RoboRater: Automating Ratings of Task-Oriented Conversations using LLMs

Sally Goldman, Yuri Vasilevski,

Google, Inc.
sgoldman@google.com, yvasilev@google.com

Abstract

We describe RoboRater, an LLM-based approach to evaluate task-oriented conversational data at scale, both to complement human raters and to accurately identify which conversations to use as training for fine-tuning or RLAIF. We combine an ensemble of LLM predictions filtered by diversity entropy to create high-quality training data, and also identify conversations where human raters disagree.

Introduction

Typically the evaluation of conversational data is performed by human raters, which is time-intensive limiting the amount of annotated data. We present RoboRater, an LLM-based approach to evaluate goal-oriented, conversational data at scale, both to complement human raters, and to accurately identify high quality conversations for fine-tuning or RLAIF. We combine an ensemble of LLM predictions, filtered by diversity entropy (DE) to create high-quality training data. Interestingly, the subset of conversations filtered out by DE, are ones in which the agreement rate of humans is also low, and thus inherently more difficult to evaluate.

Most ML-based evaluation techniques focus on text summarization and dialogue generation tasks. In contrast, we focus on goal-oriented tasks such as interacting with a virtual assistant, information seeking, and recommendation tasks. Evaluation questions are dependent on the task but generally relate to ease and satisfaction in completing the task, thus making the evaluation more of a reasoning task than a fluency task.

Related Work

There is much work applying LLMs for evaluation, especially natural language generation since metrics such as BLEU have low correlation with human evals (Deriu et al. 2021; Liu et al. 2016). Some work focuses on the use of a unified LLM (versus specialized LLMs for each dimension) (Lin and Chen 2023) and a majority vote with different LLMs (via debate, repetition or prompt differences) (Wang et al. 2022; Chan et al. 2023). Also Chain-of-Thought (CoT) reasoning (Wei et al. 2022) has been shown to improve the quality of LLM-based evaluation. (Liu et al. 2023) have the LLMs generate a CoT of detailed evaluation steps by feeding the task introduction and evaluation criteria as a prompt.

A direction of research most similar to our work presents techniques to determine which data would most benefit from human involvement. (Cai, Chang, and Han 2023) uses CoT and diversity entropy (Agarwal et al. 2020; Brinker 2003; Yang et al. 2015) to identify data for humans to annotate. (Lee et al. 2023) experimented with eliciting CoT reasoning by adding a sentence asking for thoughts and explanation prior to inferring the label. (Wang et al. 2023a) propose the Multiple Evidence Calibration framework where they address the fact that LLMs give different, inconsistent, candidates by using diversity entropy to measure the difficulty of each example and then use a human-in-the-loop as needed when the entropy is above a threshold.

There is also research that focuses on dialogue data. (Mehri and Eskenazi 2020) introduced the FED data set for evaluating automatic metrics relative to human judgment. (Wang et al. 2023b) focused on evaluation of personalized text generation. Neither of these focus on the analysis of task-oriented conversations with respect to answering evaluation questions and filtering out conversations with low confidence as done in our work.

Methodology

In our work, each data point is a task-oriented conversation between a conversational agent and a user, where immediately after the conversation, the user is given a set of evaluation questions, such as "Were you able to complete your task?", "Did the agent take your feedback into account?", "What was the quality of the recommended item?" For ease of exposition, we focus on a single evaluation question Q. We use a state-of-the-art instruction-tuned LLM with a default temperature value. In each application of the LLM, which we refer to as a *run*, it is provided with a prompt, instructions specific to the task, the full conversation, and the evaluation question. We run the LLM n times to obtain an ensemble A of answers to the evaluation question.

Our preliminary experiments are on conversations for a the task of revising user lists from PRESTO (Goel et al. 2023) along with a rating from the user as to whether they were able to reach the desired end state. We introduced two *explicit CoT* eval questions: the first asking the LLM to describe the expected end state, and a second asking the LLM

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

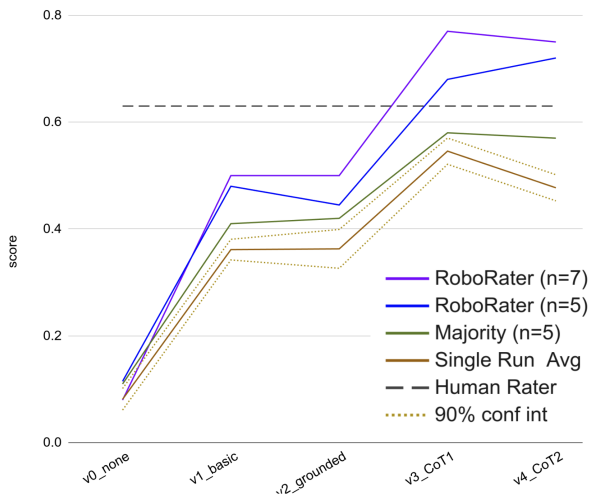

Figure 1: RoboRater performance with prompting and combining variations. A score of 1 is the highest possible.

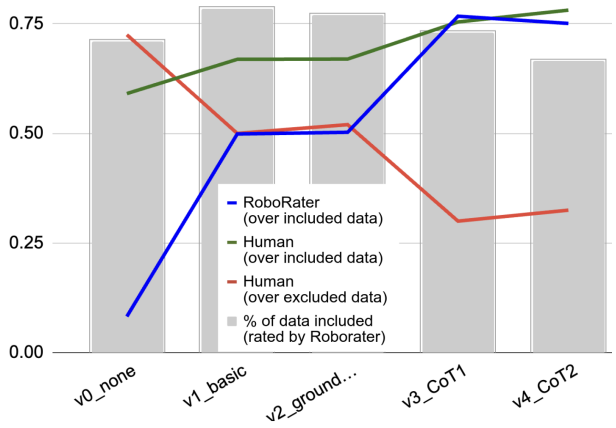

Figure 2: The performance of RoboRater and humans on portion of data rated by RoboRater ($DE(A) \leq \tau$), and also the human performance when $DE(A) > \tau$.

to describe the actual end state. Unlike a human, the LLM can answer these explicit CoT questions independently. The answer to the evaluation question depends upon on the answer to both of the CoT question answers.

Let A be the set of n answers to the evaluation question. We apply diversity entropy defined as

$$DE(A) = \sum_{a \in A} -p_a \log(p_a)$$

where p_a is the fraction of LLM runs where the ensemble predicted a. The lower the diversity entropy, the more appropriate it is to be used for fine tuning or RLAIF data. For threshold τ and $mode(A)$ defined as the most common rating of A, the predicted rating

$$r(A) = \begin{cases} mode(A) & \text{if } DE(A) \leq \tau \\ \text{NONE} & \text{if } DE(A) > \tau \end{cases}$$

We exclude conversations with high diversity entropy, since research shows that a smaller high quality data set is preferable (Team et al. 2023). In our preliminary work that had a yes/no rating question, we defined τ to corresponded to only using data with agreement over at least $n - 1$ of the n runs.

Experiments

As described above, we performed preliminary experiments with a simple digital assistant task for maintaining a set of lists (e.g. shopping list, to-do list,....). For example:

User: Add eggs and butter to the list.

System: Alright. Which list do you want to add to?

User: Shopping.

System: Ok, I added those two items.

User: Add them to the Walmart list instead.

System: Ok, I moved the two items to Walmart list.

Our data set consisted of 200 such conversations gathered via crowd-sourcing in which the worker having the conversation also answered a yes/no question as to whether they believe their desired end state was reached. We also generated a ground truth for each conversation which was yes, no, or vague where it was not clear from the conversation if the end state was met. We define a *score* for the performance of the human rater and RoboRater by taking an average over conversations where the ground truth is not vague with a score of +1 when correct and -1 when incorrect.

We used an instruction-tuned LLM with prompting:

- **v0_none** - no guidance,
- **v1_basic** - same instructions as human rater,
- **v2_grounded** - above plus asked to only rate based on conversation,
- **v3_CoT1** - above plus CoT question to describe expected end state (before rating question),
- **v4_CoT2** - above plus a 2nd CoT question about actual end state.

Figure 1 shows how RoboRater performance varies based on the method of prompting and combining. Figure 2 compares the performance of RoboRater and humans on portion of data rated by RoboRater ($DE(A) \leq \tau$), and also the human performance when $DE(A) > \tau$. Observe that with CoT2 we are able to obtain performance comparable to humans. Using an ensemble of $n = 7$ runs combined via diversity entropy gives the best performance. Furthermore, we find that the conversations in which there is not a strong consensus are inherently harder to rate (even by humans). Hence, RoboRater can identify conversations in which a highly trained rater would be valuable. Comparing the green and red lines in Figure 2 one can see with the addition of CoT, that there is a significant difference in the human performance on the subset of data in which RoboRater had high versus low diversity entropy.

References

Agarwal, S.; Arora, H.; Anand, S.; and Arora, C. 2020. Contextual diversity for active learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 137–153. Springer.

Brinker, K. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, 59–66.

Cai, Z.; Chang, B.; and Han, W. 2023. Human-in-the-Loop through Chain-of-Thought. *arXiv preprint arXiv:2306.07932*.

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; and Cieliebak, M. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54: 755–810.

Goel, R.; Ammar, W.; Gupta, A.; Vashishtha, S.; Sano, M.; Surani, F.; Chang, M.; Choe, H.; Greene, D.; He, K.; et al. 2023. PRESTO: A Multilingual Dataset for Parsing Realistic Task-Oriented Dialogs. *arXiv preprint arXiv:2303.08954*.

Lee, H.; Phatale, S.; Mansoor, H.; Lu, K.; Mesnard, T.; Bishop, C.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Lin, Y.-T.; and Chen, Y.-N. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. *arXiv preprint arXiv:2305.13711*.

Liu, C.-W.; Lowe, R.; Serban, I. V.; 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. *arXiv preprint arXiv:1603.08023*.

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, may 2023. *arXiv preprint arXiv:2303.16634*, 6.

Mehri, S.; and Eskenazi, M. 2020. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, P.; Li, L.; Chen, L.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wang, Y.; Jiang, J.; Zhang, M.; Li, C.; Liang, Y.; Mei, Q.; and Bendersky, M. 2023b. Automated Evaluation of Personalized Text Generation using Large Language Models. *arXiv preprint arXiv:2310.11593*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Yang, Y.; Ma, Z.; Nie, F.; Chang, X.; and Hauptmann, A. G. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113: 113–127.