Clarifying the Dialogue-Level Performance of GPT-3.5 and GPT-4 in Task-Oriented and Non-Task-Oriented Dialogue Systems

Shinya Iizuka*, Shota Mochizuki*, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, Ryuichiro Higashinaka

Graduate School of Informatics, Nagoya University, Japan
{iizuka.shinya.a8, mochizuki.shota.k8, ohashi.atsumoto.c0, yamashita.sanae.w7}@s.mail.nagoya-u.ac.jp,
guo.ao.i6@f.mail.nagoya-u.ac.jp, higashinaka@i.nagoya-u.ac.jp

Abstract

Although large language models such as ChatGPT and GPT-4 have achieved superb performances in various natural language processing tasks, their dialogue performance is sometimes not very clear because the evaluation is often done on the utterance level, where the quality of an utterance given context is the evaluation target. Our objective in this work is to conduct human evaluations of GPT-3.5 and GPT-4 using MultiWOZ and persona-based chat tasks in order to verify their dialogue-level performance in task-oriented and non-task-oriented dialogue systems. Our findings show that GPT-4 performs comparably with a carefully created rule-based system and has a significantly superior performance to other systems, including those based on GPT-3.5, in persona-based chat.

Introduction

Human-robot interaction has a lot to benefit from recent advancements in large language models (LLMs) such as ChatGPT and GPT-4 (OpenAI 2023; Bubeck et al. 2023) for processing dialogue, such as following instructions (Matamoros, Seib, and Paulus 2019), grounding with physical environments (Ahn et al. 2022), and conversations in general. However, one of the current challenges with dialogue processing utilizing LLMs is that, due to the interactive nature of dialogue and the cost involved in such interactive evaluation, their dialogue-level performance is sometimes not very clear with evaluations done only on the utterance level. This lack of clarity makes it difficult for researchers and developers to adopt LLMs for their dialogue applications. To the best of our knowledge, while there have been some utterance-level evaluations, there are few reports on the dialogue-level performance of GPT-3.5 and GPT-4, the most advanced models to date (Bang et al. 2023; Hudeček and Dušek 2023).

Our aim in this study is thus to clarify the dialogue-level performance of LLMs, especially GPT-3.5 and GPT-4, in human evaluation experiments using task-oriented and non-task-oriented dialogue systems. We adopted the popular dialogue task MultiWOZ (Budzianowski et al. 2018) for task-oriented dialogue and the commonly used persona-based chat (Zhang et al. 2018) for non-task-oriented dialogue and then conducted dialogue-level evaluations. Our findings showed that GPT-4 performs the best in both task-oriented and non-task-oriented dialogues: specifically, it performs on par with a carefully created rule-based system in task-oriented dialogue, and it also shows a high performance in persona-based chat, achieving better satisfaction compared to GPT-3.5.

In this paper, we open with our experiment in task-oriented dialogue and show how the systems are implemented using GPT-3.5 and GPT-4 and how the evaluation was performed. Then, we present a similar evaluation for non-task-oriented dialogue. We close with a brief summary and mention of future work. Note that the human evaluation experiment in this paper was approved by the ethical review committee of our institution.

Evaluation in Task-Oriented Dialogue

Among the many tasks typically used for task-oriented dialogue, we chose to utilize MultiWOZ (Budzianowski et al. 2018), which is the most widely studied dialogue task. MultiWOZ covers dialogues between a clerk bot and a customer in tourist information-related domains including attraction, hotel, hospital, restaurant, taxi, train, and police.

System Implementations

We implemented four systems for evaluation: a rule-based system, an end-to-end system fine-tuned with the MultiWOZ data, a system built with GPT-3.5, and a system built with GPT-4, described as follows. All systems performed text-based dialogue with users in English.

Rule-based We used the rule-based system implemented by ConvLab-2 (Zhu et al. 2020), a toolkit to build task-oriented dialogue systems with various pre-trained models. The rule-based system consists of four modules in a pipeline structure: a BERT Natural Language Understanding (NLU) (Chen, Zhuo, and Wang 2019), a rule-based Dialogue State Tracking (DST), a rule policy, and a template Natural Language Generation (NLG). All modules within the system except for NLU are crafted by experts. Studies with a user simulator[1] have shown that the

*These authors contributed equally.

[1]https://github.com/thu-coai/ConvLab-2

System	N	Success	Turn	Understanding	Appropriateness	Satisfaction
Rule-based	39	74.36	11.79	4.08	4.15	4.00
TOATOD	36	66.67	12.56	3.83	3.81	3.72
GPT-3.5	42	57.14	**11.55**	3.79	3.98	**4.05**
GPT-4	42	**76.19**	11.88	**4.26**	**4.36**	4.00

Table 1: Human evaluation in task-oriented dialogue. N is the number of workers, and Turn is the average number of turns until the end of the dialogue. The best score is shown in **bold**, and the second-best score is underlined.

rule-based system has a high task performance.

TOATOD We used the Task-Optimized Adapter for an end-to-end Task-Oriented Dialogue (TOATOD) system (Bang, Lee, and Koo 2023), which is an end-to-end dialogue system built with a T5 (Raffel et al. 2020) model fine-tuned on the MultiWOZ corpus. The TOATOD model sequentially understands the user utterance, outputs a belief state based on the dialogue history, searches the database using the content of the belief state, and finally generates a response using the search results. TOATOD is one of the latest models to achieve top scores on the leaderboard for MultiWOZ response generation[2]. As the source code and trained parameters for TOATOD are available, we utilized the distributed model for implementing the system.

GPT-3.5 Following Hudeček and Dušek (2023), we constructed a task-oriented dialogue system based on GPT-3.5 (`gpt-3.5-turbo` model via OpenAI API). In this system, a context encoder first encodes its input user utterance into a vector representation and then uses it to retrieve dialogue examples with high similarity from the MultiWOZ corpus. Next, the retrieved dialogue examples are utilized as shots by an LLM-based state tracker as a prompt to generate the belief state. After that, the state tracker retrieves the necessary information from the database using the belief state as a query, and finally, using the context and the retrieved information in the prompt, an LLM-based utterance generator generates the system response.

GPT-4 This system is the same as that for GPT-3.5 except that GPT-4 (`gpt-4-0613`, via OpenAI API) is used.

Since TOATOD, GPT-3.5 and GPT-4 only work on the utterance-level as they are, we wrote a wrapper for them to be able to converse interactively. This process includes writing lexicalization rules. Note that the original systems generate delexicalized utterances that include placeholders (e.g., `[phone]` or `[address]`) instead of actual slot values. Therefore, we had to create rules to fill such placeholders with the slot values retrieved from the MultiWOZ database so that the system can generate meaningful utterances.

Experiment

We conducted a human evaluation experiment to evaluate the above systems on the dialogue level. We recruited approximately 40 workers for each system via Amazon Mechanical Turk (AMT). To ensure the quality of the experiment, we only enrolled workers who met all of the following criteria: (1) have completed tasks on AMT more than ten times, (2) have a task approval rate of 95% or higher, (3) reside in English-speaking regions, and (4) have answered all five common-sense questions correctly.

During the experiment, each worker first read the instructions for their assigned dialogue goals and then tried to achieve these goals by chatting with a system. Each dialogue goal was randomly generated using a function implemented in ConvLab-2, and one to three domains were randomly assigned for each goal. The average number of domains used for dialogue goals in Rule-based, TOATOD, GPT-3.5, and GPT-4 were 1.72, 1.81, 1.81, and 1.81, respectively. Since the difficulty of the dialogue is predominantly determined by the number of domains, it is assumed that there was little difference in the difficulty of dialogue among the systems.

Each worker interacted only once with one of the systems with a given goal, and the worker was required to achieve their dialogue goal within 20 utterances. The task was determined a success by the worker's self-assessment; that is, if the worker considered that the dialogue goal had been achieved during the dialogue, the worker would judge the dialogue as a success and terminate it. Otherwise, the worker would judge the dialogue as a failure. In addition, when the maximum number of turns had elapsed, the dialogue would automatically be judged as a failure.

After each dialogue, the workers subjectively evaluated the system's ability in terms of language understanding, the appropriateness of the system response, and their satisfaction with the dialogue on a five-point Likert scale.

Results

Table 1 lists the dialogue performances for task success (task success rate) and subjective evaluation for the four systems. As we can see, GPT-4 reached the highest performance in task success, system language understanding, and appropriateness of system response. Its task success is on par with that of the Rule-based system, which is surprising because the Rule-based system is known for its high performance compared to neural-based models (Zhu et al. 2020). Although it is necessary to retrieve dialogue examples from the MultiWOZ dataset as shots to be included in the prompt, the fact that GPT-4 was able to achieve this task success is considered significant progress. On the other hand, GPT-3.5 had difficulty achieving the task, similarly to its reported performance on the utterance level (Bang et al. 2023). TOATOD was better than GPT-3.5, but it did not reach the level of GPT-4.

[2]https://github.com/budzianowski/multiwoz

Worker	What are the IDs of the 19 trains along with their departure times?
GPT-4	The IDs and departure times of the 19 trains are as follows: [train1_id] at [time1], [train2_id] at [time2], ..., [train19_id] at [time19]. Please let me know if you need more information.

Table 2: Typical failure case of GPT-4 with undefined placeholders.

Analysis

An example of a typical failure case of GPT-4 is when it generated placeholders such as `train1_id` and `time1` that were not defined in the MultiWOZ task by hallucinations (see Table 2 for an example). Another failure case of GPT-4 is when it failed to switch the domain during a dialogue, leading to a misunderstanding of user utterances. Although the task success rate of GPT-4 is reasonably high at 75%, these failure cases demonstrate the necessity for better understanding of the task and domains, which could presumably be done by ensuring more grounding in the dialogue task.

Evaluation in Non-Task-Oriented Dialogue

One of the main goals of non-task-oriented dialogue systems is to have a human-like conversation. To this end, the system needs to exhibit a certain personality. There have been a number of studies on persona-based chat (Li et al. 2016; Zhang et al. 2018) in which interlocutors have conversations with consistent personalities. Following this vein, we focus on persona-based chat as a type of non-task-oriented dialogue for evaluation.

Note that there are other important elements in non-task-oriented dialogue, such as social (Yu et al. 2019), emotional (Rashkin et al. 2019), and entertaining (García-Méndez et al. 2021) aspects. One previous study examined the dialogue performance of ChatGPT in empathetic response generation and emotional support conversation tasks (Zhao et al. 2023), with the evaluation done on the dialogue level. However, the performance of GPT-4 was not evaluated and the evaluation was performed in a pair-wise fashion, making it difficult to clarify the differences among systems.

System Implementations

We implemented four LLM-based systems for this evaluation: Japanese-dialog-transformers, OpenCALM-3B, GPT-3.5, and GPT-4, described as follows. Note that, since the language of the data we utilized for persona and personality traits was Japanese, the systems performed dialogue in Japanese. The dialogue was done via text chat.

Japanese-dialog-transformers We built this system based on Japanese-dialog-transformers (Sugiyama et al. 2023), the 1.6B-parameter Transformer-based encoder-decoder model for Japanese chit-chat. This is one of the most popular models for persona-based chat in Japanese. For implementing the system, we utilized the publicly available model[3] pretrained with Twitter data and fine-tuned on JPersonaChat (Sugiyama et al. 2023), the Japanese version of PersonaChat (Zhang et al. 2018). The model does not use persona as input. We utilized the default setting for generation, with five previous utterances provided as context.

OpenCALM-3B We built this system based on OpenCALM, the Transformer-based decoder-only large Japanese language model, which was recently released by CyberAgent, Inc. It has been pre-trained on the Wikipedia and Common Crawl datasets. In this experiment, we utilized a model with 3B parameters[4], which is more than the Japanese-dialog-transformers model. We built the system by LoRA-tuning (Hu et al. 2021) the model with JPersonaChat. We prepared two variants for this system: (i) persona is used as input and (ii) persona is not used as input (dialogue history only). For generation, on the basis of our preliminary experiment, the system uses the previous 12 utterances as context.

GPT-3.5 We built this system based on GPT-3.5 utilizing the `gpt-3.5-turbo` model. To assess the influence of not only persona but also personality on utterances, we built a system using the personas and personalities in RealPersonaChat (Yamashita et al. 2023). Specifically, we prompted the system to output the next utterance using the persona, personality traits, and dialogue history (all previous user and system utterances) as input. Persona is represented by ten sentences describing the profile of the interlocutor. Personality traits are a set of high/low values regarding certain personality traits, including those in the Big Five (Goldberg 1990; McCrae and John 1992; Fossati et al. 2011) and the Adult Temperament Questionnaire (ATQ) (Evans and Rothbart 2007). To ensure the consistency of persona and personality, we utilized the set of persona and personality from the same individual.

GPT-4 This model is the same as GPT-3.5 except that GPT-4 (the `gpt-4-0613` model) was used; namely, the prompt includes persona, personality traits, and dialogue history. In addition, we created three variants in which the system utilizes a prompt containing (i) persona and dialogue history, (ii) personality traits and dialogue history, and (iii) dialogue history only. This was to examine the effect of persona and personality traits on dialogue performance.

Experiment

We recruited 30 workers for each system through the CrowdWorks crowdsourcing service[5]. Each worker chatted with one of the systems only once and engaged in dialogue with the systems on a web-based interface we developed. The number of utterances per dialogue was limited to 20: ten by the worker and ten by the system, in an alternating fashion. We instructed the workers to freely chit-chat with the

[3] https://github.com/nttcslab/japanese-dialog-transformers

[4] https://huggingface.co/cyberagent/open-calm-3b

[5] https://crowdworks.jp

System	N	Coherence	Informativeness	Satisfaction
GPT-4 + persona & personality	30	**4.60**	4.50	**4.60**
GPT-4 + persona	30	4.57	**4.53**	4.53
GPT-4 + personality	30	**4.60**	4.00	4.07
GPT-4 (dialogue history only)	30	4.57	4.47	4.23
GPT-3.5 + persona & personality	30	4.37	4.47	4.23
Japanese-dialog-transformers	30	3.67	4.03	3.43
OpenCALM-3B + persona	30	3.20	3.10	2.97
OpenCALM-3B (dialogue history only)	30	3.53	3.53	3.33

Table 3: Human evaluation in non-task oriented dialogue. The best score is in **bold**, and the second-best score is underlined. The dialogue history is provided uniformly to all systems.

system. As personas for OpenCALM-3B, we prepared ten sets of personas randomly selected from JPersonaChat; note that personality traits are not included in JPersonaChat. Each persona consists of five sentences that represent the profile of an individual.

For GPT-3.5 and GPT-4, we prepared ten sets of personas and personality traits randomly selected from RealPersonaChat. The characteristics of this persona include 1) consisting of ten sentences, 2) being approximately three times longer than that in JPersonaChat, and 3) being realistic because it is collected from real individuals. For more details, refer to (Yamashita et al. 2023).

After the dialogue, the workers subjectively evaluated the system in terms of its coherence, informativeness, and satisfaction selected from the dialogue-level evaluation items described in (Mehri and Eskenazi 2020). The evaluation was done on a 5-point Likert scale.

Results

Table 3 lists the evaluation results for each system. As we can see, GPT-3.5 and GPT-4 achieved higher scores than Japanese-dialog-transformers and OpenCALM-3B for all evaluation items. This demonstrates the positive effect of the large parameter size of GPT-3.5 and GPT-4. GPT-4 was superior to GPT-3.5 for all evaluation items, and the difference was especially large for satisfaction.

Regarding Japanese-dialogue-transformer, it was better than the OpenCALM-3B + persona model even though it has fewer parameters and does not utilize persona information as input. We looked into the generated utterances of OpenCALM-3B + persona and found that persona-related information appeared frequently in the utterances, possibly leading to unnaturalness. The fact that OpenCALM-3B without the persona information performs better suggests the difficulty of naturally incorporating persona information into utterances, or possibly the unnatural nature of the JPersonaChat corpus in which the interlocutors had conversations with given personas and needed to talk about them (Yamashita et al. 2023).

Regarding GPT-4, when comparing scores depending on the information included in the prompt, the system that included both the persona and personality traits showed the highest score in coherence and satisfaction. The contribution of personality alone was not strong, leading to decreased performance for informativeness and satisfaction, which we investigate in the following subsection.

Analysis

GPT-4 showed a significant decline in satisfaction when the persona was removed from the prompt ($p < 0.05$, Mann-Whitney U test). We looked into the dialogues and found that there were cases where the personality was too strongly reflected in the utterances. For example, in some cases with introverted personalities (such as "Extroversion is low" or "Sociability is low"), we observed dialogues where the system responded to the user utterance with only terse replies, exhibiting low engagement by the system.

Seven of the dialogues by GPT-4 without personas had a satisfaction rating of three or lower. The commonality of these seven dialogues was that certain personality traits (such as extraversion in Big Five and sociability in ATQ) were low. We believe that the differences in the quality of dialogue are not due to the system's dialogue ability but rather to the non-talkative personality given to the system.

Summary and Future Work

In this paper, we investigated the dialogue-level performance of GPT-3.5 and GPT-4 in human evaluation experiments using task-oriented and non-task-oriented dialogue systems. Our findings demonstrate that, in task-oriented dialogue, GPT-4 has reached the level of dialogue that was previously only achievable by hand-crafted rules. In addition, in persona-based chat, GPT-4 greatly surpassed GPT-3.5, and it attained the highest score when using persona and personality traits. We also found that the use of personality traits may have an adverse effect depending on the traits.

In future work, we plan to further investigate the dialogue-level performance of GPT-4 in more complex dialogue tasks (such as information access (Dinan et al. 2018), negotiation (Wang et al. 2019), and sales (Smith 2020)) and in non-task-oriented dialogue in which emotional and social aspects are of greater importance. In addition, we also want to investigate how high-performance LLMs can be utilized in human-robot interaction.

Acknowledgments

This work was supported by JST Moonshot R&D Grant number JPMJMS2011.

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Ho, D.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jang, E.; Ruano, R. J.; Jeffrey, K.; Jesmonth, S.; Joshi, N.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Lee, K.-H.; Levine, S.; Lu, Y.; Luu, L.; Parada, C.; Pastor, P.; Quiambao, J.; Rao, K.; Rettinghouse, J.; Reyes, D.; Sermanet, P.; Sievers, N.; Tan, C.; Toshev, A.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yan, M.; and Zeng, A. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*.

Bang, N.; Lee, J.; and Koo, M.-W. 2023. Task-Optimized Adapters for an End-to-End Task-Oriented Dialogue System. In *Find. ACL*, 7355–7369.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv preprint arXiv:2302.04023*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proc. EMNLP*, 5016–5026.

Chen, Q.; Zhuo, Z.; and Wang, W. 2019. BERT for Joint Intent Classification and Slot Filling. *arXiv preprint arXiv:1902.10909*.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Evans, D. E.; and Rothbart, M. K. 2007. Developing a model for adult temperament. *Journal of research in personality*, 41(4): 868–888.

Fossati, A.; Borroni, S.; Marchione, D.; and Maffei, C. 2011. The big five inventory (BFI). *European Journal of Psychological Assessment*, 27: 50–58.

García-Méndez, S.; De Arriba-Pérez, F.; González-Castaño, F. J.; Regueiro-Janeiro, J. A.; and Gil-Castiñeira, F. 2021. Entertainment Chatbot for the Digital Inclusion of Elderly People Without Abstraction Capabilities. *IEEE Access*, 9: 75878–75891.

Goldberg, L. R. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6): 1216–1229.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Hudeček, V.; and Dušek, O. 2023. Are Large Language Models All You Need for Task-Oriented Dialogue? In *Proc. SIGDIAL*.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A Persona-Based Neural Conversation Model. In *Proc. ACL*, 994–1003.

Matamoros, M.; Seib, V.; and Paulus, D. 2019. Trends, Challenges and Adopted Strategies in RoboCup@Home. In *Proc. 2019 IEEE International Conference on Autonomous Robot Systems and Competitions*, 1–6.

McCrae, R. R.; and John, O. P. 1992. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2): 175–215.

Mehri, S.; and Eskenazi, M. 2020. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proc. SIGDIAL*, 225–235.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(140): 1–67.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proc. ACL*, 5370–5381.

Smith, K. T. 2020. Marketing via smart speakers: what should Alexa say? *Journal of Strategic Marketing*, 28(4): 350–365.

Sugiyama, H.; Mizukami, M.; Arimoto, T.; Narimatsu, H.; Chiba, Y.; Nakajima, H.; and Meguro, T. 2023. Empirical analysis of training strategies of transformer-based Japanese chit-chat systems. In *Proc. SLT*, 685–691.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Yamashita, S.; Inoue, K.; Guo, A.; Mochizuki, S.; Kawahara, T.; and Higashinaka, R. 2023. RealPersonaChat: A Realistic Persona Chat Corpus with Interlocutors' Own Personalities. In *Proc. PACLIC*, (to appear).

Yu, D.; Cohn, M.; Yang, Y. M.; Chen, C. Y.; Wen, W.; Zhang, J.; Zhou, M.; Jesse, K.; Chau, A.; Bhowmick, A.; Iyer, S.; Sreenivasulu, G.; Davidson, S.; Bhandare, A.; and Yu, Z. 2019. Gunrock: A Social Bot for Complex and Engaging Long Conversations. In *Proc. EMNLP-IJCNLP (System Demonstrations)*, 79–84.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zhao, W.; Zhao, Y.; Lu, X.; Wang, S.; Tong, Y.; and Qin, B. 2023. Is ChatGPT Equipped with Emotional Dialogue Capabilities? *arXiv preprint arXiv:2304.09582*.

Zhu, Q.; Zhang, Z.; Fang, Y.; Li, X.; Takanobu, R.; Li, J.; Peng, B.; Gao, J.; Zhu, X.; and Huang, M. 2020. ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. In *Proc. ACL*, 142–149.