen et al. 2023a; Liu et al. 2023; Lin and Chen 2023) study the dialogue evaluation capability of these LLMs via prompting and show that such LLMs exhibit strong zero-shot correlations with human evaluation. Yet, their study is constrained in terms of the number of meta-evaluation datasets, mode of assessment, and number of LLMs examined. Our research addresses these constraints, providing profound insights into the use of various LLMs for automatic dialogue evaluation. We conduct an extensive multi-dimensional correlation analysis involving 30 of the latest popular LLMs and introduce a range of perturbation strategies to assess their robustness.

Conclusion

In summary, we've analyzed the multi-dimensional evaluation abilities of 30 recent LLMs, covering coherence, engagingness, and more at dialogue and turn levels. To facilitate the analysis, we enrich 12 existing meta-evaluation datasets with new annotations, which will be publicly available for benchmarking new metrics. Beyond correlation analysis, we introduce various adversarial strategies to test LLM robustness, a perspective not explored in existing works. Lastly, we also examine the impact of dimension-wise and model-wise ensembles on dialogue evaluation in our work. The key insights are summarized as follows:

1. Instruction-tuned models align better with human evaluations than vanilla foundation models.
2. Proprietary models, especially GPT-4, have superior evaluation abilities compared to open-source LLMs.
3. Model size and instruction data are vital for evaluation. Only the ensemble of strong open-source models performs on par with ChatGPT and PaLM-2 Bison.
4. LLMs excel more in evaluating coherence, relevance, and overall quality than specificity and diversity. Using an ensemble of their dimension-specific scores aligns better with the overall human evaluations than a direct assessment of the overall quality.
5. None of the LLMs are robust against all the adversarial perturbations. Google's Palm-2 Bison achieves the best robustness at the turn level while Meta's LLaMA-2-Chat-13B (Touvron et al. 2023b) tops at the dialogue level.
6. LLMs show promise in automatic dialogue evaluation, but it's still an open problem, with even GPT-4 not excelling in all dimensions (achieve correlations > 0.8).

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work is supported by Human Robot Collaborative AI under its AME Programmatic Funding

Scheme (Project No. A18A2b0046), the National Natural Science Foundation of China (Grant No. 62271432, 62106038), Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002), and the Internal Project Fund from Shenzhen Research Institute of Big Data under Grant No. T00120220002. This work is also a result of the projects: ASTOUND (101071191 - HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) funded by the European Commission, BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union", and the Research Grants for Young Investigators from Universidad Politécnica de Madrid (GENIUS:APOYO-JOVENES-21-TAXTYC-32-K61X37) funded by Comunidad de Madrid.

References

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; et al. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Biderman, S.; Schoelkopf, H.; Anthony, Q.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; Skowron, A.; Sutawika, L.; and van der Wal, O. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *arXiv preprint arXiv: Arxiv-2304.01373*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; et al. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, Y.; Wang, R.; Jiang, H.; Shi, S.; and Xu, R. 2023a. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: A Preliminary Empirical Study. *arXiv preprint arXiv: Arxiv-2304.00723*.

Chen, Z.; Jiang, F.; Chen, J.; Wang, T.; Yu, F.; Chen, G.; Zhang, H.; Liang, J.; Zhang, C.; Zhang, Z.; Li, J.; Wan, X.; Wang, B.; and Li, H. 2023b. Phoenix: Democratizing ChatGPT across Languages. *arXiv preprint arXiv: Arxiv-2304.10453*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; et al. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; et al. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; et al. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv: Arxiv-2210.11416*.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2023. GPTScore: Evaluate as You Desire. arXiv:2302.04166.

Ghazarian, S.; Hedayatnia, B.; Papangelis, A.; Liu, Y.; and Hakkani-Tur, D. 2022. What is wrong with you?: Leveraging User Sentiment for Automatic Dialog Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2022*, 4194–4204. Dublin, Ireland: Association for Computational Linguistics.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *arXiv preprint arXiv: 2303.15056*.

Gupta, P.; Jiao, C.; Yeh, Y.-T.; Mehri, S.; Eskenazi, M.; and Bigham, J. 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*, 505–525. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Huynh, J.; Jiao, C.; Gupta, P.; Mehri, S.; Bajaj, P.; Chaudhary, V.; and Eskenazi, M. 2023. Understanding the Effectiveness of Very Large Language Models on Dialog Evaluation. In *The 13th International Workshop on Spoken Dialogue Systems Technology*.

Ji, T.; Graham, Y.; Jones, G.; Lyu, C.; and Liu, Q. 2022. Achieving Reliable Human Assessment of Open-Domain Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6416–6437. Dublin, Ireland: Association for Computational Linguistics.

Karpinska, M.; Akoury, N.; and Iyyer, M. 2021. The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1265–1285. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Khalid, B.; and Lee, S. 2022. Explaining Dialogue Evaluation Metrics using Adversarial Behavioral Analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5871–5883. Seattle, United States: Association for Computational Linguistics.

Lin, Y.-T.; and Chen, Y.-N. 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)*, 47–58. Toronto, Canada: Association for Computational Linguistics.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 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*, 2122–2132. Austin, Texas: Association for Computational Linguistics.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint arXiv: Arxiv-2303.16634*.

Mehri, S.; Choi, J.; D'Haro, L. F.; Deriu, J.; Eskenazi, M.; Gasic, M.; Georgila, K.; Hakkani-Tur, D.; Li, Z.; Rieser, V.;

Shaikh, S.; Traum, D.; Yeh, Y.-T.; Yu, Z.; Zhang, Y.; and Zhang, C. 2022a. Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges.

Mehri, S.; and Eskenazi, M. 2020a. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 225–235. 1st virtual meeting: Association for Computational Linguistics.

Mehri, S.; and Eskenazi, M. 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*, 681–707. Online: Association for Computational Linguistics.

Mehri, S.; Feng, Y.; Gordon, C.; Alavi, S. H.; Traum, D.; and Eskenazi, M. 2022b. Interactive Evaluation of Dialog Track at DSTC9. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5731–5738. Marseille, France: European Language Resources Association.

Ouyang, L.; and et al. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Sai, A. B.; Dixit, T.; Sheth, D. Y.; Mohan, S.; and Khapra, M. M. 2021. Perturbation CheckLists for Evaluating NLG Evaluation Metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7219–7234. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; et al. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100.

See, A.; Roller, S.; Kiela, D.; and Weston, J. 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)*, 1702–1723. Minneapolis, Minnesota: Association for Computational Linguistics.

Smith, E.; Hsu, O.; Qian, R.; Roller, S.; Boureau, Y.-L.; and Weston, J. 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*, 77–97. Dublin, Ireland: Association for Computational Linguistics.

Svikhnushina, E.; Filippova, A.; and Pu, P. 2022. iEval: Interactive Evaluation Framework for Open-Domain Empathetic Chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 419–431. Edinburgh, UK: Association for Computational Linguistics.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. *ARXIV*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; and other. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv: 2307.09288*.

Wang, Y.; Ivison, H.; Dasigi, P.; Hessel, J.; Khot, T.; Chandu, K. R.; Wadden, D.; MacMillan, K.; Smith, N. A.; Beltagy, I.; and Hajishirzi, H. 2023. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. *arXiv preprint arXiv: 2306.04751*.

Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; brian ichter; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022b. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Xu, C.; Guo, D.; Duan, N.; and McAuley, J. 2023a. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. *arXiv preprint arXiv: 2304.01196*.

Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; and Jiang, D. 2023b. WizardLM: Empowering Large Language Models to Follow Complex Instructions. *arXiv preprint arXiv: Arxiv-2304.12244*.

Yeh, Y.-T.; Eskenazi, M.; and Mehri, S. 2021. A Comprehensive Assessment of Dialog Evaluation Metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, 15–33. Online: Association for Computational Linguistics.

Zhang, C.; D'Haro, L. F.; Friedrichs, T.; and Li, H. 2022a. MDD-Eval: Self-Training on Augmented Data for Multi-Domain Dialogue Evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10): 11657–11666.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; et al. 2022b. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv: Arxiv-2205.01068*.

Zhao, T.; Lala, D.; and Kawahara, T. 2020. Designing Precise and Robust Dialogue Response Evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 26–33. Online: Association for Computational Linguistics.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.