Adversarial Evaluation of Dialogue System Metrics

Justin Vasselli and Taro Watanabe
Nara Institute of Science and Technology
{vasselli.justin_ray.vk4,taro}@is.naist.jp

Abstract

Effective dialogue system evaluation requires metrics that align with human judgment and are resilient against adversarial attacks. We present a benchmark to evaluate the robustness of evaluation metrics against adversarial attacks, including generic responses, ungrammatical replies, and context repetition. We find the metric with the highest correlation with human annotation, GPT-4, is effective against generic responses and context repetition, while metrics with lower human correlation like UniEval outperform GPT when countering ungrammatical responses. The proposed benchmark serves as a valuable tool for assessing the robustness of dialogue evaluation metrics.

1 Introduction

Dialogue systems have seen significant advancements with the use of Large Language Models (LLMs). However, evaluating the quality of these systems remains a complex task that heavily relies on human judgment. There are several factors that pose challenges during the evaluation process, including the myriad of valid possible responses to user inputs which makes reference-based metrics a poor choice [1].

Recognizing the limitations of reference-based metrics in accommodating equally valid responses, researchers have turned to reference-free metrics as a promising alternative [2, 3]. However, it' s possible that these metrics are not as reliable as we need them to be. For instance, [4] identified vulnerabilities in DialogRPT [3]. They discovered that adding the seemingly harmless "teacher:" prefix to each response had a significant impact on their system's performance in the BEA2023 shared task [5], resulting in a 5-place jump on the leaderboard. These findings emphasize the need to test the reliability of metrics like DialogRPT against subtle manipulations. They also raise concerns about the ability of evaluation methods to accurately assess dialogue systems' complexities.

To establish trust in reference-free metrics, it is important to test their resistance to attacks and ensure they align with human judgment. Effective evaluation metrics should not only reward responses that are engaging, relevant, and grammatical but also penalize generic and nonsensical replies. In this study, we undertake a comparative analysis of evaluation metrics, evaluating their correlation with human judgment using the DailyDialog subset benchmark [6]. This study pioneers an experimental framework for robustness evaluation, subjecting metrics like DialogRPT [3], UniEval [7], and GPT-3.5/4 to adversarial responses. Our investigation highlights strengths, weaknesses, and proposed enhancements for advancing dialogue system evaluation methodologies. The analysis contributes to the ongoing discourse on developing more robust and reliable evaluation methods.

2 Related works

2.1 Reference-free metrics

Evaluating dialogue systems poses unique challenges due to the varied nature of dialogue responses. Dialogue responses can differ significantly while remaining valid, making it insufficient to rely on a single reference evaluation [1]. However, it is not practical to cover all possible valid responses. Many newer methods of evaluation are reference-free, but these metrics don't always correlate well with human evaluation, and can be unreliable [5].

2.2 Adversarial robustness

Adversarial robustness of the evaluation metric is crucial, but underrepresented in the literature. [8] introduced the Evaluator Reliability Error score, which quantifies the deviation of an evaluator model from gold standard accuracy values across different scenarios. [9] investigated adversarial responses, including the removal of stopwords

and the provision of generic or irrelevant answers, to assess the performance of ADEM. Our work expands upon this analysis by incorporating additional adversarial elements, such as speaker tags, fixed generic responses, nonsensical outputs, and contextual repetitions. We evaluate these adversarial responses using more recent and fine-grained reference evaluation approaches.

3 Method

We tested different evaluation metrics looking at correlation with human evaluation and robustness against adversarial attacks.

3.1 Evaluation dataset

The DailyDialog subset [6][1)] comprises 100 conversations from DailyDialog, including ground-truth responses and 8 additional responses generated by different models and decoding methods. These responses have been human-annotated for content, grammaticality, and relevance.

The dataset was provided pretokenized and lowercased, which is not required for some of the evaluation methods we are assessing, so the dataset was detokenized and truecased using regular expressions as a preprocessing step.

3.2 Evaluation metrics

3.2.1 DialogRPT

DialogRPT is a collection of five models that have been fine-tuned on Reddit data. They have been trained to measure content related aspects such as if a responses is likely to be upvoted (updown), be replied to by many users (width), start a long discussion (depth). The final two models measure the relevance of the response (human-vs-random), and how natural the response sounds (human-vs-machine).

When measuring correlation, we try to find the closest mapping to the aspect scores from the DailyDialog subset. We used a combination of updown, depth, and width as the `content` score, human-vs-machine as the `grammar` score, and human-vs-random as the `relevance` score. The composite DialogRPT score is used as the `overall` score.

3.2.2 UniEval

UniEval [7] is a single model that has been trained to evaluate dialogue system responses on five different aspects. This T5 based model was finetuned on synthetic data to be able to answer yes/no questions about each aspect, and can be extended to new aspects with finetuning, or zero-shot by changing the provided prompt.

UniEval was trained on TopicalChat [10][2)] which evaluates the response given a fact in addition to the dialogue context. Because DailyDialog is not grounded, we made some zero-shot adjustments. Our altered UniEval uses a new prompt that does not refer to the fact for `content`, the original naturalness prompt for `grammar`, and the coherence prompt for `relevance`. As we are only using three of the original five sub-metrics due to the ungroundedness of the current test and to maintain simplicity, we rebalance the composite score to give more weight to engagingness and coherence. The composite score is: content $*$ 0.4 + grammar $*$ 0.2 + relevance $*$ 0.4.

See Appendix A for the ablation analysis of our altered UniEval metric.

3.2.3 GPT 3.5/4

Several studies have utilized OpenAI's GPT-3.5 or GPT-4 as an evaluation metric for natural language generation [11, 12, 13]. Both GPTScore [12] and G-Eval [11] use the deprecated log-probs feature that will be discontinued as of January 2024. For this reason, we opted to use a direct assessment that is unweighted by probabilities as explored in [13]. We experimented with a single prompt per metric, but found higher correlation with human annotation when using a combined prompt that describes all three sub-metrics (content, grammar, and relevance) and requests scores for each. The prompt additionally asks for an overall score, which is not used directly, but averaged along with the sub-metrics, to generate the composite score for the response. We use `gpt-3.5-turbo-1106` and `gpt-4-1106-preview`. The full prompt is provided in Appendix B.

3.3 Adversarial attacks

We design four categories of adversarial attacks to evaluate the vulnerability of the evaluation metrics:

Speaker tags We prepend the response with a speaker name as a prefix. Different speaker tags, such as "teacher," "agent," and "user," are tested, as DialogRPT already has a documented vulnerability in this area.

1) https://github.com/ZHAOTING/dialog-processing

2) http://shikib.com/usr

	Content		Grammar		Relevance		Overall	
Metrics	ρ	τ	ρ	τ	ρ	τ	ρ	τ
DialogRPT	*-0.013*	*-0.008*	-0.115	-0.082	0.233	0.162	*-0.024*	*0.016*
UniEval	0.387	0.273	0.170	0.117	0.535	0.381	0.285	0.198
UniEval-altered	0.448	0.322	0.170	0.117	0.535	0.381	0.491	0.347
GPT-3.5	0.369	0.302	0.526	0.437	0.648	0.525	0.637	0.484
GPT-4	**0.490**	**0.411**	**0.564**	**0.477**	**0.719**	**0.583**	**0.703**	**0.548**

Table 1 Turn level Spearman's ρ, and Kendall's τ correlations of different metrics on the Daily Dialog Subset. *Italicized* values are not statistically significant ($p > 0.05$). The highest value in each column is bolded.

Generic responses We replace the response with fixed generic phrases such as "Hello" and "I don't know." that are grammatical yet uninteresting. We also explore more complicated responses like "I don't know, what do you think?" and "fantastic! how are you?" that may be more engaging, but less likely to be contextually appropriate. Finally, we consider ungrammatical responses such as "I will do" and "I don't know, what do you think? I think."

Ungrammatical responses Following [9] we alter the ground truth response in various was to make it ungrammatical, including removing punctuation, removing stopwords, retaining only nouns, changing the order of tokens, and repeating random words with a probability of 0.2.

Context responses This category utilizes the context in the response by either repeating the last utterance or prefixing the response with a repetition of the last utterance in the context. As all of the evaluation methods contain at least one sub-metric dedicated to the relevance or dialogue level coherense of the response, usually trained with random responses from other dialogue histories as the negative sample, we hypothesis that many systems will mark word for word repetitions of previous utterances as highly relevant.

For each adversarial response, we calculate the success rate of the attack, which represents the frequency with which the adversarial response scores higher than or equal to the reference response.

As GPT-3.5 and GPT-4 result in many ties due to the discrete nature of the scores, it's important to note that we count a tie between an attack and the ground truth as a successful attack. If used as a reranking metric, a tie would still result in the possibility of the attack being chosen over a more desirable response, so any possibility is counted.

4 Results

4.1 Correlation with human evaluation

The DialogRPT composite score does not correlate with the overall human annotated score of the DailyDialog subset. There is a negative correlation between human-vs-machine and `grammar`, but a small significant correlation between human-vs-random and `relevance`.

GPT-4 has the highest Spearman's ρ and Kendall's τ on all metrics, but GPT-3.5 is also an improvement on most scores, except `content`, where UniEval remains competitive. For full results, please see Table 1

4.2 Vulnerability against adversarial attacks

DialogRPT is the most vulnerable to many of our attacks, but particularly to speaker tag attacks and randomizing the order of words in the utterance.

UniEval is incredibly robust against nearly all attacks. Only context repetition attacks fool the metric in the majority of cases. A perfect copy of the dialogue history can fool most metrics by appearing very relevant. As most relevance-based sub-metrics train the model to discriminate between ground truth responses and a ground truth response from a random other conversation in the dataset, they learn to reward similarity between the last utterance and the candidate response. Either the relevance metric or a content based metric should penalize utterances that offer nothing new to the conversation.

Both GPT-3.5 and 4 are more susceptible to the speaker tag attacks than UniEval. It's possible that they simply consider it a change of speaker. Interestingly, **teacher:** triggers lower scores than **agent:** or **user:**. Neither GPT is particularly sensitive to changes in punctuation, but GPT-3.5 seems to give more weight to the nouns in the response when scoring, as evidenced by the higher number of ungrammatical responses containing only the nouns of the

Attacks	DialogRPT	UniEval	GPT-3.5	GPT-4
Speaker tags				
teacher: prefix	0.99	**0.07**	0.38	0.17
agent: prefix	0.99	**0.08**	0.55	0.67
user: prefix	1.00	**0.07**	0.75	0.67
Generic Responses				
”Hello”	0.16	**0.00**	0.04	0.01
”Cucumber”	0.03	**0.00**	0.01	**0.00**
”I don’t know”	0.15	0.02	**0.01**	0.02
”I don’t know, what do you think?”	0.43	0.14	0.18	**0.03**
”I don’t know, what do you think? I think”	0.35	0.02	**0.01**	**0.01**
”I’m sorry, can you repeat?”	0.49	0.12	0.19	**0.03**
”I will do”	0.24	0.02	0.07	**0.01**
”fantastic! how are you?”	0.23	0.13	**0.03**	**0.03**
Ungrammatical Responses				
no punctuation	0.38	**0.12**	0.48	0.32
no stopwords	0.18	**0.06**	0.16	0.09
only nouns and verbs	0.02	**0.01**	0.10	0.06
only nouns	0.18	**0.03**	0.23	0.04
jumbled words	0.80	**0.01**	0.02	0.02
reversed words	0.15	0.02	0.02	**0.01**
repeat words	0.62	**0.00**	0.09	0.04
Context Repetition Responses				
previous utterance	0.19	0.67	0.60	**0.01**
previous utterance prefix	0.66	0.85	0.78	**0.08**

Table 2 The success rate of each attack against different evaluation metrics. The lower the number, the more resistant to attack. The best performing system for each attack is bolded.

utterance scoring at least as high as the ground truth. GPT-4 is able to notice the repetition present in the context repetition attacks much better than GPT-3.5. It’s possible that GPT-4 is better able to pay attention to the speaker of each utterance. See Table 2 for the vulnerability of each system against each type of adversarial response.

5 Conclusion and Future Work

We propose a novel stress test for evaluate dialogue response metrics, uncovering weaknesses not captured by assessing correlation with human judgment alone. The application of this adversarial test across diverse metrics has yielded insightful findings.

DialogRPT’s susceptibility to speaker tag attacks and context injection serves as a cautionary note, emphasizing the importance of evaluating metrics against a spectrum of potential vulnerabilities. UniEval, showcasing resilience against ungrammatical responses and speaker tag attacks, stands out as a robust metric, at times outperforming GPT-3.5/4, despite lower correlation scores. Notably, GPT-4 was the only metric that successfully penalized responses that repeated parts of the dialogue history.

Contrary to expectations, metrics that correlate more with human judgment are not robust against all types of attacks. Our findings show that attacks based on generic responses and repetitive context are deterred by high human judgment metrics like GPT-4. However, attacks involving speaker tags and ungrammatical responses are better countered by lower human judgment metrics, such as UniEval. This difference highlights the complexity of evaluating dialogue systems and stresses the need to comprehensively understand metric performance across different adversarial scenarios.

While our focus remains on the evaluation of responses given only a dialogue history, UniEval, GPT-3.5, and GPT-4 have all shown promise in grounded dialogue scenarios [11], opening avenues for future research. We leave developing attacks for grounded dialogue response evaluation metrics to future work.

References

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

[2] Liliang Ren, Mankeerat Sidhu, Qi Zeng, Revanth Gangi Reddy, Heng Ji, and ChengXiang Zhai. C-PMI: Conditional pointwise mutual information for turn-level dialogue evaluation. In Smaranda Muresan, Vivian Chen, Kennington Casey, Vandyke David, Dethlefs Nina, Inoue Koji, Ekstedt Erik, and Ultes Stefan, editors, **Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering**, pp. 80–85, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. Dialogue response ranking training with large-scale human feedback data. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 386–395, Online, November 2020. Association for Computational Linguistics.

[4] Yann Hicke, Abhishek Masand, Wentao Guo, and Tushaar Gangavarapu. Assessing the efficacy of large language models in generating accurate teacher responses. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**, pp. 745–755, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. The BEA 2023 shared task on generating AI teacher responses in educational dialogues. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**, pp. 785–795, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. Designing precise and robust dialogue response evaluators. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 26–33, Online, July 2020. Association for Computational Linguistics.

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

[8] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2157–2169, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[9] Ananya Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. Re-evaluating ADEM: A deeper look at scoring dialogue responses. **CoRR**, Vol. abs/1902.08832, , 2019.

[10] Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 681–707, Online, July 2020. Association for Computational Linguistics.

[11] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[12] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. **arXiv preprint arXiv:2302.04166**, 2023.

[13] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? a preliminary study. In Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini, editors, **Proceedings of the 4th New Frontiers in Summarization Workshop**, pp. 1–11, Hybrid, December 2023. Association for Computational Linguistics.

A UniEval

We modified UniEval, originally designed for the TopicalChat benchmark, to better suit the DailyDialog subset. Given that the **groundedness** submetric only makes sense in the grounded dialouge setting, we removed it. Additionally, **understandability** was excluded for simplicity. The results of ablation experiments are listed in Table 3.

	Content		Grammar		Relevance		Overall	
Metrics	ρ	τ	ρ	τ	ρ	τ	ρ	τ
UniEval	0.387	0.273	0.170	0.117	0.535	0.381	0.285	0.198
- Groundedness	0.387	0.273	0.170	0.117	0.535	0.381	0.315	0.219
- Understandability	0.387	0.273	0.170	0.117	0.535	0.381	0.340	0.236
+ new content prompt	0.448	0.322	0.170	0.117	0.535	0.381	0.477	0.337
+ weighted average	0.448	0.322	0.170	0.117	0.535	0.381	0.491	0.347

Table 3 Turn level Spearman's ρ, and Kendall's τ correlations of different metrics on the Daily Dialog Subset for different versions of the UniEval metric.

B GPT Prompt

The full prompt for evaluation was as follows:

You will be given a conversation between two individuals. You will then be given one potential response for the next turn in the conversation.
Your task is to rate the response on a series of metrics: content quality, grammaticality, and relevance. Finally, you will assign an overall score (not an average).
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Content Quality(1-5) - How compelling is the content of the response, and to what extent does it actively contribute to the ongoing conversation?
- A score of 1 (generic or boring) suggests that the response lacks interesting content and fails to contribute meaningfully to the conversation, potentially coming across as generic or dull.
- A score of 3 (moderately engaging) indicates that the response contains some interesting elements, contributing somewhat to the conversation, but there is room for improvement.
- A score of 5 (interesting and engaging) signifies that the content is exceptionally interesting, capturing attention and actively enhancing the overall conversation, demonstrating a high level of originality and contribution.

Grammaticality(1-5) - How grammatical is the response? Consider only the response itself, not the conversation history.
- A score of 1 (ungrammatical) indicates that the response is confusing, lacks coherence, and is difficult to comprehend.
- A score of 3 (somewhat grammatical) suggests that the response is moderately clear but may contain some ambiguous or convoluted elements.
- A score of 5 (grammatical) indicates that the response is exceptionally clear, logically organized, and easy to understand.

Relevance (1-5) - How well does the response align with the current conversational context and contribute meaningfully to the ongoing discourse? Pay close attention to the speaker.
- A score of 1 (irrelevant) suggests that the response is not related to the current conversation or significantly deviates from the established context.
- A score of 3 (somewhat relevant) indicates a partial alignment with the conversation but may contain elements that are not entirely pertinent to the ongoing discourse.
- A score of 5 (highly relevant) signifies that the response is directly related to the current conversation, seamlessly fitting into the established context without introducing unnecessary tangents or needless repetition.

Overall Score (1-5) - How would you rate the response overall?.
- A score of 1 (poor) indicates that the response is of unrelated, boring, generic, or nonsensical.
- A score of 3 (average) suggests that the response is reasonably appropriate for the conversation, passably understandable, and somewhat interesting.
- A score of 5 (excellent) signifies that the response is exceptionally interesting and engaging, relevant to the conversation, and easy to understand.

Conversation History:
(The following is a conversation between Alice and Bob.)
Alice: Well, how does it look?
Bob: It's a perfect fit.
Alice: Let me pay for it now.

Response:
Bob: Cash, credit card, or debit card?

Evaluation Form (scores ONLY):