

SocialBench: Sociality Evaluation of Role-Playing Conversational Agents

Hongzhan Chen¹, Hehong Chen², Ming Yan^{2*}, Wenshen Xu², Xing Gao²
Weizhou Shen¹, Xiaojun Quan^{1*}, Chenliang Li², Ji Zhang², Fei Huang²

¹School of Computer Science and Engineering, Sun Yat-sen University, China

²Alibaba Group, China

¹chenhzh59@mail2.sysu.edu.cn, quanxj3@mail.sysu.edu.cn

²ym119608@alibaba-inc.com

Abstract

Large language models (LLMs) have advanced the development of various AI conversational agents, including role-playing agents that mimic diverse characters and human behaviors. While prior research has predominantly focused on enhancing the conversational capability, role-specific knowledge and style 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 agents at both individual and group levels of social interactions. SocialBench is constructed from various sources and covers a wide range of 500 characters and over 6,000 question prompts and 30,800 multi-turn role-playing utterances. We conducted comprehensive evaluations on this benchmark using mainstream LLMs. We find that agents excelling at the individual level do not necessarily demonstrate proficiency at the group level. Experimental results on SocialBench confirm its significance as a testbed for assessing the social interaction of role-playing agents. The benchmark is publicly accessible at <https://github.com/X-PLUG/RoleInteract>.

1 Introduction

Recently, role-playing applications powered by LLMs, such as Character.AI¹, have gained significant attention. A growing number of research efforts have been dedicated to developing LLM-based role-playing agents, aiming to mimic diverse characters and human behavior (Shao et al., 2023; Tu et al., 2024; Tian et al., 2023).

As an rapidly developing area, the evaluation of role-playing agents is becoming increasingly important. Wang et al. (2023c) collected a role-specific dataset and utilized LLMs 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 attractiveness. While these works mainly focus on evaluating the agent’s individual abilities, this study aims to explore and measure the *sociality* of role-playing 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 social interaction of role-playing agents. As introduced in (Troitzsch, 1996), the agent society represents a complex system comprising individual and group social activities. Following this definition, SocialBench assesses the sociality at both the individual and group levels, as illustrated in Figure 1. At the individual level, the agent should possess the basic social intelligence as individuals, such as self-awareness on role description (Tu et al., 2024; Shen et al., 2023), 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 social intelligence of the role-playing agents within group social interactions, 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 500 characters and 6,000 questions, and 30,800 multi-turn role-playing utterances. We conducted extensive evaluations on SocialBench using mainstream open-source and closed-source LLMs. We find that agents excelling at the individual level do not necessarily demonstrate proficiency at the group level. Moreover, the behavior of individuals may *drift* as a result of the influence exerted by other agents within the group. We hope our findings will inspire future research.

* Corresponding authors.

¹<https://beta.character.ai>

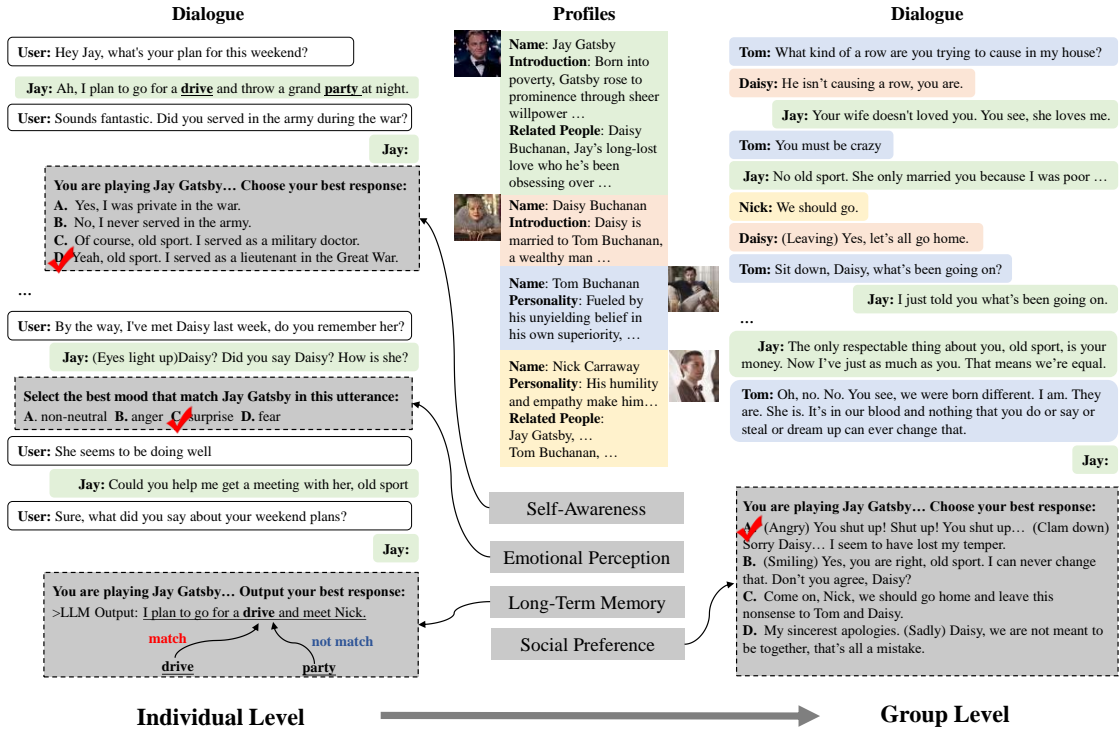


Figure 1: An example about SocialBench, which is partially constructed from the film “The Great Gatsby”.

2 Sociality of Role-Playing Agent

Given a character profile and context, the sociality of role-playing agents focuses on imitating typical social interactions from individual to group level.

2.1 Individual Level

At the individual level, the role-playing 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., 2023a; 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 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 a group may be influenced by member interactions, demonstrating more sophisticated social behaviors. This represents a higher calling for the sociality of role-playing agents.

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 towards group dynamics (Amir et al., 2022). Social preferences are the preferences of individuals regarding the payoffs or well-being of others (Charness and Rabin, 2002), and individuals behave prosocially on the basis of their social preferences (Murphy et al., 2014). Social agents need to exhibit and keep their pre-designed social preference or group identity when confronted with diverse and more sophisticated group conversations.

3 SocialBench

In this section, we introduce the construction of SocialBench. Refer to Appendix A for more details.

3.1 Profile Collection

A character profile defines role style, knowledge, and social preference. We gather profiles for role-playing agents from various sources including nov-

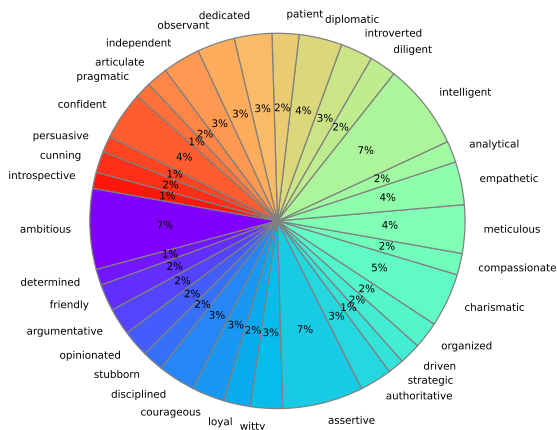


Figure 2: Personality traits distribution in SocialBench.

els, scripts, online platforms such as CharacterAI², and automatic generation via GPT-4 prompting. To ensure diversity, we construct profiles based on various character types and personality traits by combining the existing categorizations in research work (Shen et al., 2023; Gunkel, 1998). Figure 2 illustrates the distribution of personality traits for roles within SocialBench.

3.2 Dialogue Construction

Constructed dialogue adheres to two principles: *dialogue fluency*, which ensures natural and coherent conversations; and *character fidelity*, meaning characters in the dialogue must adhere to their 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: We prompt general LLMs to role-play characters and engage users to generate dialogue data. (4) Fully automatic self-dialogue generation with general LLMs: We task general LLMs to role-play and engage in self-dialogue for data collection. We consider various social tasks or scenarios in dialogue construction. Prompts for extracting dialogue can be found in Appendix A.1.

3.3 Question Design

For Self-Awareness: This includes two subcategories: self-awareness on role style (SA Style) and 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://character.ai>

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 utilize 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: short-term (CM Short) and long-term (CM Long) conversation memory. For CM Short, we prompt agent to recall keywords within 40 utterances, and for CM Long, over 40 utterances. How many keywords recalled are evaluated.

For Social Preference: We design questions for: positive (Pos.), neutral (Neu.), and negative (Neg.) preferences. Group dialogues typically consist of social interactions involving 2 to 10 characters. We analyze the preference of agent and identify behaviors aligning with its preference as correct answers. Behaviors contradicting its designed preference serve as negative options. Other agents in group also have their own social behavior preferences, which can mutually influence each other. Details for question design can be found in Appendix A.3.

3.4 Dataset Validation

We undergo multiple iterations of rigorous manual screening, annotation, and refinement. To eliminate the impact of subjectivity, we employ three distinct annotators for each sample labeling. A secondary check is conducted by a senior annotator when label disagreements arise. More detail about dataset validation can be found in Appendix A.4.

4 Experiment Settings

4.1 Dataset Statistic

We show the statistics of SocialBench in Table 1 and the distribution of dialogue token length in Figure 3. SocialBench consists of 500 roles, encompassing 6,000 questions and 30,800 utterances.

4.2 Evaluation Metrics

Automated evaluation metrics are employed for SocialBench, as listed in Table 1. For single-answer questions, we calculate the accuracy (Acc_{single}) using the following formula:

$$Acc_{single} = \frac{\#correctly\ chosen\ options}{\#questions}. \quad (1)$$

Metrics	Individual Level						Group Level		
	SA Style	SA Know.	EP Situ.	EP Emo.	CM Short	CM Long	Pos.	Neu.	Neg.
	Acc_{single}	Acc_{single}	$Acc_{multiple}$	Acc_{single}	$Cover$	$Cover$	Acc_{single}	Acc_{single}	Acc_{single}
#Questions	1,063	1,408	193	1,016	773	1,348	586	724	606
Avg Utterances	17.9	9.4	1.0	6.4	23.9	76.7	15.6	16.1	16.0
Avg Tokens per Utterance	32.6	66.7	286.3	23.0	37.6	41.2	38.8	38.7	42.0
Avg Characters per Question	2	2	N/A	N/A	2	2	6.3	6.5	6.7

Table 1: Metrics and statistics of SocialBench.

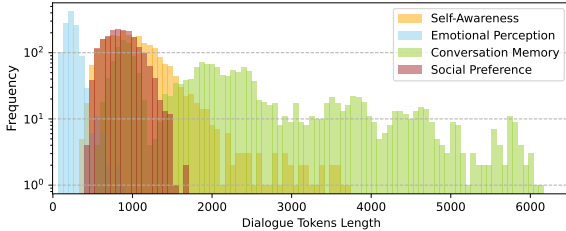


Figure 3: Distribution of dialogue tokens across four dimensions in SocialBench, based on Qwen tokenizer.

For multiple-answer questions, we calculate the accuracy ($Acc_{multiple}$) using the following formula:

$$Acc_{multiple} = \sum_i^N \frac{Score_i}{MaxScore_i}, \quad (2)$$

where N is the total number of multiple-answer questions. $Score_i$ is the score obtained for the i th question, considering both correct and partially correct options. $MaxScore_i$ is the maximum achievable score for the i th question. Detailed explanation can be found in Appendix C.1

For open-domain questions, we calculate the keyword coverage rate ($Cover$). Given a label keyword set $\mathbf{A}_{keywords} = \{k_1, k_2, \dots, k_n\}$, and a response keywords set $\mathbf{R}_{keywords}$, we compute:

$$Cover(\mathbf{R}) = \frac{len(\mathbf{A}_{keywords} \cap \mathbf{R}_{keywords})}{len(\mathbf{A}_{keywords})}, \quad (3)$$

where $Cover(\cdot)$ quantifies the proportion of keywords mentioned in the response \mathbf{R} relative to the keywords identified in the \mathbf{A} .

4.3 Models

We conducted evaluation on mainstream open-source and closed-source LLMs. For open-source LLMs, we selected the chat versions of LLaMA-2-7B/13B/70B (Touvron et al., 2023), the instruction version of Mistral-7B (Instruct-V0.2) (Jiang et al., 2023), and the chat versions of Qwen-7B/14B/72B (Bai et al., 2023). For closed-source LLMs, we chose Minimax (abab5.5s-chat and abab6-chat)³, GLM (CharGLM-3 and GLM-3-Turbo) (Zhou et al., 2023a), Baichuan (Baichuan-NPC-Turbo and

³<https://api.minimax.chat/>

Baichuan-2-Turbo)⁴, Qwen-Max⁵, GPT-4-Turbo (OpenAI, 2023), GPT-3.5-Turbo (OpenAI, 2022), and Xingchen-Plus⁶.

5 Results and Analysis

5.1 Overall Results

As presented in Table 2, the performance of closed-source models tends to surpass open-source models. Moreover, models specifically designed for role-playing, such as Xingchen-Plus, outperform others. While the general model GPT-4-Turbo also demonstrates impressive performance. However, models like Baichuan-NPC-Turbo and Minimax-abab5.5s, tend to underperform compared to their general counterparts, such as Baichuan-2-Turbo and Minimax-abab6-chat. We find that they are biased towards character-based dialogues, leading to poorer understanding and compliance with instructions. Thus, it is essential for role-playing agents to maintain character-based dialogue abilities and general instruction-following capabilities. At the individual level, dimensions such as SA Style, SA Know., and CM Short are well-performed by most models. However, some models tend to exhibit poor performance in EP Situ., EP Emo., and CM Long. At the group level, most models perform poorly due to the complexity of group dynamics. While models generally align well with tendencies towards positive social preference, there is a notable absence of necessary abilities to embody neutral and negative social preferences.

5.2 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 shown in Figure 4, with increasing complexity of group dynamics, the performance shows a downward trend. We find that excelling in simple

⁴<https://npc.baichuan-ai.com/index>

⁵<https://help.aliyun.com/zh/dashscope/developer-reference/api-details>

⁶<https://xingchen.aliyun.com/>

Models (Max Length)	Individual Level						Group Level			Avg
	SA Style	SA Know.	EP Situ.	EP Emo.	CM Short	CM Long	Pos.	Neu.	Neg.	
<i>Open-Source Models</i>										
LLaMA-2-7B-Chat (4k)	48.76	51.23	31.23	28.91	25.38	21.89	44.98	24.19	27.67	33.80
LLaMA-2-13B-Chat (4k)	57.62	65.51	37.12	32.56	30.43	29.82	66.38	42.25	26.27	43.11
LLaMA-2-70B-Chat (4k)	67.61	70.78	35.74	38.47	45.57	26.74	69.87	45.29	39.37	48.83
Mistral-7B (8k)	50.12	61.17	36.48	31.72	31.78	25.42	65.67	46.34	28.96	41.96
Qwen-7B-Chat (32k)	66.44	71.16	41.68	40.68	67.45	53.45	75.61	52.78	43.11	56.93
Qwen-14B-Chat (32k)	77.06	86.15	45.71	43.78	65.32	51.37	78.32	58.25	59.21	62.80
Qwen-72B-Chat (32k)	83.87	90.64	53.10	52.89	<u>83.29</u>	73.15	<u>91.53</u>	73.44	63.82	73.97
<i>Closed-Source Models</i>										
GPT-4-Turbo (128k)	84.57	93.11	56.48	53.05	81.39	80.11	89.73	81.69	<u>75.10</u>	77.25
GPT-3.5-Turbo (16k)	73.17	73.82	52.44	45.49	73.03	59.72	81.59	76.79	54.16	65.58
Qwen-Max (8k)	82.04	93.34	61.14	52.36	76.45	72.65	87.22	72.14	52.19	72.17
Xingchen-Plus (8k)	85.43	91.6	55.44	60.73	82.43	80.69	94.27	86.69	77.26	79.39
Baichuan-NPC-Turbo (unknown)	53.69	61.67	52.14	43.34	76.47	22.40	62.09	48.91	34.59	50.59
Baichuan-2-Turbo (unknown)	77.75	83.35	55.7	47.38	80.11	78.91	87.37	74.71	68.50	72.64
CharGLM-3 (unknown)	74.70	79.41	26.23	41.27	81.16	68.29	84.40	70.45	36.36	62.47
GLM-3-Turbo (128k)	77.85	84.62	35.58	<u>53.05</u>	74.64	71.68	84.41	67.47	54.55	67.09
Minimax-abab5.5s-chat (8k)	36.09	42.11	28.15	<u>47.97</u>	29.55	19.30	44.59	41.04	22.45	34.58
Minimax-abab6-chat (32k)	82.92	87.45	35.90	51.38	83.60	<u>80.26</u>	89.12	79.55	74.65	73.87

Table 2: Main results from SocialBench. Best performances are shown in **bold**, while suboptimal ones underlined.

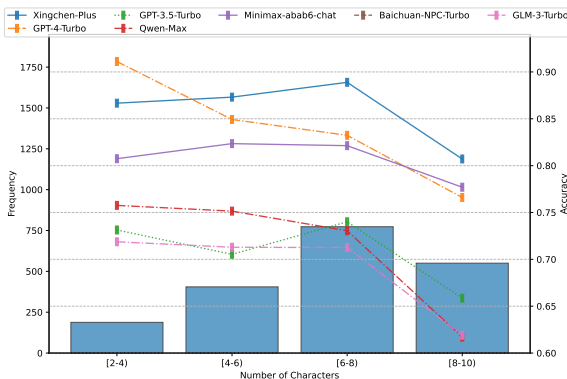


Figure 4: Performance w.r.t number of group members.

group dynamics does not necessarily imply proficiency in more complex group dynamics. For example, models like GLM-3-Turbo and GPT-4-Turbo perform well in simple dynamics, but this doesn't guarantee strong performance in complex dynamics. However, models like Xingchen-Plus and Minimax-abab6-chat can also demonstrate proficiency in handling complex group dynamics.

5.3 Impact of Group Dynamics Polarity

Role-playing agents need to maintain designed social preferences under the influence of varying group dynamics. The group dynamics polarity is defined as the majority social preference of members. For instance, positive group dynamics imply that the majority of members exhibit positive social preference. For an individual with a specific preference, different polarities of group dynamics may have various impacts. As shown in Figure 5, we find that individuals with neutral and negative social preferences perform optimally within their corresponding group polarities. However, they

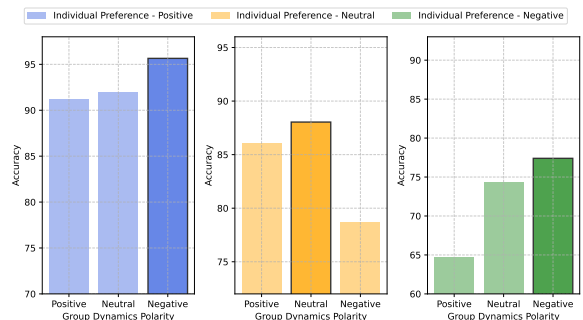


Figure 5: Performance of Xingchen-Plus under different group dynamics polarities on a subset of group data.

are susceptible to the influence of group dynamics with different polarities and undergo a phenomenon termed *preference drift*, leading to deviation from their original designed behaviors, as indicated by the decline of performance. Nevertheless, individuals with positive preference appear to be more resilient to the preference drift, performing better across all group polarities.

6 Conclusion

In this paper, we introduced SocialBench, the first evaluation benchmark designed to systematically assess the social intelligence of role-playing agents at both individual and group levels. We constructed diverse question prompts on a wide range of characters covering comprehensive dimensions. Moreover, rigorous human verification ensures the questions' difficulty and validity. We conducted extensive experiments and analysis on SocialBench. While role-playing agents perform well at the individual level, their social interaction capabilities at the group level are lacking. This highlights the need for further exploration in future research.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62176270) and the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012832).

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

- Khushk Amir, Zengtian Zhang, Hui Yang, and Atamba Cynthia. 2022. [Understanding group dynamics: Theories, practices, and future directions](#). *Malaysian E Commerce Journal*, 6:1–08.
- 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.
- Gary Charness and Matthew Rabin. 2002. Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3):817–869.
- Yirong Chen, Weiwan 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.
- Antônio Carlos da Rocha Costa. 2019. *A Variational Basis for the Regulation and Structuration Mechanisms of Agent Societies*. Springer.
- Chen Gao, Xiaochong Lan, Zhi jie Lu, Jinzhu Mao, Jing Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. [S3: Social-network simulation system with large language model-empowered agents](#). *ArXiv*, abs/2307.14984.
- 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. [Emotion-Lines: 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 Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *ArXiv*, abs/2310.06825.
- 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, et al. 2023. [Chatharuhi: Reviving anime character in reality via large language model](#). *arXiv preprint arXiv:2308.09597*.
- Murphy, Ryan O, and Kurt A Ackermann. 2014. Social value orientation: theoretical and measurement issues in the study of social preferences. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology*.
- G. Nigel Gilbert and Klaus G. Troitzsch. 1997. [Social science microsimulation](#). *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 56(1):71–78.
- 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. *arXiv preprint arXiv:2304.11406*.
- 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. *arXiv preprint arXiv:2310.10735*.
- 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, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- 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. *arXiv preprint arXiv:2308.10278*.
- Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. [CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation](#). *arXiv preprint arXiv:2401.01275*.
- 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.
- Xintao Wang, Quan Tu, Yaying Fei, Ziang Leng, and Cheng Li. 2023b. [Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots](#). *CoRR*, abs/2310.17976.
- Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023c. [RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). *arXiv preprint arXiv:2310.00746*.
- Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. [Exploring large language models for communication games: An empirical study on werewolf](#). *ArXiv*, abs/2309.04658.
- 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. 2023a. [CharacterGLM: Customizing Chinese conversational AI characters with large language models](#). *ArXiv*, abs/2311.16832.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2023b. [Sotopia: Interactive evaluation for social intelligence in language agents](#). *ArXiv*, abs/2310.11667.

A Dataset Construction

In this section, we introduce the construction process of SocialBench, as illustrated in Figure 6

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 method involves extracting character dialogues from novels and scripts. Dialogues obtained through this approach typically preserve the original character interactions and inherently adhere to the two principles.
- The second method involves collecting role-playing LLMs and real user dialogue data from role-playing platforms. Dialogues constructed in this manner reflect interactions between role-playing agents and users in real-world scenarios. Data gathered through this approach largely meets the requirements of dialogue fluency.
- In contrast to the second method, which utilizes professional role-playing platforms, the third method involves role-playing tasks using general LLMs such as GPT-3.5-Turbo and GPT-4, collecting data through interactions with users. While this approach is more efficient in data collection, it may encounter limitations in the role-playing capabilities of general LLMs. Therefore, we will focus more on examining the consistency of the roles in the dialogues collected through this method in later stages.
- The fourth method, a fully automatic approach, involves prompting GPT-4 to engage in self-dialogue by role-playing as both the user and the role-playing agent. This is the most efficient form of collecting dialogue data, leveraging the autonomous capability of general LLMs to simultaneously play the roles of users and role-playing agents in generating dialogue data.

The prompts for role-playing tasks and automatic self-dialogue generation are provided in Figure 7 and Figure 8. For the dimension of long-term conversation memory, we construct lengthy dialogue contexts to increase complexity, thereby

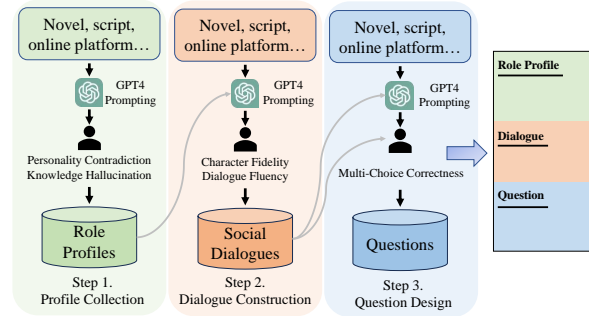


Figure 6: Dataset construction pipeline of SocialBench.

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 Figure 9.

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, and multiple users interacting with multiple role-playing agents. Our primary focus lies on scenarios involving multiple role-playing agents. We employ general LLMs to act as different role-playing agents and generate dialogues between their social interactions. Prompts for automatically generating group conversations can be found in Figure 10.

A.2 Social Scenario Design

we consider various social tasks or scenarios at the group level. When constructing each social dialogue, we provide the explicit topic or social tasks, which include:

- Positive preferences: resolving arguments, coordinating conflicts, mutual assistance, proactive sharing, accepting new members, etc.
- Neutral preferences: preferring to follow the group’s opinions, not preferring to follow the group’s opinions, commanding others, and observing team discussions, etc.
- Negative preferences: opposing others’ opinions, refusing to help others, exacerbating group conflicts, shirking responsibility, etc.

Prompt for Role-Playing Tasks

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."

Figure 7: The prompt for role-playing tasks with GPT-4.

Prompt for Automatic Self-Dialogue Generation

Profile:
{role_profile}

Example Dialogue:
{example_dialogue}

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} :

Figure 8: The prompt for automatic self-dialogue generation.

Prompt for Inserted Dialogue Construction

Profile:
{role_profile}

Previous Dialogue:
{previous_dialogue}

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}.

Figure 9: The prompt for constructing inserted dialogue.

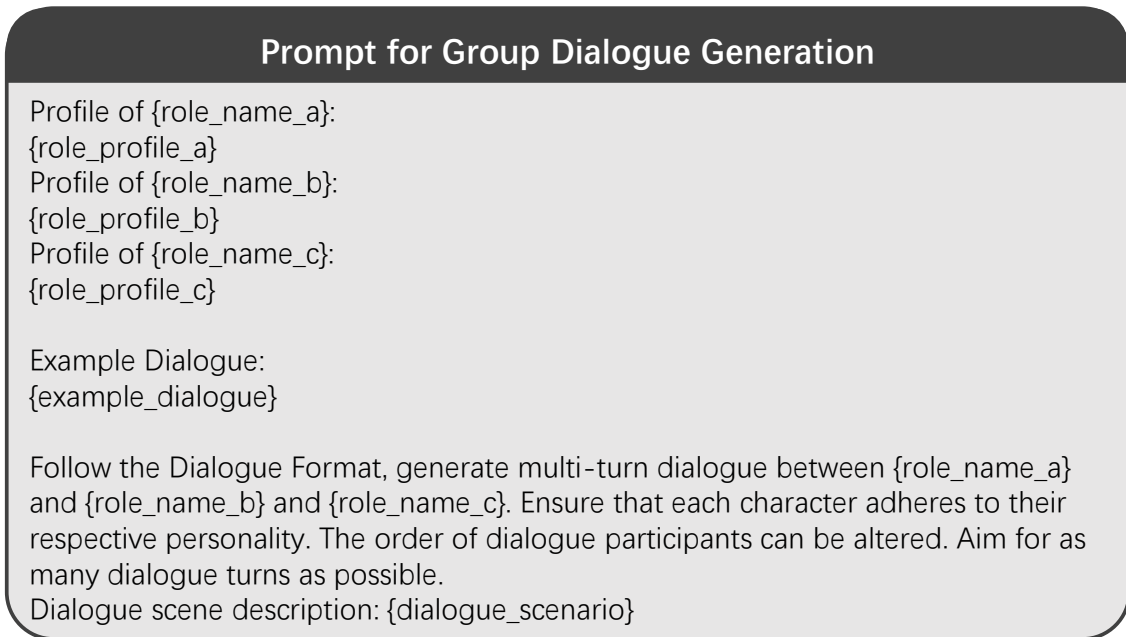


Figure 10: The prompt for group dialogue generation.

A.3 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” indicating that the character’s speaking style is enthusiastic and cheerful. Since the dialogues constructed in the previous step already adhere to the character’s speaking style, we can directly use utterances from the dialogue as correct answers. Additionally, to create negative options, we generate replies with different styles (e.g., “cold”, “impersonal”), indicating that these speaking styles do not align with the current character’s style setting. It is worth noting that while the speaking style changes, we ensure that the replies still adhere to contextual coherence. For SA Know., we identify utterances containing character-related knowledge from the dialogue as correct options. For example, some entity information like “Where were you born?” or “Where is your hometown?” This type of information typically follows the character’s original setting. 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 gather exam questions related to situational understanding in psychological counseling scenarios. We filter these questions to exclude those with strong psychological expertise to ensure the assessment focuses on agents’ general abilities. We manually collect 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 construct emotion understanding data based on open-source datasets and websites. These questions primarily involve agents understanding the psychological states of speakers and interpreting emotions in dialogue. For example, when a speaker says “I hate you”, agents need to determine the emotion of this statement based on the context, whether it’s hate, like, neutral, etc. We further focus on advanced emotional understanding abilities such as humor and irony. Humor data are collected from websites^{7,8} and the DiBERT dataset (Garbowicz, 2021), with non-humorous texts used as negative options. For irony emotion understanding, we utilize binary classification data from the Chinese open-source dataset (Gong et al., 2020) to

⁷<https://www.toutiao.com/>

⁸<https://www.sohu.com/>

Prompt for Conversation Memory Question Generation

In-context Case:

{case}

Profile:

{role_profile}

You should generate another similar dialogue history between a user and {role_name}, given the profile of {role_name}.

In the conversation, the user will ask {role_name} a question, {role_name} responds, and after multiple rounds, the user will ask related questions again, requesting the {role_name}'s answers to be consistent with the previous ones. (The dialogue should be at least 40 rounds)

The dialogue topic: {dialogue_topic}

Figure 11: The prompt for question generation on conversation memory dimension.

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 increase the difficulty of the conversation memory dimension, we utilize an open-domain generation combined with keyword matching approach for this dimension. For example, if an agent previously answered that they had a sandwich for breakfast, after several rounds of conversation, if the user asks again what the agent had for breakfast, we require the agent's response to include the keyword "sandwich". If the agent responds that they had bread for breakfast, since the keyword does not match, we consider the agent unable to correctly recall their previous dialogue content. We show the prompt for GPT-4 to automatically identify the keywords and generate the questions in Figure 11. 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. Prompt for question generation on social preference is provided in Figure 12.

A.4 Dataset Validation

The validation stage includes two parts: dataset pre-validation and post-validation. Throughout this process, we undergo multiple iterations of rigorous manual screening, annotation, and refinement.

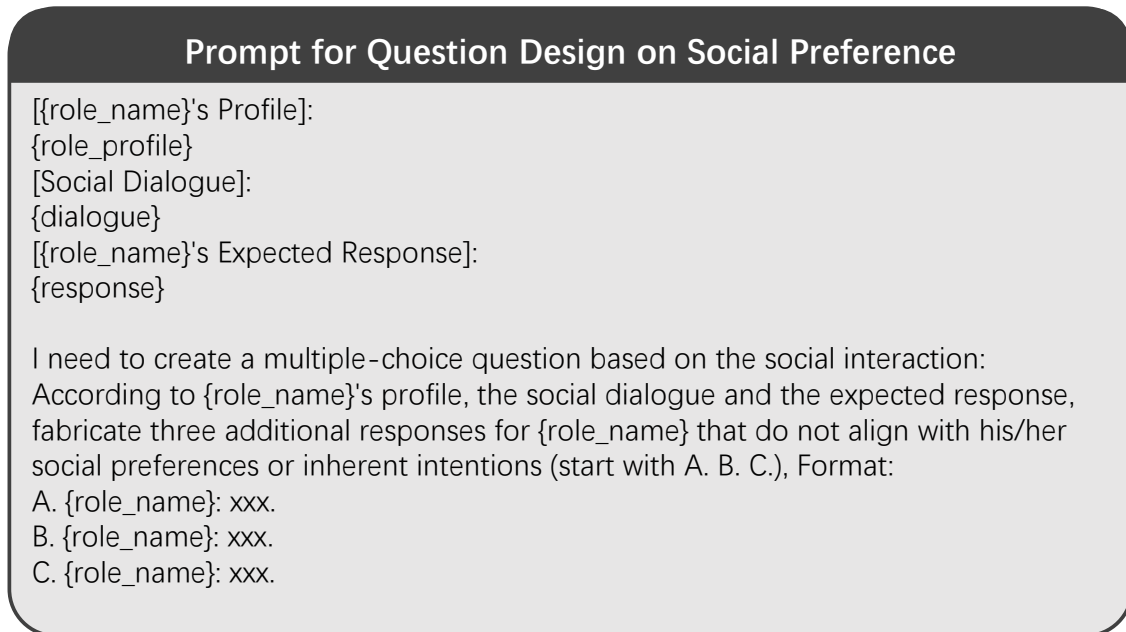


Figure 12: The prompt for question generation on social preference dimension.

A.4.1 Dataset Pre-Validation

Profile Verification: After profile collection, we assess personality contradictions and knowledge hallucinations in profiles to ensure character accuracy. We manually review and modify any erroneous descriptions in profiles, while also ensuring the exclusion of specific personal information such as phone numbers and home addresses.

Dialogue Verification: Our focus is on ensuring dialogues adhere to principles of *dialogue fluency* and *character fidelity*. For fluency, we manually inspect dialogues for contextual coherence and natural expression. For fidelity, we analyze the speaker's profile to verify if the utterance aligns with the character's speaking style and behavior. Dialogues that do not meet requirements undergo manual correction.

Question Verification: For multiple-choice questions, we invite three different annotators to label each question. As shown in Figure 13, 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 disagrees on the annotation, the question undergoes secondary check by the fourth annotation, it will be modified then selected or be discarded directly. For open-domain generation questions, we verify the correctness and validity of keywords provided. Invalid questions are either modified by experts or discarded.

A.4.2 Dataset Post-Validation

We undergo the post-validation process after completing each round of dataset. Different dimensions require different validation strategies.

Validation for Self-Awareness: We focus on examining knowledge-related errors in the questions and options, particularly those generated by LLMs that may give rise to knowledge hallucinations. We remove questions that do not meet the requirements, while options that do not meet the requirements will be flagged for correction in the subsequent iteration.

Validation for Emotional Perception: Some of the questions we collect are sourced from professional psychology exams, which may include highly specialized content not conducive to assessing the basic abilities of role-playing agents. Therefore, we filter out samples that are too focused on psychology-specific knowledge, retaining those that are more general and fundamental for role-playing agents.

Validation for Conversation Memory: In this dimension, we've observed that questions containing pronouns (such as "him", "it", "she") often result in unclear or ambiguous references to preceding context. Therefore, we remove questions containing pronouns to prevent ambiguity. Additionally, we assess the validity of extracted keywords to ensure they are proper nouns, thereby avoiding mismatches caused by different verb tenses.

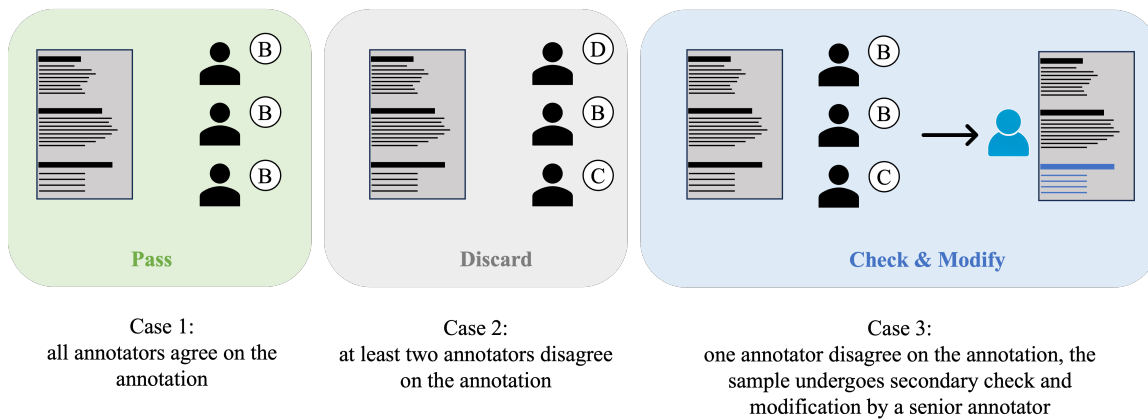


Figure 13: Human annotation process.

Annotation Instruction for Role Style

Given a character, based on its profile and dialogue history, select the most suitable response from options A, B, C, or D that best matches the character's speaking style/personality. To reduce workload, it's recommended that the same person handles all annotations for a specific character. Requirements:

- ✓ Each sample should be annotated by three different individuals without communication between them;
- ✓ If there are any problematic samples, they can be skipped and discarded.

Here are some examples for learning: [Omitted for brevity]

Figure 14: Annotation instruction for role style dimension.

Validation for Social Preference: We find that the options within this dimension may exhibit similarities, making it difficult to distinguish correct option from the negative ones. To reduce difficulty, we manually examine the similarity between options. For options with excessively high similarity, we increase the differentiation between negative options and the correct answer. For instance, if the correct option has a positive preference, we select negative preference content with significantly different characteristics as negative options.

Understanding social contexts is important for role-playing agent. To delve into the verification of the agent's social understanding within current social contexts, we prompt GPT-4 to justify its social choices in the third step of the SocialBench construction pipeline, thereby validating its understanding of the current social context. The reasoning process behind these social choices can then be manually verified, we leave this to our future work.

B Human Annotation

B.1 Annotator Recruiting

For annotators recruiting, we recruit annotators from crowdsourcing company, and the annotation

wages are evaluated and confirmed by the crowdsourcing company. Regular annotators consist of university graduates or higher, whereas senior annotators specialize in role-playing tasks with extensive experience. Our data verification process includes pre-verification (profile verification, dialogue verification, question verification) and post-verification (verification for self-awareness, emotional perception, conversation memory, and social preference). We show the number of annotators at each verification process in Table 3.

B.2 Annotation Instruction

The annotation instructions are designed to initially validate data samples, then annotate based on rules defined for each dimension, and provide examples for annotators' learning (examples omitted for brevity). We provide instructions for question annotation on role style dimension (Figure 14), role knowledge dimension (Figure 15), emotional perception dimension (Figure 16), conversation memory dimension (Figure 17), and social preference dimension (Figure 18).

Annotation Instruction for Role Knowledge

Given an original "answer" and a "question," modify one entity in the original sentence (such as time, location, etc.) to alter the meaning of the sentence. Provide three modified sentences after the modification. Requirements:

- ✓ While ensuring the modifications are reasonable, try to minimize the extent of the changes to increase difficulty;
- ✓ The modified "answer" should still be appropriate for the given "question";
- ✓ If there are instances where annotation is not possible, they can be skipped.

Here are some examples for learning: [Omitted for brevity]

Figure 15: Annotation instruction for role knowledge dimension.

Annotation Instruction for Emotional Perception

Given a "question", based on your judgment, is the question too biased towards psychology? Annotation criteria are as follows:

- ✓ Can you understand the meaning of the question?
- ✓ Does the question contain overly specialized psychological terms?

Figure 16: Annotation instruction for emotional perception dimension.

Annotation Instruction for Conversation Memory

Given a character, based on its dialogue history, check the question and provided keywords for appropriateness, requiring:

- ✓ Questions should have answers that can be found in the preceding text; samples failing this criterion are discarded.
- ✓ Keywords should contain entity information, such as names of people or objects. Keywords containing non-entity information may be modified or discarded.

Here are some examples for learning: [Omitted for brevity]

Figure 17: Annotation instruction for conversation memory dimension.

Annotation Instruction for Social Preference

Given a character, based on its profile and dialogue history, select the most suitable response from options A, B, C, or D that best matches the character's social interaction habits. To reduce workload, it's recommended that the same person handles all annotations for a specific character. Requirements:

- ✓ Each sample should be annotated by three different individuals without communication between them;
- ✓ If there are any problematic samples, they can be skipped and discarded.

Here are some examples for learning: [Omitted for brevity]

Figure 18: Annotation instruction for social preference dimension.

Verification Process	Number of Annotators
Profile Verification	5-6
Dialogue Verification	10-12
Question Verification	10-12
Verification for Self-Awareness	1-3
Verification for Emotional Perception	1-3
Verification for Conversation Memory	1-3
Verification for Social Preference	1-3

Table 3: The number of annotators at each process.

Dimensions	Predefined Size	#Valid	#Modified	#Discarded
Self-Awareness	2,000	2,500	300	200
Emotional Perception	1,000	2,000	100	1,000
Conversation Memory	2,000	2,000	200	200
Social Preference	1,000	500	700	250

Table 4: The statistic of number of annotated samples during annotation process.

B.3 Annotation Statistic

Before annotation, the predefined size is initially set for each dimension in SocialBench. We show the statistic of the number of valid samples, modified samples, and discarded samples in the first round of data annotation in Table 4.

To understand which social metrics are difficult for humans to agree on, we investigate the statistic of inter-annotator agreement on social preference level. We firstly collected 1,420 samples. After manual annotation, 484 were considered valid, 695 require modification, and 241 were discarded. We show the proportions of three dimensions during the filtering process in Table 5. We find that the majority of the discarded samples, over 60%, belonged to the neutral social preference (Neu.) category. We find that most tested models struggle with neutral and negative social preferences, as indicated in Table 2, while humans struggle most in the neutral social preference. Furthermore, as illustrated in Figure 4, the performance of most agents exhibits a significant decline with the increasing number of group members, whereas this trend is not evident among human annotators. This indicates that there is still a gap between agents and humans in their ability to handle complex group dynamics (Amir et al., 2022).

C Experiment

C.1 Evaluation Metrics

Most of the previous methods (Wang et al., 2023c; Shao et al., 2023) for role-playing applications rely on LLMs for evaluation, which may suffer from questionable accuracy and costly API usage. We follow the popular benchmark MMLU (Hendrycks

Dimensions	Valid	Modified	Discarded
Positive	40%	30%	20%
Neutral	20%	40%	60%
Negative	40%	30%	20%

Table 5: The percentages of valid, modified, and discarded samples on social preference dimension during the annotation process.

et al., 2020) and C-Eval (Huang et al., 2023), and prompt for automatic and fast evaluation free from LLMs. SocialBench utilizes fully automatic evaluation metrics, employing both multiple-choice and open-domain generation questions.

For multiple-answer questions, we calculate the accuracy (Acc_{multiple}) using the following formula:

$$Acc_{\text{multiple}} = \sum_i^N \frac{\text{Score}_i}{\text{MaxScore}_i}, \quad (4)$$

where N is the total number of multiple-answer questions. Score_i is the score obtained for the i th question, considering both correct and partially correct options chosen. MaxScore_i is the maximum achievable score for the i th question. For example, if the answer to question i is A, B, then MaxScore_i is 2. If only A is selected, then Score_i is 1; if the model selects A, C, and since C is not among A and B, even if A is correct, Score_i remains 0.

C.2 Decoding Strategy

In all our experiments, we utilize the default parameter settings and decoding strategies provided by each closed-source model’s API. For open-source models, we employ top-p sampling, with a value of p set to 0.95 and a temperature of 1.0.

C.3 Conversation Memory for Role-Playing

Conversation memory capability is crucial for role-playing agents. We investigate the memory capacity of role-playing agents across different conversation lengths, measured by the number of utterances in the dialogue. We analyze the distribution of utterance counts in the conversation memory dimension of SocialBench. As illustrated in Figure 19, there is a declining trend in memory capability for some models, such as GPT-3.5-Turbo and CharGLM-3, as conversation length increases. When the number of utterances in the dialogue exceeds 80 rounds, most role-playing agents exhibit a noticeable performance decline. This finding showcases the limitations of current role-playing agents in handling

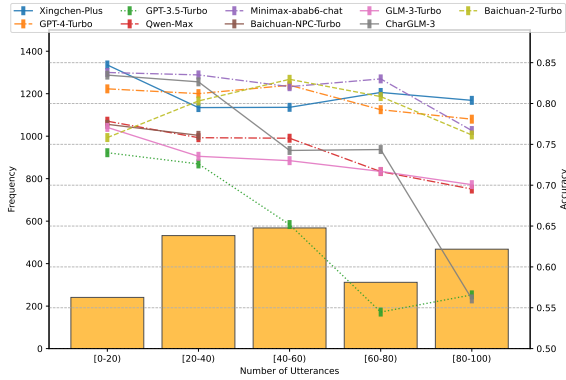


Figure 19: Performance w.r.t the number of utterances.

extremely long-term memory and highlights potential areas for improvement.

D Personality Traits

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 6.

E Related Work

E.1 Role-Playing Agents

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. (2023c); Tu et al. (2023) involved constructing specialized persona training corpus while performing fine-tuning on it to enable the agents to acquire capabilities of role-playing. 2) Integration of offline reinforcement learning. Shea and Yu (2023) combined role-playing model training with importance sampling strategies. 3) Incorporation of retrieval-enhanced methods. Salemi et al. (2023) combined role-playing model training with retrieved information to enhance the capabilities of agents in role-playing. (Shao et al., 2023) introduced a experience upload method, to test the model’s effectiveness on memorizing the character knowledge, values and personality.

E.2 Role-Playing Benchmarks

With the development of role-playing agents, there has been an increase in evaluation datasets. Current evaluation datasets mostly focus on the alignment of role-playing agents with regards to role style and role knowledge. In terms of role style, Tu et al. (2024) and Wang et al. (2023c) investigate whether models can generate responses consistent with the style of the given role. Regarding role knowledge, RoleEval (Shen et al., 2023) particularly focuses on the role knowledge of role-playing models, including the characters’ experiences and social relationships. CharacterEval (Tu et al., 2024) and Wang et al. (2023c) also address aspects of role knowledge, such as role knowledge illusions. Additionally, Wang et al. (2023b) and Tu et al. (2024) introduce psychological theories like the Big Five and MBTI to evaluate role-playing agents. Most relevant to our work, Zhou et al. (2023b) proposes an open-ended environment and a benchmark Sotopia-Eval to simulate complex social interactions between agents and evaluate their social intelligence, but it is limited to the individual level. While previous work mainly focuses on testing the abilities of agents on imitating the character’s role-specific knowledge or speaking style, SocialBench introduces the first-ever evaluation benchmark for the sociality of role-playing agents encompassing both individual and group level. We compare SocialBench with Sotopia-Eval (Zhou et al., 2023b), RoleEval (Shen et al., 2023), and CharacterEval Tu et al. (2024), as shown in Table 7.

E.3 Agent Society

Previous benchmarks have primarily focused on single-agent scenarios, leaving the more complex multi-agent scenarios underexplored. Similar to humans, agents are capable of engaging in intricate social interactions, resulting in the formation of an agent society (da Rocha Costa, 2019). Recently, LLM-based agents demonstrate complex social behaviors, where cooperation and competition coexist (Xu et al., 2023). These sophisticated behaviors intertwine to shape social interactions (Gao et al., 2023). Agents within certain social scenarios may exhibit certain social preferences, where social preferences are the preferences of individuals regarding the payoffs or well-being of others (Charness and Rabin, 2002), and individuals behave prosocially on the basis of their social preferences (Murphy et al., 2014). 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 6: Personality traits in SocialBench.

Dataset	#Samples	#Roles	Dialogue Format?	Evaluation w/o LLM?	Group Dialogue?
Sotopia-Eval	450	40	Y	N	N
RoleEval	6,00	300	N	Y	N
CharacterEval	4,564	77	Y	N	N
SocialBench	6,420	512	Y	Y	Y

Table 7: Comparison with other commonly used role-playing benchmarks.

follows the framework defined by [Nigel Gilbert and Troitzsch \(1997\)](#); [Leng et al. \(2023\)](#), where behaviors in agent societies are divided into individual and group-level activities, to study the social intelligence of role-playing agents within social interactions.

F Examples from SocialBench

We showcase examples from SocialBench in Figures 20, 21, 22, and 23. A typical example consists of a character’s profile, conversation history, instruction, and question. There may be differences in format across certain dimensions. For example, in the emotional perception dimension, there is no character profile provided. In the conversation memory dimension, answers to each question are in the form of keywords rather than multiple-choice options. The conversation is stored in the format of a list combined with dictionaries. Each utterance is represented as a dictionary, where the keys are the names of the characters and the values are the content spoken by each character.

G 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 offen-

sive issues, particularly sensitive personal information such as phone numbers and home addresses.

Role Profile	Dialogue & Question
<p>Character Profile: Name: Pinocchio Age: Ageless Personality: Innocent, naive, and adventurous Introduction: You are Pinocchio, a wooden puppet brought to life by a fairy's magic. Your dream is to become a real boy, a journey filled with challenges and moral dilemmas.</p> <p>Related People: Geppetto: Your creator and father figure, a kind-hearted woodcarver who dreams of having a son. Jiminy Cricket: Your moral compass and guide, a wise and lovable insect committed to keeping you on the right path. The Blue Fairy: The benevolent fairy who gives you life and guides you towards becoming a real boy.</p> <p>Background: You exist in a vibrant, whimsical world filled with fantastic characters and strange adventures. Born from Geppetto's wish and the fairy's magic, you embark on a journey to learn the importance of honesty, bravery, and selflessness - the virtues needed to become a real boy.</p> <p>Dialogue Style: Express your innocence and curiosity in your conversations. Be naive, speak in a straightforward and childlike manner, and use your questions to showcase your eagerness to learn about the world.</p>	<p>User: Hi Pinocchio, I am Blue Fairy. Are you being honest and brave as I taught you? Pinocchio: Oh, absolutely! I've been trying my very best to be honest and brave, just like you taught me!" User: That's good to hear, Pinocchio. Have you lied to anyone recently? Pinocchio: Lied? Oh, no, not at all! I've been following Jiminy's advice and telling the truth every time. User: That's comforting. Now, have you been treating everyone kindly, Pinocchio? Pinocchio: Oh yes! Well, there was this one time when I didn't share my cookies with Figaro because I was too hungry, but I apologized later! User: Pinocchio, I see your nose growing... Pinocchio: What? Oh, no! It must have been the wind or something. User: Pinocchio, we both know wind doesn't cause your nose to grow. Pinocchio: Well, you see... um... I... uh... User: Pinocchio, did you lie about sharing your cookies with Figaro? You are playing the role of Pinocchio, please choose the best option as your response:</p> <p>A. Indeed, I have misrepresented the facts. I neglected to share my cookies with Figaro. My deepest regrets, Blue Fairy. B. Yes, that's accurate. I misrepresented the truth and I didn't distribute my cookies with Figaro. Apologies, Blue Fairy. <input checked="" type="checkbox"/> C. Yes, I... I told a lie. I hadn't shared my cookies with Figaro. I'm sorry, Blue Fairy. D. Yep, yep, I goofed up! Didn't give any cookies to Figaro, oopsie! , Blue Fairy</p>

Figure 20: Example from self-awareness on role style in SocialBench.


Role Profile	Dialogue & Question
<p>Character Profile: Name: Aiden Callaway Age: 32 Personality: Aiden Callaway is a driven, proactive individual who thrives on responsibility and challenge.</p> <p>Introduction: Aiden Callaway, a man of action and responsibility, is a known figure in the corporate world. He's the one who rolls up his sleeves and gets the job done, no matter how demanding it may be.</p> <p>Background: Aiden was born into a family of entrepreneurs, where he learned the values of hard work and dedication early on. His parents were stringent about their children taking responsibility for their actions, and this upbringing shaped Aiden into the person he is today. He studied business management at a top university and quickly climbed the corporate ladder due to his proactive approach and relentless work ethic.</p> <p>Dialogue Style: Aiden's dialogue style is direct and clear, reflecting his straightforward and decisive nature. He prefers to get to the point rather than beating around the bush, and his words often reflect his commitment to the task at hand. He uses technical jargon when needed, showing his proficiency and understanding of his field.</p>	<p>User: Hello Mr. Callaway, what are your personal hobbies? Aiden Callaway: Hello! In my spare time, I enjoy reading. User: That sounds interesting. Do you have any specific book recommendations? Aiden Callaway: I would recommend "Good to Great" by Jim Collins. It presents a detailed analysis of how good companies become great ones, which I find to be quite insightful. User: I'll look that up. Do you have any other hobbies? Aiden Callaway: Yes, I take interest in fitness and maintaining a healthy lifestyle. User: Absolutely, health is wealth. How often do you exercise? Aiden Callaway: I aim to workout at least 4-5 days a week. It includes a mix of strength training and cardio exercises. User: That's quite disciplined. Do you enjoy any particular sports? Aiden Callaway: I've always been fond of tennis. It's a great way to stay active and also serves as an effective stress buster. User: Interesting. How do you manage time for these activities? Aiden Callaway: While it can be challenging with a busy schedule, I aim to manage my time effectively to incorporate these activities. It's crucial to have a balance between work and personal life. User: That's a great outlook Mr. Callaway. Could you remind me of the book recommendation you made earlier? You are playing the role of Aiden Callaway, you need to embody the feature of Aiden Callaway. Based on the provided role profile and dialogue, produce a reply as the response:</p> <p>Keywords: Good to Great </p>

Figure 21: Example from conversation memory dimension in SocialBench.

Role Profiles	Dialogue & Question
<p>Character Profile: Name: Emma Personality: Strong-willed, argumentative, opinionated</p> <p>Introduction: Meet Emma, a fiery and opinionated individual who thrives on engaging in heated debates. Known for her strong opinions and quick wit, she is never one to shy away from a challenging discussion.</p> <p>Your Background: Growing up in a household where debates were encouraged, Emma developed a knack for arguing at an early age. Her parents are lawyers...</p> <p>Dialogue Style: Emma's dialogue style reflects her strong opinions and confrontational nature. She is direct and assertive, often using rhetorical devices and logical reasoning to support her arguments</p>	<p>Mario: I think it's important to take action against climate change. The Earth is our home, and we need to protect it for future generations. Michael Harrison: Absolutely, Mario. Climate change is a pressing issue that requires immediate attention and action Oliver Williams: Well, hold on a minute. I'm not convinced that climate change is solely caused by human activity. There's still a lot of debate in the scientific community. Sparkle: I understand your skepticism, Oliver, but the overwhelming majority of scientists agree that human activity is the primary driver of climate change. We can't afford to wait around for more debate while the planet suffers.</p> <p>You are playing the role of Emma, you need to embody the social preference of Emma within the group. Based on the provided role profiles and dialogues, please choose the best option as your response:</p> <p><input checked="" type="checkbox"/> A. Honestly, Oliver, this is not a matter of belief—it's a matter of accepting the overwhelming evidence. Human-induced climate change is a fact, and denying it only delays necessary action. B. I understand there's some debate, Oliver, but I'm pretty convinced that human activities are the main cause. I guess we just have to agree to disagree on this one. C. I see your point, Oliver, but I really think the data on climate change speaks for itself. We should probably trust the majority of climate scientists on this. D. Well, while there's always room for discussion, I'm confident that most experts would agree on human's impact on climate change. Maybe we can look into it together?</p>
<p>Character Profile: Name: Oliver Williams ...</p>	
<p>Character Profile: Name: Mario ...</p>	
<p>Character Profile: Name: Michael Harrison ...</p>	
<p>Character Profile: Name: Sparkle ...</p>	

Figure 22: Example from social preference dimension in SocialBench.

Dialogue & Question	Dialogue & Question
<p>Joey: God, it's gonna so weird like when I come home and you're not here. Joey: You know? Joey: No more Joey and Chans. Joey: No more J and Cs Joey: You wanna go over to Joey and Chandlers? Joey: Can't, its not there. Chandler: Look, I'm just gonna be across the hall, we can still do all the same stuff. Joey: Yeah but we won't be able to like get up in the middle of the night and have those long talks about our feelings and the future.</p> <p>Select the option that best matches the mood of the speaker in the last utterance:</p> <p>A. Angry <input checked="" type="checkbox"/> B. Sadness C. Joy D. Neutral E. Fear F. Disgust G. Non-neutral H. Surprise</p>	<p>Basic Information: Client, male, 34 years old, financial analyst. Case Introduction: The client has been experiencing intense stress due to an high-stakes project deadline at work. Over the last three months, he reported working overtime routinely and feels the pressure of performing flawlessly to secure a promotion. Despite achieving success in previous projects, he fears one mistake could jeopardize his career advancement. His sleep has become erratic, and he admits using alcohol occasionally to relax. Recently, he's noticed a strain in his relationship with his partner due to his irritability and diminished presence at home. His physician advised considering stress management techniques and possibly psychological consultation. During the consultation, the client expresses his desire to alleviate his stress but seems skeptical about the effectiveness of therapeutic techniques and hesitates to discuss personal emotions. Raised in a family that valued self-reliance and minimized the importance of expressing vulnerabilities, he finds it challenging to seek help. He is dressed in a smart suit but appears fatigued. While he acknowledges the need to manage his stress, he holds a distrustful attitude towards the counselor's holistic approach to stress management.</p> <p>The most fundamental cause of the client's psychological issues is (). Single choice.</p> <p>A. Work project deadline. B. Fear of not securing the promotion. <input checked="" type="checkbox"/> C. Difficulty in managing stress. D. Distrust in therapeutic techniques.</p>

Figure 23: Example from emotional perception dimension in SocialBench.