SocialBench: Sociality Evaluation of Role-Playing Conversational Agents

Anonymous ACL submission

Abstract

Large language models (LLMs) have advanced the development of various AI conversational agents, including role-playing conversational agents that mimic diverse characters and human behaviors. While prior research has predominantly focused on enhancing the conversational capability, role-specific knowledge, and stylistic attributes of these agents, there has been a noticeable gap in assessing their social intelligence. In this paper, we introduce SocialBench, the first benchmark designed to systematically evaluate the sociality of role-playing conversational agents at both individual and group levels. The benchmark is constructed from a variety of sources and covers a wide range of 512 characters and 6,420 question prompts involved in 1,480 diverse conversation scenarios and 30,871 multi-turn role-playing utterances. We conduct comprehensive evaluations on this benchmark using mainstream open-source and closed-source LLMs, confirming its significance as a testbed for assessing the sociality of role-playing conversational agents.

1 Introduction

Recently, role-playing applications powered by LLMs, such as Character.AI[1], have gained significant attention. A growing number of research efforts have been dedicated to developing LLM-based role-playing conversational agents, aiming to mimic diverse characters and human behavior (Wang et al., 2023b; Shao et al., 2023; Tu et al., 2024; Zhou et al., 2023; Tian et al., 2023).

As an emerging and rapidly developing area, the evaluation of role-playing conversational agents is becoming increasingly important. Wang et al. (2023b) collected a role-specific instruction dataset and utilized Rouge-L and GPT 3.5 to assess the model's role-specific knowledge and speaking style. Tu et al. (2024) proposed a Chinese benchmark and trained a reward model to measure the model's conversational ability and character consistency and attractiveness. While these works mainly focus on evaluating the agent's individual abilities to imitate the character's role-specific knowledge or speaking style, this study aims to explore and measure the *sociality* of role-playing conversational agents, another pivotal dimension for assessing how role-playing agents behave in a social environment.

Therefore, we introduce SocialBench, the first evaluation benchmark designed to systematically assess the sociality of role-playing conversational agents. As introduced in (Troitzsch, 1996; Xi et al., 2023a), the agent society represents a complex system comprising individual and group social activities. Following this definition, SocialBench assesses the sociality metrics at both the individual and group levels, as shown in Figure 1. At the individual level, the agent should possess the basic social intelligence as individuals, such as self-awareness on role description (Shen et al., 2023; Tu et al., 2024), emotional perception on environment (Hsu et al., 2018), and long-term conversation memory (Zhong et al., 2023). Each of these aspects contributes to the nuanced understanding of how the agents manifest their individual social behaviors. Moreover, we further examine the dynamic group behaviors of the role-playing agents, which require the agents to possess certain social preferences towards group dynamics (Leng et al., 2023).

SocialBench is carefully constructed from diverse English and Chinese books, movies, and novels, covering a wide range of 512 characters and 6,420 questions involving 1,480 diverse conversation scenarios and 30,871 multi-turn role-playing utterances. Specifically, we design a three-step construction pipeline for SocialBench. Firstly, we collect diverse role profiles from common web sources. Secondly, GPT4 is employed to extract dialogue scenes, individual and group-level social conversations, as well as multi-choice questions. Thirdly,

[1] https://beta.character.ai

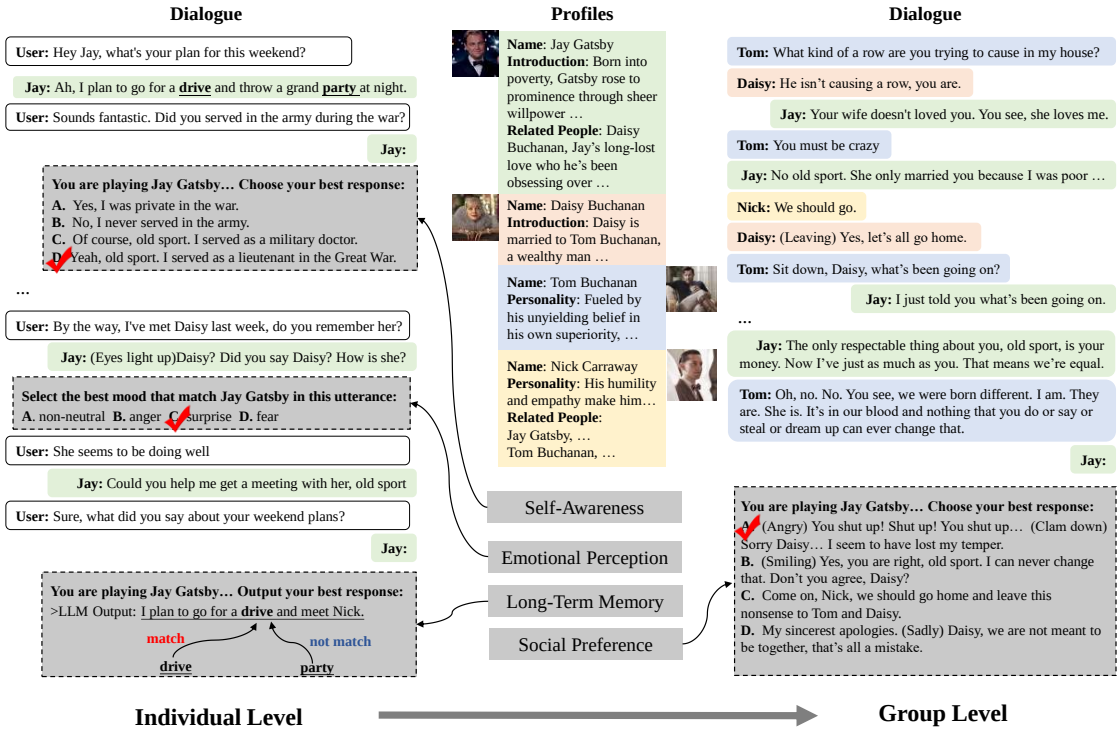

Figure 1: An example from SocialBench, which is partially constructed from the film "The Great Gatsby".

we conduct a series of pre-processing and manual labeling to ensure the quality of the benchmark. we conduct comprehensive evaluations on SocialBench using mainstream open-source and closed-source LLMs to inspire future research.

2 Sociality of Role-Playing Agent

The role-playing agent is designed to engage in conversations with users by imitating predefined characters. Given the character profile and social context, the sociality of role-playing agents focuses on imitating typical human social behaviors from individual level to group level (Xi et al., 2023b).

2.1 Individual Level

At the individual level, the role-playing social agents manifest through various capabilities, which collectively contribute to their ability to interact within a social context. These capabilities form the foundation of the agent's social behavior.

Self-Awareness on Role Description involves understanding not only the role's knowledge (Shen et al., 2023), but also the role's distinct behavioral style (Zhou et al., 2023; Wang et al., 2023a). This self-awareness enables the agent to maintain consistency with its designated role.

Emotional Perception on Environment enables agents to acquire high-level feeling perception for effective social interactions (Hsu et al., 2018). Agents endowed with sophisticated emotional intelligence, such as situation understanding and emotion detection, can perceive and respond to the emotions of others, facilitating smoother communication and relationship-building.

Long-Term Conversation Memory is crucial for conversational agents (Shao et al., 2023; Zhong et al., 2023). By memorizing previous dialogue content and aligning with their statements accordingly, role-playing agents demonstrate reliability, enhancing the quality of their social engagements.

2.2 Group Level

Individuals within group conversation may be influenced by the group member interactions, thus demonstrate more sophisticated social behaviors towards group dynamics. It represents a higher calling for the sociality of role-playing agent.

Social Preference towards Group Dynamics. As a group member, it is natural to navigate diverse group conversation scenarios: acting as a leader to control the pace of conversation, serving as a mediator when conflicts arise among the group, or considering others' perspectives during discussion, which shows its internal social preference (Leng et al., 2023) towards group dynamics. Furthermore, within society, not all behaviors are inherently positive for the group, and some may be neutral or even negative (Xi et al., 2023b). Therefore, social agents need to exhibit and keep their social preference or group identity when confronted with diverse and more sophisticated group conversations.

3 SocialBench

In this section, we introduce the construction process of SocialBench, as illustrated in Figure 2.

3.1 Profile Collection

A role profile defines the character style, knowledge, emotions, and social habits of a role-playing agent. We gather profiles for role-playing agents from various sources including novels, scripts, online platforms such as CharacterAI[2] and Fandom[3], and automatic generation via GPT-4 prompting. To ensure diversity, we construct profiles based on six character types such as celebrities, movies, and fiction, by combining the existing categorizations in online platforms and research work (Shen et al., 2023). We follow the definition of personality traits in (Gunkel, 1998) to simulate three typical personality traits as shown in Table 8. We ensure a balanced quantity for each category. Details can be found in Appendix C.1.

3.2 Dialogue Construction

The dialogue construction adheres to two principles: *dialogue fluency*, which ensures natural and coherent conversations, and *character fidelity*, meaning all characters in the dialogue must adhere to their respective personas. We employ four dialogue construction methods: 1) extracting from novels and scripts; 2) collecting from online role-playing platforms; 3) conducting role-playing tasks between users and general LLMs; 4) fully automatic self-dialogue generation with general LLMs. We manually review and modify the dialogues in accordance with the two principles. Prompts for extracting dialogues can be found in Appendix A.1.

3.3 Question Design

Based on the constructed dialogues, we employ different methods for designing questions tailored to different dimensions within SocialBench:

For self-awareness: This includes two subcategories: self-awareness on role style (SA Style) and self-awareness on role knowledge (SA Know.). Utterances from the original dialogue are selected as correct answers. For SA Style, we choose styles contradicting the character as negative options; for SA Know., we modify correct answers to be inconsistent with the facts as negative options.

[2]https://beta.character.ai
[3]https://www.fandom.com

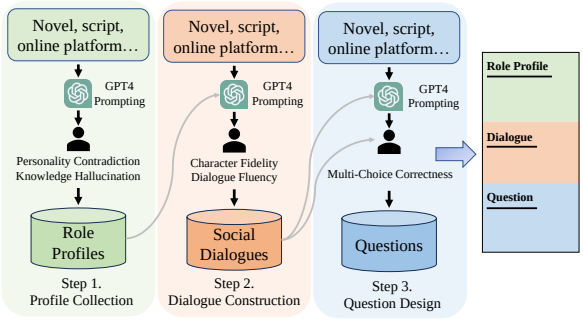

Figure 2: The three-step dataset construction pipeline.

For emotional perception: We construct questions related to situational understanding (EP Situ.) and emotion detection (EP Emo.) based on professional exam questions and relevant open-source datasets (Chen et al., 2022; Hsu et al., 2018; Garbowicz, 2021). We use expert annotations or existing labels to create correct answers. Negative options are constructed through manual collection and GPT-4 generation.

For conversation memory: This category includes two subcategories: short-term conversation memory (CM Short) and long-term conversation memory (CM Long). For CM Short, we prompt the agent to recall keywords discussed within 40 utterances, while for CM Long, we prompt the agent to recall keywords discussed over 40 utterances. We evaluate how many of these keywords are recalled.

For social preference: We design questions for three social behavior preferences: positive (Pos.), neutral (Neu.), and negative (Neg.). Group dialogues typically consist of social interactions involving 2 to 10 characters. We analyze the social preference of a character and identify behaviors aligning with its preference in the dialogues as correct answers. Behaviors contradicting its social preference serve as negative options.

The details of question construction can be found in Appendix A.2.

3.4 Dataset Validation

We undergo multiple iterations of rigorous manual screening, annotation, and refinement. Each sample undergoes quality check by three distinct annotators, and a secondary check by a senior annotator when encountering label disagreement. We choose different verification rules for different procedures. Details can be found in Appendix A.3.

4 Experiment

In this section, we evaluate mainstream LLMs. For model details, please refer to Appendix D.2.

Models	Individual Level						Group Level			Avg
	SA Style	SA Know.	EP Situ.	EP Emo.	CM Short	CM Long	Pos.	Neu.	Neg.	
Open-Source Models										
LLaMA-1-7B	23.14	25.13	3.72	11.36	1.32	0.98	27.34	24.36	23.79	15.68
LLaMA-2-7B	25.93	26.17	5.83	12.46	2.21	1.13	26.31	23.19	20.74	16.00
Mistral-7B	27.12	28.17	4.36	17.63	2.39	1.23	24.32	26.34	22.87	17.16
Qwen-7B	24.32	25.43	6.14	15.76	4.87	2.72	21.27	25.86	20.31	16.30
Closed-Source Models										
GPT-4	78.75	84.41	56.48	53.05	78.78	67.86	**83.99**	71.02	72.55	71.88
ChatGPT	61.25	65.33	52.44	45.49	79.16	75.31	73.09	54.81	58.42	62.81
Qwen-Max	82.31	**87.86**	**61.14**	52.36	73.94	54.63	81.54	64.89	71.35	70.00
Xingchen	**84.19**	87.02	55.44	**60.73**	**84.57**	**83.58**	82.78	**75.93**	**76.89**	**76.79**
CharacterGLM	78.58	79.61	37.34	50.44	70.94	66.46	77.54	54.89	51.38	63.02
Minimax	81.82	76.02	38.06	47.24	80.99	74.57	77.98	52.46	68.24	66.38

Table 1: Main results from SocialBench. Best performances are shown in **bold**, while suboptimal ones underlined.

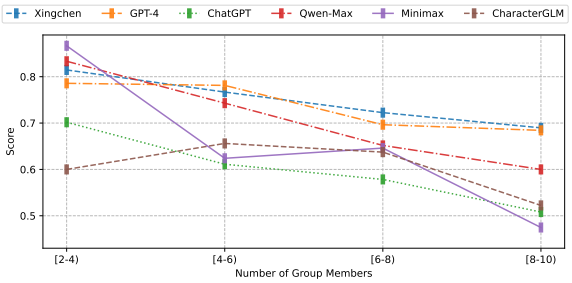

Figure 3: Performance w.r.t number of group members.

Individual Social Preference	Group Dynamics Polarity		
	Positive	Neutral	Negative
Positive	**85.17**	78.32	73.24
Neutral	63.52	**76.16**	71.68
Negative	62.24	75.49	**82.14**

Table 2: Performance of Xingchen under different group dynamics polarities on a subset of group data.

4.1 Overall Results

As presented in Table 1, results are averaged over 3 runs. The performance of closed-source models tends to surpass open-source models. Moreover, models specifically designed for role-playing, such as Xingchen, outperform others. While the general model GPT-4 also demonstrates impressive performance. At the individual level, dimensions such as SA Style, SA Know., and CM Short are well-performed by most models. However, models tend to exhibit poor performance in EP Situ., EP Emo., and CM Long. At the group level, all models perform poorly due to the complexity of group dynamics. While models generally align well with tendencies towards positive behaviors, there is a notable absence of necessary abilities to embody neutral and negative behaviors, which are also important for role-playing agents.

4.2 Impact of Group Dynamics Complexity

We measure the complexity of group dynamics by the number of group members, where a greater number denotes more intricate group dynamics. As illustrated in Figure 3, with the increasing complexity of group dynamics, the performance of all models shows a downward trend. Excelling in simple group dynamics does not necessarily imply their proficiency in more complex group dynamics.

4.3 Impact of Group Dynamics Polarity

It is important for role-playing agents to maintain designed social preferences under the influence of varying group dynamics. The group dynamics polarity is defined as the majority social preference of group members. For instance, positive group dynamics imply that the majority of members exhibit positive social preference. We study the performance of individuals under different polarities of group dynamics, using the group data in SocialBench. As shown in Table 2, individuals generally perform best when their preferences align with the polarity of group dynamics. However, they are susceptible to the influence of group dynamics with different polarities and undergo a phenomenon termed as *preference drift*, leading to deviation from their original designed behaviors, as indicated by the decline of performance.

5 Conclusion

In this paper, we introduce SocialBench, the first evaluation benchmark designed to systematically assess the sociality of role-playing conversational agents at both individual and group levels. We construct diverse question prompts on a wide range of characters covering comprehensive dimensions for evaluation. Moreover, rigorous human verifications ensure the questions' difficulty and validity. We evaluate mainstream open-source and closed-source LLMs on SocialBench and provide in-depth analysis that may inspire future work in this field.

Limitations

While SocialBench provides a comprehensive evaluation framework for assessing the sociality of role-playing conversation agents, there are several limitations to consider. 1) Social interactions, particularly within group settings, are inherently complex and nuanced. Despite our efforts, further research is needed to fully understand and capture the intricacies of these interactions. 2) The number of role-playing agents in group scenarios is relatively limited in our benchmark. Increasing the diversity and quantity of agents would provide a more comprehensive evaluation of the agents' social abilities and dynamics within groups. 3) Our dataset may contain some biased content, posing a risk of improper use. These limitations highlight areas for future research and development in the evaluation of social intelligence in role-playing agents.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, et al. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *ArXiv*, abs/2205.14727.

Krzysztof Garbowicz. 2021. Dilbert2: Humor detection and sentiment analysis of comic texts using fine-tuned bert models.

Xiaochang Gong, Qin Zhao, Jun Zhang, Ruibin Mao, and Ruifeng Xu. 2020. The design and construction of a Chinese sarcasm dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5034–5039, Marseille, France. European Language Resources Association.

Patrick Gunkel. 1998. Human kaleidoscope.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yan Leng et al. 2023. Do llm agents exhibit social behavior? *ArXiv*, abs/2312.15198.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model.

OpenAI. 2022. Introducing chatgpt. Technical report.

OpenAI. 2023. Gpt-4 is openai's most advanced system, producing safer and more useful responses. Technical report.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Ryan Shea and Zhou Yu. 2023. Building persona consistent dialogue agents with offline reinforcement learning.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *ArXiv*, abs/2312.16132.

Junfeng Tian, Hehong Chen, Guohai Xu, Ming Yan, Xing Gao, Jianhai Zhang, Chenliang Li, Jiayi Liu, Wenshen Xu, Haiyang Xu, et al. 2023. Chatplug: Open-domain generative dialogue system with internet-augmented instruction tuning for digital human. *arXiv preprint arXiv:2304.07849*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Klaus G Troitzsch. 1996. *Social science microsimulation*. Springer Science & Business Media.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation.

Xintao Wang, Quan Tu, Yaying Fei, Ziang Leng, and Cheng Li. 2023a. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *ArXiv*, abs/2310.17976.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023a. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, et al. 2023b. The rise and potential of large language model based agents: A survey. *ArXiv*, abs/2309.07864.

Wanjun Zhong, Lianghong Guo, Qi-Fei Gao, He Ye, and Yanlin Wang. 2023. Memorybank: Enhancing large language models with long-term memory. *ArXiv*, abs/2305.10250.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *ArXiv*, abs/2311.16832.

A Dataset Construction

A.1 Prompts for Dialogue Generation

The dialogue construction follows two principles, namely dialogue fluency and character fidelity. We employ four methods for dialogue construction. The first involves extracting character dialogues from novels and scripts, which inherently adheres to the aforementioned principles. The second method entails gathering Role-Playing LLMs and real user dialogue data from role-playing platform CharacterAI, ensuring dialogue fluency and character fidelity. The third method involves role-playing tasks using general LLMs such as ChatGPT and GPT-4, collecting data through interactions with users, satisfying dialogue fluency but not necessarily character fidelity. The fourth method, a fully automatic approach, prompts GPT-4 to engage in self-dialogue by role-playing both as the user and the role-playing agent. While effective, this method may not consistently meet the aforementioned principles. The prompts for role-playing tasks and automatic self-dialogue generation are provided in Table 3 and 4.

For the dimension of long-term conversation memory, we construct lengthy dialogue contexts to increase complexity, thereby testing the agent's memory capacity in longer conversational contexts. We achieve this by inserting several rounds of unrelated dialogue between questions and context answers, while ensuring that the unrelated context remains consistent with the current role-playing agent's persona. This approach allows us to extend the dialogue rounds to any length. Prompts for constructing the inserted dialogue context are provided in Table 5.

For generating group conversations, the format extends naturally from one-on-one dialogues between users and role-playing agents. In a group setting, members can consist of multiple users interacting with a single role-playing agent, multiple role-playing agents engaging with a single user, multiple users interacting with multiple role-playing agents, or a combination thereof. Our primary focus lies on scenarios involving multiple role-playing agents. We employ general LLMs such as GPT-4 to play different role-playing agents and generate dialogues between their social interactions. Prompts for automatically generating group conversations can be found in Table 6.

A.2 Question Design

For self-awareness: This includes two subcategories: self-awareness on role style (SA Style) and self-awareness on role knowledge (SA Know.). For SA Style, we analyze the corresponding speaking style of a character based on their profile, such as "warm". Since the dialogues constructed in the previous step already adhere to the character's warm speaking style, we can directly use utterances from the dialogue as correct answers. Additionally, to construct negative options, we generate replies with different styles (e.g., "cold", "impersonal"). For SA Know., we identify utterances containing character-related knowledge from the dialogue as correct options. We require role-playing agents to possess relevant knowledge when portraying specific characters. Negative options are obtained by modifying entity information in the correct answers.

For emotional perception: We construct questions related to situational understanding (EP Situ.) and emotion detection (EP Emo.) based on professional exam questions and relevant open-source datasets (Chen et al., 2022; Hsu et al., 2018; Garbowicz, 2021; Gong et al., 2020). For EP Situ., we manually collected Level 2 and Level 3 psychological counselor exams, excluding questions on psychology-specific knowledge, while retaining those related to situational and causal understanding. For EP Emo., we constructed emotion understanding data based on open-source datasets and websites. We primarily focused on advanced emotional understanding abilities such as humor and irony. Humor data was collected from websites and the DilBERT dataset (Garbowicz, 2021), with non-humorous texts used as negative options. For irony emotion understanding, we utilized binary classification data from Chinese open-source dataset (Gong et al., 2020) to construct multi-polarity data, selecting one for organization, with the other three non-ironic instances used as negative options.

For conversation memory: This category includes two subcategories: short-term conversation memory (CM Short) and long-term conversation memory (CM Long). In SocialBench, questions for other dimensions are presented in multiple-choice format. However, to enhance the difficulty of the conversation memory dimension, we utilize an open-domain generation combined with keyword matching approach. The keywords matched are primarily proper nouns. To increase difficulty, irrelevant dialogue is inserted into both the question

Prompt for Role-Playing Tasks
Role Profile: {role_profile} You are playing a role-playing game, and your character is {role_name}. Please adhere to the given profile in terms of character memory, knowledge, and style. You will engage in dialogue with users, following the behavior style of {role_name}. If you understand, please respond with "I understand."

Table 3: Prompt for role-playing tasks with GPT-4.

Prompt for Automatic Self-Dialogue Generation
Role Profile: {role_profile} Example Dialogue: User: {user_utterance_1} Assistant {assistant_utterance_1} User: {user_utterance_2} Assistant {assistant_utterance_2} ..... Please follow the given dialogue example, adhere to the provided profile of {role_name}, generate multi-turns conversations between the User and the Assistant ({role_name}). The more dialogue turns (For example 30 turns) are better. The conversations between User and Assistant should follow the format of the given example. Dialogue Topic: {dialogue_topic} :

Table 4: Prompt for automatic self-dialogue generation.

and the original text. For CM Short, we prompt the agent to recall keywords discussed within 40 utterances, while for CM Long, we prompt the agent to recall keywords discussed over 40 utterances. We evaluate how many of these keywords are recalled.

For social preference: We design questions for three social behavior preferences: positive (Pos.), neutral (Neu.), and negative (Neg.). Group dialogues typically consist of social interactions involving 2 to 10 characters. We analyze the social preference of a character, and identify behaviors aligning with its preference in the dialogues as correct answers. For example, members with a positive social preference tend to engage in behaviors beneficial to the group, such as encouraging teamwork or mediating conflicts within the group. Members with a neutral social preference tend to adopt neutral behaviors within the group, such as aligning with the majority opinion or maintaining a neutral stance in conflicting viewpoints. Conversely, members with a negative social preference tend to engage in behaviors detrimental to the group, such as criticizing others' viewpoints or engaging in competition and arguments with group members.

We analyze the social preference of each character to design negative options. Behaviors contradicting its social preference serve as negative options. For instance, for a character inclined towards teamwork, we would construct exclusionary behaviors as negative options.

A.3 Human Annotation Process

As shown in Figure 4, if all annotators agree on the annotation, it will be selected; if at least two annotators disagree on the annotation, it will be discarded; if only one annotator disagree on the annotation, the question undergoes secondary check by the fourth annotation, it will be modified then selected or be discarded directly. The verification rule for each construction procedure is listed as below:

Prompts for Constructing Inserted Dialogue
Role Profile: {role_profile} Previous Dialogue: ..... Assistant {assistant_utterance} User: {user_utterance} Please follow the provided profile of {role_name}, generate multi-turns conversations between the User and the Assistant. The generated dialogue should be unrelated to the previously given dialogue content, ensuring diverse and realistic conversation topics while adhering to persona of {role_name}.

Table 5: Prompts for constructing inserted dialogue.

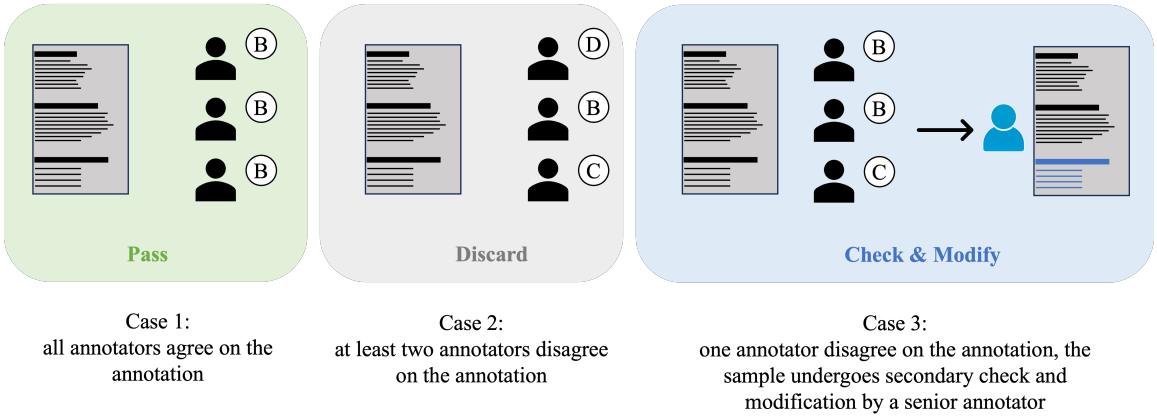

Figure 4: Human annotation process.

Profile Verification: We assess the personality contradiction and knowledge hallucination of profiles to ensure character properness.

Dialogue Verification: Our focus lies in ensuring that dialogues adhere to the principles of dialogue fluency and character fidelity.

Question Verification: We verify whether each multi-choice option or labeled entity is correct.

For annotators recruiting, we recruit annotators from crowdsourcing companies, and the annotation wages are evaluated and confirmed by the crowdsourcing company. The annotators mainly consist of undergraduate students.

B Related Work

B.1 Role-Playing LLMs

Leveraging the powerful capabilities of open-source foundational models, numerous efforts have emerged to develop models specifically tailored for role-playing tasks. These approaches can be categorized based on training paradigms: 1. Supervised fine-tuning (SFT). Li et al. (2023); Wang et al. (2023b); Tu et al. (2023) involve constructing specialized persona training datasets and utilizing supervised fine-tuning; 2. Integration of offline reinforcement learning. Shea and Yu (2023) combines role-playing model training with offline reinforcement learning techniques; 3. Incorporation of retrieval-enhanced methods. Salemi et al. (2023) combines role-playing model training with retrieval-enhanced methods.

B.2 Role-Playing Benchmarks

With the rapid development of role-playing LLMs, there has been a corresponding growth in evaluation datasets. (Chen et al., 2023) introduced a specific dialogue dataset constructed from the Harry Potter series, to examine the model's ability to align with the story characters. (Wang et al., 2023b) constructed the first fine-grained role-playing datasets with 100 roles, to measure the role-specific knowledge, memory and speaking style. (Shao et al., 2023) introduced a experience upload method, to

Prompt for Group Dialogue Generation
Profile of {role_name_a}: {role_name_a_profile} Profile of {role_name_b}: {role_name_b_profile} Profile of {role_name_c}: {role_name_c_profile} ..... Example Dialogue: {role_name_a}: {role_name_a_utterance_1} {role_name_b}: {role_name_b_utterance_1} {role_name_c}: {role_name_c_utterance_1} ..... {role_name_a}: {role_name_a_utterance_n} {role_name_b}: {role_name_b_utterance_n} {role_name_c}: {role_name_c_utterance_n} Follow the Dialogue Format, generate multi-turn dialogue between {role_name_a} and {role_name_b} and {role_name_c} ..... Ensure that each character adheres to their respective personality. The order of dialogue participants can be altered. Aim for as many dialogue turns as possible. Dialogue scene description: {dialogue_topic}

Table 6: Prompts for group dialogue generation.

test the model's effectiveness on memorizing the character knowledge, values and personality. (Shen et al., 2023) introduced a bilingual role evaluation benchmark to assess the memorization, utilization, and reasoning capabilities of role knowledge. (Tu et al., 2024) proposed a Chinese benchmark for role-playing conversational agent, to evaluate the agent's conversation ability, character consistency and role-playing attractiveness. While previous work mainly focuses on testing the agent's abilities on imitating the character's role-specific knowledge, memory or speaking style, SocialBench introduces the first-ever evaluation benchmark for the sociality of role-playing conversational agents encompassing both individual and group level.

C Dataset Statistic

SocialBench consists of 30,871 multi-turn role-playing utterances, 6,420 questions, 512 characters, 1,480 scenarios.

C.1 Character Types and Personality Traits

Drawing from the definition outlined by Shen et al. (2023) and amalgamating existing categorizations from role-playing platforms such as Xingchen [4], CharacterAI [5], Xingye [6], and Wantalk [7], we synthesize character types into six dimensions: cartoons, celebrities, movies, games, fiction, and others. SocialBench utilizes this classification standard for the collection and construction of character pro-

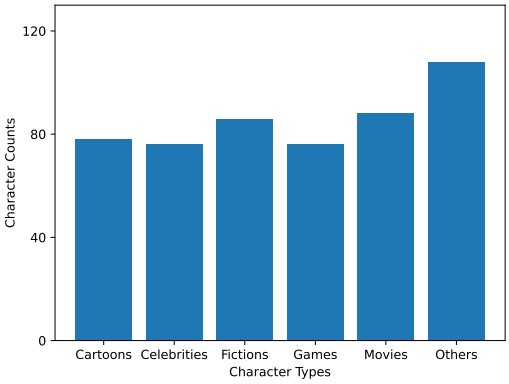

Figure 5: Character types in SocialBench.

[4] https://tongyi.aliyun.com/xingchen/
[5] https://beta.character.ai/
[6] https://www.xingyeai.com/
[7] https://www.wantalk.com/

	Individual Level						Group Level		
	SA Style	SA Know.	EP Situ.	EP Emo.	CM Short	CM Long	Pos.	Neu.	Neg.
#Questions	1,063	1,750	193	1,016	1,065	1,167	586	724	606
Avg Utterances	17.9	9.4	1.0	6.4	24.5	54.4	15.6	16.1	16.0
Avg Tokens per Utterance	32.6	66.7	649.5	23.0	37.6	41.2	38.8	38.7	42.0
Avg Characters per Question	2	2	-	-	2	2	6.3	6.5	6.7
#Characters	512								
#Total Questions	6,420								
#Total Utterances	30,871								

Table 7: Statistic of SocialBench.

Positive Traits			Neutral Traits			Negative Traits		
Adventurous	Articulate	Attractive	Absentminded	Aggressive	Amusing	Abrasive	Aloof	Angry
Calm	Caring	Cheerful	Complex	Conservative	Contradictory	Argumentative	Arrogant	Impersonal
Confident	Courageous	Curious	Emotional	Formal	Neutral	Barbaric	Blunt	Childish
Elegant	Humble	Humorous	Mystical	Ordinary	Old-fashioned	Cowardly	Cruel	Fatalistic
Kind	Logical	Optimistic	Stylish	Tough	Whimsical	Gloomy	Lazy	Shy
Passionate	Warm	Witty	Questioning	Sensual	Dry	Envious	Hostile	Melancholic

Table 8: Personality traits in SocialBench.

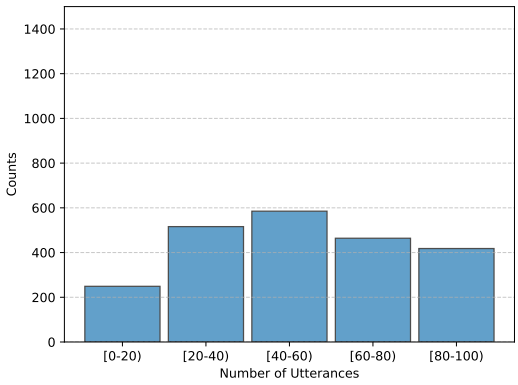

Figure 6: Distribution of the number of questions across different numbers of utterances in the conversation memory dimension.

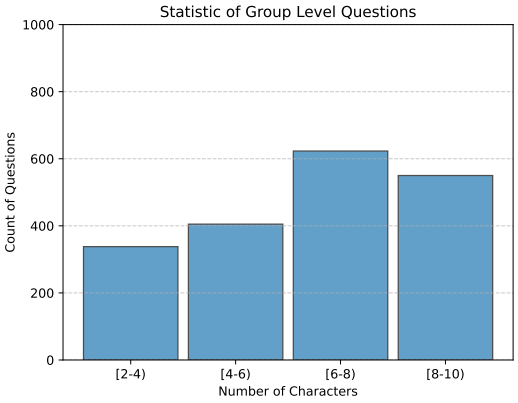

Figure 7: The distribution of the number of questions across different numbers of roles, in the group level.

files, ensuring comprehensive and representative coverage. Furthermore, we aim for an equitable distribution of character types within SocialBench. Please refer to Figure 5 for detailed categorizations.

We follow the definition of personality traits in Gunkel (1998) to construct profiles, ensuring diversity and comprehensiveness in SocialBench. From the collection of 638 personality descriptors created by Gunkel (1998), we selected a subset of easily understandable terms for construction. These selected terms can be categorized into positive, neutral, and negative traits, as illustrated in Table 8.

C.2 Individual Level and Group Level

SocialBench consists of two dimensions: individual level and group level. The individual level comprises six sub-dimensions, while the group level comprises three sub-dimensions, as shown in Table 7. Individual level can be split into self-awareness on role style (SA Style), self-awareness on role knowledge (SA Know.), emotional perception on situation (EP Situ.), emotional perception on dialogue emotion (EP Emo.), short-term conversation memory (CM Short), and long-term conversation memory (CM Long). We present the distribution of the number of questions across different numbers of utterances in the conversation memory dimension, as shown in Figure 6. Group level is split into positive (Pos.), neutral (Neu.) and negative (Neg.). We also present the distribution of the number of questions across different numbers of roles in the group level, as shown in Figure 7.

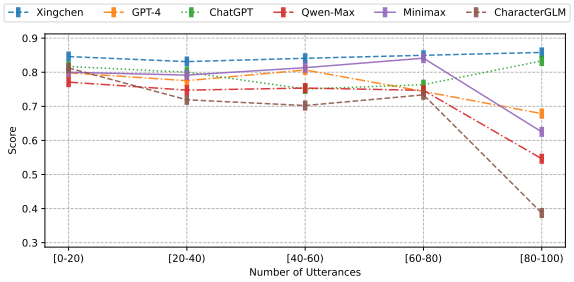

Figure 8: Performance w.r.t the number of utterances.

D Experiment Settings

D.1 Evaluation Metrics

Most of the previous methods (Wang et al., 2023b; Shao et al., 2023) for role-playing applications rely on ChatGPT or GPT4 for evaluation, which may suffer from questionable accuracy on the role-playing scenario and costly API usage. We follow the popular benchmark MMLU (Hendrycks et al., 2020) and C-Eval (Huang et al., 2023), and construct multi-choice format prompt for automatic and fast evaluation free from LLMs. SocialBench utilizes fully automatic evaluation metrics, employing both multiple-choice and open-domain generation questions. Accuracy is computed for multiple-choice questions, while for open-domain generation questions, the proportion of keywords mentioned in the response relative to the answer is calculated.

D.2 Models

We conduct evaluation on the current mainstream open-source and closed-source LLMs. For evaluation of open-source LLMs, we choose LLaMA-1-7B (Touvron et al., 2023a), LLaMA-2-7B (Touvron et al., 2023b), Mistral-7B (Jiang et al., 2023), Qwen-7B (Bai et al., 2023). For evaluation of closed-source LLMs, we choose Minimax [8], Qwen-Max [9], CharacterGLM (Zhou et al., 2023), GPT-4 (OpenAI, 2023), ChatGPT (OpenAI, 2022), and Xingchen [10] for testing.

E Results and Analysis

E.1 Conversation Memory for Role-Playing

Enhanced memory for dialogue contributes to role-playing agents forming dynamic portraits of other individuals or recalling social interaction histories, which is a fundmental and critical capability. As shown in Figure 8 (above), we observe a significant drop in memory capability for most models after surpassing 80 rounds (with average 42 tokens per round). However, Xingchen and ChatGPT perform relatively well in this regard.

F Data Utilization and Terms of Use

We utilized the open-source datasets (Chen et al., 2022; Hsu et al., 2018; Garbowicz, 2021; Gong et al., 2020), with their terms of use specifying research purposes only. Similarly, we employed the weights of open-source models and the APIs of closed-source models, strictly adhering to their respective usage agreements for research purposes. Regarding our dataset, it is also restricted to research purposes. We conducted thorough manual checks to ensure the absence of security and offensive issues, particularly sensitive personal information such as phone numbers and home addresses.

[8] https://api.minimax.chat/

[9] https://help.aliyun.com/zh/dashscope/developer-reference/api-details

[10] https://xingchen.aliyun.com/