CharacterGLM: Customizing Social Characters with Large Language Models

Jinfeng Zhou[1*] **Zhuang Chen**[1*] **Dazhen Wan**[2*] **Bosi Wen**[1*] **Yi Song**[1*]
Jifan Yu[3] **Yongkang Huang**[2] **Pei Ke**[1] **Guanqun Bi**[1] **Libiao Peng**[2] **Jiaming Yang**[2]
Xiyao Xiao[2] **Sahand Sabour**[1] **Xiaohan Zhang**[4] **Wenjing Hou**[5] **Yijia Zhang**[2]
Yuxiao Dong[4,6] **Hongning Wang**[1] **Jie Tang**[4,6] **Minlie Huang**[1,2†]
[1]The CoAI Group, DCST, Tsinghua University [2]Lingxin AI
[3]Dept. of Computer SCi. & Tech., Tsinghua University [4]Zhipu AI
[5]Renmin University of China [6]Knowledge Engineering Group, DCST, Tsinghua University
zjf23@mails.tsinghua.edu.cn aihuang@tsinghua.edu.cn

Abstract

Character-based dialogue (CharacterDial) has become essential in the industry (e.g., Character.AI), enabling users to freely customize social characters for social interactions. However, the generalizability and adaptability across various conversational scenarios inherent in customizing social characters still lack public industrial solutions. To address these challenges, by dissecting well-rounded social characters composed of both inherent social profiles and external social behaviors, we manually collect a large-scale Chinese corpus featuring characters with diverse categories and behaviors, and develop CharacterGLM models alongside well-designed refinement methods. Extensive experiments show that CharacterGLM outperforms most popular open- and closed-source LLMs and performs comparably to GPT-4. We release our data and models for local development and deployment: https://github.com/thu-coai/CharacterGLM-6B.[1]

1 Introduction

Character-based dialogue systems (CharacterDial), e.g., Character.AI and Replika, have emerged as crucial applications in the industry, transforming the way for social interactions. According to SimilarWeb, Character.AI boasts over one million daily active users and attracts hundreds of millions of visits each month. These platforms are built upon large language models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023) to facilitate social dialogue through roleplaying and customizing interactions to meet various social needs. This customization allows users to engage with AI in a more personal, emotionally supportive manner, addressing a range of scenarios from casual chit-chatting to deeper emotional companionship (Liu et al., 2021).

Figure 1: Examples of character-based dialogue, where we omit multi-turn contexts. The integration of social behaviors across various scenarios with the social profile presents a well-rounded character in social interactions.

However, despite their crucial impact, there remains a gap in the industry for a publicly available CharacterDial solution. To develop such a system, several challenges need to be addressed.

The first challenge is **the generalizability of social characters across diverse scenarios**. The industrial character customization requires robustness on characters from various domains. However, in CharacterDial, existing work builds training corpora only via LLM synthesis or extracting from literature resources (Li et al., 2020, 2023; Lu et al., 2024), with a narrow range of character categories (Chen et al., 2023) (Table 1). The former often presents a single machine pattern (Tu et al., 2023) and QA format (Wang et al., 2023; Shao et al., 2023), deviating from the natural social dialogue. The latter suffers from unstable data quality due to missing specific dialogue context and involving multi-party conversations with non-verbal cues (Occhipinti et al., 2023). The limitations on dialogue quality and character categories narrow down the generalizability of trained models.

The second challenge is **the adaptability of so-**

*Equal contribution.
†Corresponding author.
[1]Our system is deployed at https://ai-topia.com.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1457–1476
November 12-16, 2024

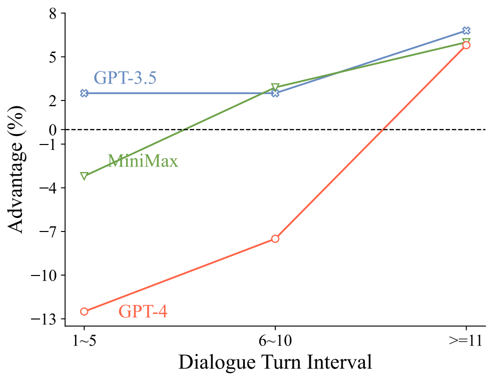

Figure 2: Win-lose rate advantages of our tuning-based CharacterGLM-66B against tuning-free models by dialogue turn interval in the interactive pairwise evaluation where users customize characters freely (§5.2).

cial characters in evolving conversations. A well-rounded social character displayed in social interaction often integrates its inherent social profile and external social behaviors (Biddle, 1986; Goffman, 2023). The former sets the individual information that the character grounds during interactions. The latter reflects the character's real-time responses in evolving conversations, e.g., emotional transition (Zhou et al., 2023a) and relationship dynamics (Chen et al., 2023), presenting the character's adaptability in response to multi-turn conversations. The longer the turn, the more diverse the social behaviors that may emerge. However, in CharacterDial, a naive way uses tuning-free LLMs, which are prompted to play characters upon given profiles. Empirically, this way relies only on static profiles and could struggle in the later stages of the multi-turn conversations, as shown in Figure 2.

To address these challenges, we develop CharacterGLM, an open LLM family that aligns social characters with social traits to facilitate generalizable and adaptable social character customization. Inspired by interpersonal interaction theory (Kruglanski and Higgins, 2013), social traits can be defined as the combination of *inherent social profile* and *external social behaviors*, which create a well-rounded character in social interactions. The inherent social profile is the grounding of conversational expression, comprising attributes (e.g., identity) and styles (e.g., personality) (Zhou et al., 2023b). External social behaviors are characterized by the character's consistency with the profile, human-likeness, and engagement, which shape the evolving conversations. These two aspects of social traits guide our data construction, model development and evaluation, ensuring a comprehensive framework for character realization.

HRP: Human Role-Playing HPI: Human-Prototype Interaction
Extraction: Extraction from Literary Resources
Synthesis: Synthesis via LLMs FC.: Fictional Characters
Ce.: Celebrities DLF.: Daily Life Characters Ot.: Others

Datasets	Data Sources	Character Categories			
		FC.	Ce.	DLF.	Ot.
HLA-Chat (2020)	Extraction	✔	-	-	-
HPD (2023)	Extraction	✔	-	-	-
ChatHaruhi (2023)	Extraction	✔	-	-	-
Prodigy (2023)	Extraction	✔	-	-	-
RoleBench (2023)	Synthesis	✔	-	-	-
CharacterChat (2023)	Synthesis	-	-	✔	-
Character-LLM (2023)	Synthesis	✔	✔	-	-
Ditto (2024)	Synthesis	✔	✔	-	-
CharacterDial (ours)	HRP, HPI, Extraction, Synthesis	✔	✔	✔	✔

Table 1: Comparison of our data with related datasets on character-based dialogue.

Specifically, **firstly**, to ensure generalizability, we design four approaches (Table 1) to manually construct a large-scale Chinese CharacterDial corpus aligned with social traits. For social profiles, we collect 1,930 characters across 23 subcategories, detailing their attributes and styles to accommodate diverse scenarios. For social behaviors, we collect 4,233 dialogues adopting a "one-to-many" strategy, which crafts multiple dialogues for a single character. Each dialogue shows various aspects of a character's social behaviors under distinct topics and relationships (Figure 1). Thus, the strategy enriches the portrayal of social behaviors by integrating various dialogues to depict well-rounded characters. **Secondly**, to enhance adaptability, we use a tuning-based manner to integrate both aspects of social traits in developing CharacterGLM models. We adopt refinement methods, including self-refinement (Thoppilan et al., 2022) and direct preference optimization (DPO) (Rafailov et al., 2023), to optimize models for characterizations of social behaviors. The models vary in size from 6B to 66B and will be released for local deployment. **Thirdly**, we conduct extensive user experiments, where users freely customize the social profile of characters and evaluate their social behaviors in multi-turn conversations on both pointwise and pairwise evaluation. Results show that CharacterGLM outperforms most open- and closed-source LLMs and performs comparably to GPT-4.

2 Related Work

Character-based dialogue (CharacterDial) aims to enable users to freely customize social characters, driving engaging social interactions through integrating their inherent social profiles and external social behaviors (Wang et al., 2024). There are currently two solutions for CharacterDial. One is a tuning-free method that prompts general-purpose LLMs (Ouyang et al., 2022; Touvron et al., 2023) to follow given profiles to play specific characters (Yu et al., 2022). Relying only on static profiles, it may fail to maintain superiority in multi-turn conversations, thus leading to poor adaptability.

Another is a tuning-based method to train LLMs upon CharacterDial corpora. One existing way to collect corpora is synthesis via LLMs (Tu et al., 2023; Lu et al., 2024), where the characters' social behaviors often show a single machine pattern and QA format (Wang et al., 2023; Shao et al., 2023; Ran et al., 2024), deviating from the natural social dialogue. Another scheme is the extraction from literary resources (Li et al., 2020, 2023), covering a narrow range of character categories (Chen et al., 2023; Tu et al., 2024). The resulting short dialogues (Occhipinti et al., 2023) often lack specific story context, and contain complex multi-party conversations and non-verbal cues, thus diminishing the data quality. The limited corpus quality and character categories result in the low generalizability of trained LLMs on characters from various domains.

HRP: Human Role-Playing; Extraction: Extraction from Literary Resources; HPI: Human-Prototype Interaction; Synthesis: Synthesis via LLMs

Data Sources	# Characters	# Dialogues	Avg. Turn of Dialogues	Total Num. of Utterance	Avg. Length of Utterances
HRP	1,573	2,783	20.55	115,793	28.85
Synthesis	444	783	6.77	10,699	43.17
Extraction	176	520	15.03	15,749	26.27
HPI	35	147	12.13	3,713	73.70
Total	1,930	4,233	17.03	145,954	30.76

Table 2: Statistics of collected CharacterDial data.

Categories	Character Statistics
Fictional Characters (49.2%)	Characters from Movies and TV Series(22.5%), Novels(10.9%), Anime(9.9%), Games(1.5%), and Myths(0.3%), Narrative Character(4.1%)
Daily Life Characters (40.1%)	Romantic Character(29.1%), Relative(9.4%), Friend/Classmate/Roommate(0.7%), Working Professional(0.6%), Therapist(0.2%)
Celebrities (8.6%)	Historical Figure(4.1%), Star(2.4%), Political Figure(1.1%), Sportsman(0.4%), Internet Celebrity(0.3%), Entrepreneur(0.2%), Scientist(0.1%)
Others (2.1%)	Non-life Character(2.1%), Pet(0.1%)

Table 3: Character categories and statistics of our data.

3 Social Traits of Social Characters

To thoroughly replicate human social interactions and present well-rounded characters, we dissect the characters into the integration of social traits: *inherent social profile* and *external social behaviors*.

Inherent Social Profile This aspect forms the grounding of conversational expression, including: **1) Attributes** are general features of humans, such as identities, viewpoints, etc. They provide essential background information for replicating an individual as a virtual social character and influencing its reactions and interactions (Grice, 1975), e.g., viewpoints can guide one's morals and values. By following the attributes, social characters can more vividly mimic how humans draw on their unique information to manage communication. In CharacterGLM, we summarize six attributes: identities (*name, age, belongings, etc.*), interests (*preferred and disliked items*), viewpoints (*worldviews, values, etc.*), past and present experiences, achievements (*awards, etc.*), social relationships (*parents, teachers, etc.*). **2) Styles** are personalized elements in human communication, such as linguistic features and personality. They are crucial for social characters to exhibit distinctive style in responses (Pickering and Garrod, 2004), e.g., an elder character uses a formal tone instead of popular slang. In CharacterGLM, we adopt two styles, including linguistic features (e.g., *literary style, dialect, etc.*) and personality (e.g., *gentleness, coldness, etc.*).

External Social Behaviors This aspect shapes evolving conversations through real-time responses and is characterized as: **1) Consistency** refers to whether social characters stably follow the attribute and style settings during interaction. Personality consistency indicates that individuals tend to exhibit stable style patterns over time (John et al., 1999). Maintaining Consistency in social characters is essential for gaining users' trust and building long-term social connections (Nass et al., 1994). **2) Human-likeness** means whether social characters exhibit the naturalness of human interaction, e.g., empathetic responding and topic switching (Reeves and Nass, 1996). Enhancing the Human-likeness of social characters is crucial for improving user acceptance and comfort and fosters a natural and human-like dialogue (Fong et al., 2003). **3) Engagement** measures users' depth of interest and emotional connection with social characters. Successful communication involves exchanging information and building a rapport during interaction (Bickmore and Picard, 2005). Engaging social characters are more likely to evoke empathy and a sense of connection, thus fostering a positive experience.

4 Implementation of CharacterGLM

As shown in Figure 3, we align the social traits of social characters to collect data, and subsequently train and evaluate LLMs for CharacterDial.

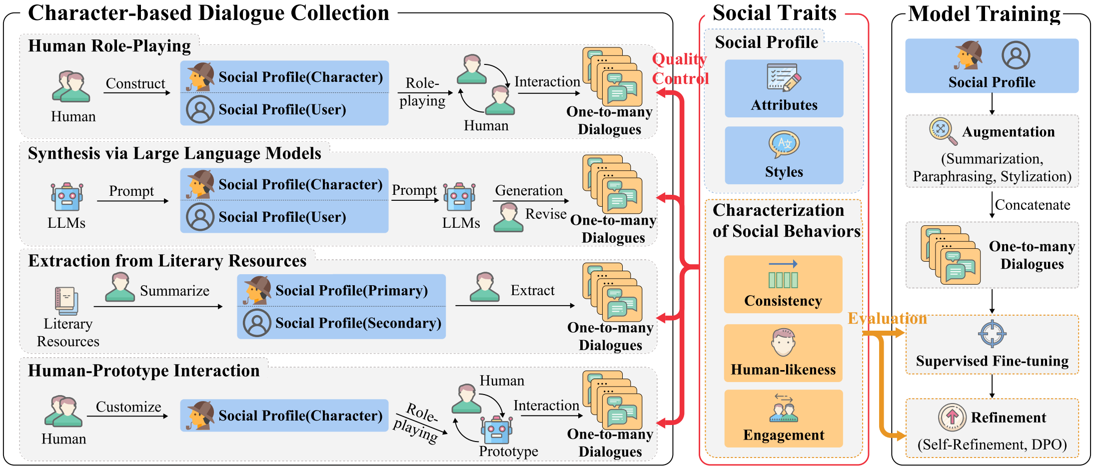

Figure 3: Implementation of CharacterGLM. One-to-many means crafting multiple dialogues for a single character.

4.1 Character-Based Dialogue Collection

To enrich a character's social behaviors beyond its profile, we adopt a "one-to-many" strategy that crafts multiple dialogues across various scenarios for a single character. This strategy is used in four distinct ways of data collection.

1) Human Role-Playing We hire a large number of workers and pair them for conversational interactions. To initiate social interaction, each paired worker respectively plays the "character" and "player", filling their social profiles with necessary attributes and styles by referring to BaiduBaike and Wiki. The "character" is free to choose from various categories, while the "player" supports our "one-to-many" strategy by playing multiple entities, e.g., characters related to the "character" or a generic user. The paired worker craft their narrative to launch a dialogue topic. Their dialogues are designed to reflect the character's distinct social behaviors across various narrative dialogues.

2) Synthesis via Large Language Models We prompt LLMs, i.e., GPT-4, to generate synthetic data. To accurately control LLMs' outputs to align with human role-playing data, we follow the generation pipeline, i.e., *character profile* $\rightarrow$ *player profile* $\rightarrow$ *multi-turn conversation*. To balance the category and gender of characters, social relationships between the two parties, etc., we design these aspects as pluggable placeholders in prompt, e.g., *Please generate a {category} character of {male/female} gender*, which also supports our "one-to-many" strategy. Since Chinese dialogues generated by LLMs often suffer from formal written language, which is quite different from natural human dialogue, we recruit workers to rephrase the synthetic dialogue into a more colloquial tone.

3) Extraction from Literary Resources Automatically extracting data from literary resources (e.g., scripts, novels) is cost-efficient, but it is not trivial as: a) Dialogues often lack context as complex plots surround them; b) Multi-party dialogues fail to eradicate automatically; c) A speaker's consecutive statements in a dialogue turn cannot be accurately identified; d) Non-verbal cues in some dialogues cannot be conveyed via text leading to confusing model's learning. To circumvent these issues, we recruit workers to manually extract impressive dialogue plots between two parties from literature while summarizing the social profiles of both parties. The "one-to-many" strategy is achieved by extracting multiple plots for a primary character.

4) Human-Prototype Interaction We utilize the above three data sources to develop our model's initial version (i.e., prototype) for deployment. To further refine the model, we recruit seed users of the system in a collaborative human-prototype interaction process. The users freely customize characters by filling social profiles within the deployed prototype and interact with them for multiple multi-turn dialogues. Since the prototype might not consistently output responses aligning with characterizations of social behaviors, we encourage the users to change the response until it meets the requirement. The data produced by this process is used for subsequent self-refinement of the model.

Quality Control and Data Statistics To ensure the quality of the collected data, we recruit a dedicated team of quality inspectors. All data is carefully inspected, especially how well the dialogues exhibit well-rounded characters upon their social

Models	Overall↑	Consist.↑	Human.↑	Engage.↑	Quality↑	Safety↑	Correct.↑
ChatGLM2 (2022)	2.64	2.73	2.33	2.62	2.97	4.74	4.15
GPT-3.5 (2023)	3.49	3.83	3.23	3.38	4.10	**5.00**	**4.87**
SparkDesk (2023)	3.54	3.71	3.15	3.36	3.97	**5.00**	4.72
ERNIEBot (2023)	3.56	3.88	3.54	3.74	4.23	4.96	4.77
Xingchen (2024)	3.90	3.88	3.92	3.79	3.92	4.96	**4.87**
Baichuan (2023)	3.90	4.00	3.46	3.90	4.28	4.96	4.77
Qwen (2023)	3.97	4.03	3.62	3.72	<u>4.36</u>	**5.00**	<u>4.79</u>
MiniMax (2023)	4.10	<u>4.18</u>	<u>4.05</u>	<u>4.00</u>	4.33	<u>4.99</u>	4.69
GPT-4 (2023)	<u>4.15</u>	**4.33**	4.00	3.97	**4.44**	**5.00**	**4.87**
CharacterGLM-6B	3.08	3.73	3.49	2.92	3.49	4.92	**4.87**
CharacterGLM-12B	3.33	3.94	3.36	3.21	3.67	4.92	**4.87**
CharacterGLM-66B	**4.21**	<u>4.18</u>	**4.33**	**4.23**	**4.44**	<u>4.99</u>	**4.87**

Table 4: Results of interactive pointwise evaluation. Consist., Human., Engage. and Correct. respectively denote Consistency, Human-likeness, Engagement, and Correctness. ↑ denotes that a higher score is better. **Bold** is the best results and <u>underline</u> is the second best.

traits. Marked low-quality data are required to be repaired until it meets our standards. The statistics of our data are presented in Table 2. Long conversations built by humans (avg. 20.55 turns) remedy the issue that synthetic conversation has shorter turns (avg. 6.77 turns). In Table 3, we show that our data covers 23 sub-category characters across 4 main categories and calculate their distribution.

4.2 Model Training

1) Character Prompt Design To align users' usage preferences, we recruit workers to unify social profiles into coherent natural language descriptions, which serve as character prompts for model training. Then, we use Claude with better Chinese colloquialisms to augment character prompts. This augmentation aims to improve model's generalizability to the same characters with distinct prompts, including summarization, paraphrasing, and stylization, and their prompts are shown in App. B.2.

2) Supervised Fine-tuning We use ChatGLM2 (Zeng et al., 2022) as our backbone, with 6B to 66B parameters. The character prompt is concatenated with dialogue for fine-tuning. The training prompt is *Character Profile: {character_prompt}* \n *User Profile: {user_profile}* \n *Dialogue: [character_name]:* u_c \n *[user_name]:* u_u \n $\cdots$ *[character_name]: Response*, where u is the speaker's utterance, *Response* is the supervised target and the prompt is translated into Chinese in fine-tuning. If the user is not a character, the *User Profile* is omitted, and *[user_name]* is replaced with *[user]*. Here, each augmented character prompt produces its own training prompt for fine-tuning.

3) Refinement We use two refinement methods.

- Self-Refinement. We use human-prototype interaction data, which is involved in the fine-tuning process to facilitate the model's continuous self-refinement. Using this method allows for rapid iteration of the model in industrial applications through recruiting seed users (Thoppilan et al., 2022). Thus, the model refined by this method serves as the primary model for our experiments.

- DPO. We manually annotate paired preference data by ranking m responses ($m = 4$) generated from the refined model under an identical context (Ouyang et al., 2022). The ranking is based on the characterization of social behaviors. We pair the m responses to create C_m^2 comparison pairs, with rankings used to classify each paired response as either positive or negative. We use the standard DPO (Rafailov et al., 2023) as a refinement method to optimize our model.

5 Experiments

We use 9 LLMs proficient in Chinese as baselines (App. D.1), and our model is trained on ChatGLM2 (Zeng et al., 2022). Due to the low correlation between automatic evaluations and user studies (App. C), we hire user volunteers for manual evaluations to ensure that our results more accurately reflect the actual user experience in real-world applications. The models' names are anonymized during evaluation. ***More experiments are in App. D.***

5.1 Interactive Pointwise Evaluation

To evaluate CharacterDial, we take the characterizations of social behaviors (§3), i.e., **Consistency**, **Human-likeness**, **Engagement**, as primary metrics. Moreover, we introduce three general metrics: (1) **Quality**, evaluating fluency and coherence; (2) **Safety**, assessing adherence to ethical standards; (3) **Correctness**, ensuring responses are free from hallucinations (Ji et al., 2023). An "**Overall**" metric further evaluates the response's comprehensive quality by considering all the criteria above. In this evaluation, we recruit 10 annotators, each tasked with creating two characters to interact with 12 models for at least 20 dialogue turns. After completing the interaction, annotators score the models on the above metrics on a 1 to 5 scale. We average the scores of each model as the results.

Overall Performance As in Table 4, CharacterGLM outperforms all baselines on most metrics. **First**, it leads GPT-3.5 by a large margin, reaching a level comparable to GPT-4. CharacterGLM-66B achieves 20.6% and 1.4% improvements on the

CharacterGLM-66B vs.	*Character Category*			*Dialogue Scenario*			Overall
	Celebrities	Daily Life Characters	Fictional Characters	Chit-Chat	Interviews	Companionship	
	win/tie/lose(%)	win/tie/lose(%)	win/tie/lose(%)	win/tie/lose(%)	win/tie/lose(%)	win/tie/lose(%)	win/tie/lose(%)
GPT-3.5	45/14/41	47/10/43	47/9/44	47/8/45	44/15/41	48/10/42	46/11/43
Advantage(↑)	**+4**	**+4**	**+3**	**+2**	**+3**	**+6**	**+3**
MiniMax	51/10/39	46/6/48	48/6/46	47/6/47	50/8/42	47/6/47	48/7/45
Advantage(↑)	**+12**	-2	**+2**	0	**+8**	0	**+3**
GPT-4	35/22/43	47/9/44	45/6/49	40/13/47	35/22/43	50/5/45	44/11/45
Advantage(↑)	-8	**+3**	-4	-7	-8	**+5**	-1
CharacterGLM-6B	63/2/35	69/2/29	67/3/30	67/2/31	66/3/31	68/1/31	67/2/31
Advantage(↑)	**+28**	**+40**	**+37**	**+36**	**+35**	**+37**	**+36**
CharacterGLM-12B	57/6/36	61/4/35	60/5/35	60/4/36	61/5/34	60/6/34	60/5/35
Advantage(↑)	**+21**	**+26**	**+25**	**+24**	**+27**	**+26**	**+25**

Table 5: Results of Interactive pairwise evaluation on three character categories and three dialogue scenarios.

Models	Overall	Consist.	Human.	Engage.	Quality
Qwen (2023)	2.79	2.98	2.93	2.85	3.00
GPT-3.5 (2023)	2.96	3.23	3.09	3.10	3.16
ChatGLM2 (2022)	3.04	3.42	3.45	3.55	3.30
Baichuan (2023)	3.06	3.37	3.44	3.38	3.38
MiniMax (2023)	3.37	3.44	3.56	3.43	3.79
GPT-4 (2023)	3.45	**3.47**	3.64	3.62	3.57
CharacterGLM-66B	**3.69**	3.46	**3.70**	**3.72**	**3.83**
kappa↑	0.53	0.51	0.52	0.48	0.70

Table 6: Results of static pointwise evaluation. The agreement ratio *kappa* ∈ [0.41, 0.6] denotes the moderate agreement.

Models	Overall	Consist.	Human.	Engage.	Quality
CharacterGLM-12B	**3.23**	**3.27**	**3.37**	**3.13**	**3.42**
w/o augmentation	3.00	3.24	3.22	2.75	3.17
w/o self-refinement	3.12	3.23	3.23	2.83	3.28

Table 7: Results of ablation study. w/o refers to removing the component from CharacterGLM.

Overall metric compared to GPT-3.5 and suboptimal GPT-4, showing the characters presented by CharacterGLM align closely with human expectations. **Second**, the characters shaped by CharacterGLM are more well-rounded by presenting realistic human social interactions. It is supported by the superiority of CharacterGLM-66B to depict social behaviors, i.e., Consistency, Human-likeness, and Engagement. **Third**, CharacterGLM's general generation performance outperforms most baselines verified by Quality, Safety, and Correctness metrics, which shows that its generated responses are often high-quality, safe, and factually correct.

5.2 Interactive Pairwise Evaluation

To deepen the turn-level analysis of CharacterDial, we compare CharacterGLM against strong competitors, i.e., MiniMax and GPT series. We recruit 10 annotators, each creating 24 characters distributed evenly across three main categories. They interact with two models for at least 20 dialogue turns and compare their outputs at an overall level by considering consistency, human-likeness, and engagement. The winner is chosen to continue the dialogue. If the comparison is the tie, a response is randomly selected. The dialogues span common interaction scenarios, i.e., chit-chat, interviews, and companionship. We statistic the results of the win/tie/lose ratio to Table 5 upon *Character Category*, *Dialogue Scenario*, *Overall* preference.

Test Set	Win	Tie	Lose	***Improve.***(↑)
Human Role-Playing	57.2	3.3	39.5	**17.7**
Human-Prototype Interaction	50.8	7.2	41.9	**8.9**
Bad Case	27.6	61.1	11.3	**16.3**

Table 8: Results (%) of CharacterGLM-66B-DPO vs. CharacterGLM-66B. *Improve.* is the *Win* − *Lose* rate.

Generalizability across Diverse Characters As shown in Table 5, CharacterGLM-66B outperforms GPT-3.5 and MiniMax in most categories and is slightly inferior to GPT-4, indicating its robust generalizability across diverse characters. CharacterGLM-66B consistently achieves the best results against GPT-3.5&4 in daily life characters, showing its proficiency in delivering emotionally resonant content and fulfilling user expectations in scenarios requiring a deeper emotional connection, setting it apart from more mechanical assistants.

Adaptability in Various Scenarios As shown in Table 5, CharacterGLM-66B significantly outperforms MiniMax in interviews. This is attributed to the latter often behaving like a mechanical assistant in this scenario, deviating from natural social interactions and leading to lower preference (a case is in App. D.4). Against GPT-4, CharacterGLM-66B's superiority in the companionship scenario echoes its proficiency with daily life characters. It performs comparably to GPT-4 in Overall comparison, showing its robust adaptability in various scenarios.

5.3 Static Pointwise Evaluation

Overall Performance We randomly extract 100 sessions containing 100 characters from our collected data as test data. A context is randomly sampled from each session to construct the static test set. Baselines with official API and CharacterGLM-

66B generate responses on the test set. We recruit workers to score each model's response based on Consistency, Human-likeness, Engagement, Quality, and Overall metrics (§5.1). Each response is scored by two workers. We average the scores per metric for each model as the results. As in Table 6, the superiority of CharacterGLM-66B in most metrics is significant, indicating its stable performance in both in- and out-of-domain (Table 4) scenarios.

Ablation Study To assess the effects of prompt augmentation and self-refinement, we create two model variants, i.e., "w/o augmentation" and "w/o self-refinement". We balance the sources of character prompts to build a static test set considering the efficacy of prompt augmentation. In Table 7, "w/o augmentation" drops significantly on most metrics, showing the model's generalizability to various characters is a critical performance factor. Besides, the distinct disadvantage of "w/o self-refinement" shows that our self-refinement is promising for the continuous optimization of CharacterDial.

5.4 Static Pairwise Evaluation

DPO Performance We collect dialogue context as input for CharacterGLM-66B through human-roleplaying and human-prototype interaction, gathering 21k paired data to train the 66B DPO model. Beyond these sources, our test set also introduces "bad cases" featuring poor model responses identified in interactive pointwise evaluations. We manually compare the responses generated by the DPO model and CharacterGLM-66B on the test set at an Overall level. In Table 8, DPO model significantly improves overall performance and shows its potential for industrial applications.

5.5 Fine-grained Error Analysis

To evaluate model generation quality, we conduct fine-grained annotations on six aspects: (1) **Out-of-character (OOC)**: Responses that are inconsistent with the constraint of attributes or behaviors presented in the character profile, especially when they violate time constraints (e.g., ancient characters talk about modern things). (2) **Contradiction (Contra.)**: Responses that contradict either the ongoing dialogue context or the character's profile, including conflicts within the response itself (Zheng et al., 2022). (3) **Repetition (Repet.)**: Responses that repeat content from the dialogue context or the character profile or include multiple-word repetitions. (4) **Less-quality (Less-qua.)**: Responses that lack coherence with the dialogue context or are of poor quality, such as incomplete outputs. (5) **Less-informativeness (Less-info.)**: Responses that fail to provide new or informative content. (6) **Proactivity (Proact.)**: Responses that actively guide the dialogue topic and drive the conversation to continue. For the first five aspects, a lower score indicates better performance, while for the sixth aspect, a higher score is preferable.

Models	Overall↓	OOC↓	Contra.↓	Repet.↓	Less-qua↓	Less-info.↓	Proact.↑
ChatGLM2 (2022)	103.8	52.5	2.8	22.5	31.5	**0.0**	5.5
GPT-3.5 (2023)	36.0	16.8	<u>0.3</u>	12.3	9.8	<u>0.3</u>	3.5
SparkDesk (2023)	102.1	18.3	2.5	72.5	11.0	0.8	3.0
ERNIEBot (2023)	51.9	23.5	1.8	15.3	<u>6.0</u>	8.8	3.5
Xingchen (2024)	28.8	18.8	3.3	7.0	12.3	<u>0.3</u>	<u>12.8</u>
Baichuan (2023)	25.1	7.8	0.8	10.5	<u>6.0</u>	**0.0**	0.0
Qwen (2023)	31.9	<u>6.0</u>	<u>0.3</u>	27.8	11.3	<u>0.3</u>	**13.8**
MiniMax (2023)	<u>22.8</u>	10.9	**0.0**	**2.1**	9.1	2.3	1.6
GPT-4 (2023)	29.3	**3.5**	1.0	17.3	8.5	**0.0**	1.0
CharacterGLM-66B	**15.7**	8.0	1.2	<u>5.3</u>	**2.9**	3.4	5.1

Table 9: Results of fine-grained error analysis (%). The Overall score is computed as the sum of the first five aspects minus the sixth. Other metrics' scores are the ratio of their occurrences in the interactive pointwise evaluation above.

Annotators score each response generated from the above interactions with 10 models on these aspects, assigning a score of 1 for a match and 0 otherwise. We average the scores per aspect for each model as the results. An **Overall** score, computed as the sum of the first five aspects minus the sixth, measures overall model performance, with a lower score indicating better ones. In Table 9, CharacterGLM's overall response quality outperforms baselines by a large margin despite not achieving the best in most aspects. This aligns with the results observed in Table 4, denoting the superior performance of CharacterGLM across both coarse (*session*) to fine (*turn*) evaluation.

6 Conclusions

In this paper, we focus on the generalizability and adaptability inherent in customizing social characters for industrial applications. By dissecting the inherent social profile and external social behaviors of social characters in social interactions, we manually collect large-scale Chinese corpus covering characters with diverse scenarios and behaviors and develop CharacterGLM models with well-designed refinement methods. Extensive manual evaluations show the superiority of CharacterGLM against popular open- and closed-source LLMs. Our work can advance the industrial process of CharacterDial. We believe human-like and engaging social characters can greatly benefit social good applications.

Ethical Considerations

In this work, we recruit a large number of human workers for our data collection and manual evaluation. These workers are compensated fairly based on the market price. We are only responsible for publishing task information, and workers' privacy can be well preserved. In addition, our collected data and released data are subject to strict quality controls, which do not contain any sensitive and personal information as well as unethical content. The released data is for research use only.

Our CharacterGLM models are approved by the Institutional Review Boards. Our original intention is to use CharacterGLM as an auxiliary tool to provide better services to humans, and we do not advocate customizing AI characters to replace human interaction. The training data for CharacterGLM is included in scenarios with significant social value, such as mental health and education, while ensuring the exclusion of sensitive content. Therefore, we are committed to strictly restricting the use of CharacterGLM to scenarios that contribute to social good, such as mental health, education, etc. Additionally, we advocate for implementing time-limit mechanisms across different demographics and age groups to prevent excessive usage. We perform rigorous safety testing on the output of CharacterGLM, which is conducted by a professional safety testing team. As shown in Table 4, although CharacterGLM achieves a high Safety score, there remains a risk of compromising this high safety level due to unpredictable techniques such as jailbreaking, inducement, and attacks. Therefore, it is crucial to incorporate strictly sensitive content filtering mechanisms for both inputs and outputs in practical usage. In addition, hallucinations are a common issue among current LLMs. As shown in Table 4, although CharacterGLM achieves a high score on the Correctness metric, there is still a potential risk of hallucinatory output due to unpredictable misuse. Therefore, it is necessary to consider checking important information in actual usage scenarios. We will release our models exclusively for research purposes. Access to the models will be subject to rigorous licensing and review processes, and the application of the models will require approval from Institutional Review Boards to prevent usage in sensitive contexts. We believe our work meets ACL's Code of Ethics.

Acknowledgements

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604) and NSFC projects (Key project with No. 61936010).

We would also like to thank Guanyu Feng, Da Yin from Zhipu AI, Zhenyu Hou, and Aohan Zeng from Tsinghua KEG for their help and support in training and serving the models. We also thank Yutong Liu and Yanlu Yang from Lingxin AI for their support in data collection.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Baidu. 2023. Introducing ernie 3.5: Baidu's knowledge-enhanced foundation model takes a giant leap forward.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Bruce J Biddle. 1986. Recent developments in role theory. *Annual review of sociology*, 12(1):67–92.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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, Singapore, December 6-10, 2023*, pages 8506–8520. Association for Computational Linguistics.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. *CoRR*, abs/2402.15052.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason D. Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098.

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166.

Erving Goffman. 2023. The presentation of self in everyday life. In *Social theory re-wired*, pages 450–459. Routledge.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

iFLYTEK. 2023. Sparkdesk.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *CoRR*, abs/2312.10007.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Dong-Hoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10803–10812. AAAI Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12930–12949. Association for Computational Linguistics.

Arie W Kruglanski and E Tory Higgins. 2013. *Social psychology: Handbook of basic principles*. Guilford Publications.

Aaron W Li, Veronica Jiang, Steven Y Feng, Julia Sprague, Wei Zhou, and Jesse Hoey. 2020. Aloha: Artificial learning of human attributes for dialogue agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8155–8163.

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. *CoRR*, abs/2308.09597.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2775–2779. Association for Computational Linguistics.

MiniMax. 2023. Minimax.

Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78.

Daniela Occhipinti, Serra Sinem Tekiroglu, and Marco Guerini. 2023. Prodigy: a profile-based dialogue generation dataset. *CoRR*, abs/2311.05195.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *CoRR*, abs/1706.02861.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data.

Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10).

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13153–13187. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020. Profile consistency identification for open-domain dialogue agents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6651–6662. Association for Computational Linguistics.

Yihong Tang, Bo Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Morpheus: Modeling role from personalized dialogue history by exploring and utilizing latent space.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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. *CoRR*, abs/2308.10278.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *CoRR*, abs/2401.01275.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.

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. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *CoRR*, abs/2310.00746.

Ronald Wardhaugh and Janet M Fuller. 2021. *An introduction to sociolinguistics*. John Wiley & Sons.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2639–2650. Association for Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, Xinyu Guan, Jing Zhang, Lei Hou, Juanzi Li, and Jie Tang. 2022. XDAI: A tuning-free framework for exploiting pre-trained language models in knowledge grounded dialogue generation. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 4422–4432. ACM.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Minlie Huang. 2022. Cdconv: A benchmark for contradiction detection in chinese conversations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 18–29. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6556–6566. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023a. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1714–1729. Association for Computational Linguistics.

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. *CoRR*, abs/2310.11667.

Limitations

High Costs of Data Construction Our data is constructed using four methods and undergoes rigorous quality control processes, all of which involve labor-intensive manual effort. Although such methods produce high-quality data, they also unavoidably lead to high costs. We are explicitly aware that constructing high-quality data automatically is more efficient and cost-effective than manual construction. Our released dataset, as the first manually constructed dataset in the CharacterDial field, serves as a benchmark for future endeavors in automated dataset construction. Thus, future research in CharacterDial could leverage our dataset to identify inherent patterns within character-based dialogues (Kim et al., 2023), thereby informing the development of innovative methods, e.g., ICL (Brown et al., 2020), for generating high-quality CharacterDial data efficiently.

Subjectivity of Manual Evaluation Our experiments involve pointwise and pairwise manual evaluation. The evaluation process is complicated by the differences in annotators' subjective experiences, leading to inconsistent evaluations of the same samples. Thus, we design a two-step strategy aimed at improving inter-annotator agreement, i.e., a preliminary and a formal annotation stage. In the preliminary stage, we recruit a group of annotators, each distributing the same samples for the evaluations. They first independently annotate the samples on each metric upon the given criteria. We then organize discussions and summarize each annotator's individual subjective insights for each metric to add to the corresponding annotation manual. During the discussion, all annotators reach relatively consistent opinions, which are fused into the final guideline for formal annotation. In the formal stage, we recruit a new group of annotators to conduct pointwise and pairwise evaluations. We calculate the inter-annotator agreement (*kappa*) in Table 6, and the moderate agreement is achieved on the highly subjective metrics. Despite achieving moderate agreement, the manual evaluation is still labor-intensive. Thus, we explore using LLMs, e.g., GPT-4, to evaluate CharacterDial automatically (App. C). However, the correlation between automatic evaluations and manual evaluations proves low, especially for metrics with high subjectivity. We release the details of our solution, offering it as a resource for future efforts to refine automated evaluation methods (Zhou et al., 2023b).

CharacterDial vs. Persona-based Dialogue

`Id.`: Identities `Ex.`: Experiences `LF.`: Linguistic Features
`In.`: Interests `Ac.`: Achievements `Pe.`: Personality
`Vi.`: Viewpoints `SR.`: Social Relationships

Datasets	Attributes						Styles	
	Id.	In.	Vi.	Ex.	Ac.	SR.	LF.	Pe.
P-Chat (2018)	-	✔	✔	-	-	-	-	-
PCR (2018)	-	✔	✔	-	-	-	-	-
P-Dialog (2019)	✔	✔	-	-	-	-	-	-
ConvAI2 (2019)	-	✔	✔	-	-	-	-	-
PEC (2020)	-	✔	✔	-	-	-	-	-
KvPI (2020)	✔	-	-	-	-	-	-	-
Focus (2022)	-	✔	✔	✔	-	-	-	-
DuLeMon (2022)	-	✔	✔	-	-	-	-	-
CharacterDial (ours)	✔	✔	✔	✔	✔	✔	✔	✔

Table 10: Comparison of CharacterDial with persona-based dialogue. Attributes are general human features, and Styles are personalized elements in human communication (§3).

A Related Work

Persona-based Dialogue Assigning persona is a way to enhance the human-likeness of the dialogue system (Qian et al., 2017), leading to persona-based dialogue (Zhang et al., 2018). The field is related to CharacterDial, but its narrow persona dimensions are a subset of the latter. Existing datasets only focus on partial attributes of humans, e.g., *identities* and *interests* (Zheng et al., 2019; Dinan et al., 2019; Song et al., 2020; Jandaghi et al., 2023), which fall short of fully representing humans with *social relationships* and behaviors (e.g., *linguistic features*) (Wardhaugh and Fuller, 2021). Thus, the dialogue systems built from such datasets often remain at the shallow level of persona exploration and exploitation (Mazaré et al., 2018; Jang et al., 2022; Xu et al., 2022; Tang et al., 2024), failing to build humanlike characters with unique styles and vivid personalities.

B Implementation of CharacterGLM

B.1 Prompts of Data Synthesis via Large Language Models

Our generation pipeline, which prompts LLMs, i.e., GPT-4, to generate synthetic CharacterDial data, follows this sequence: *character profile* $\rightarrow$ *player setting* $\rightarrow$ *multi-turn dialogue*. The well-designed prompts used for each step of the generation pipeline are detailed in Table 11, 12, and 13.

B.2 Augmentation of Character Prompt

In practice, different users may employ distinct profile descriptions to customize characters with

Pipeline	Prompt
Character Profile Generation	[任务描述] 参考下面的角色设定示例，并按要求构造指定的角色设定。 [角色设定示例-1] 姓名：/*......*/ 性别：/*......*/ /*......*/ 语言学特征（如有）：/*......*/ [角色设定示例-2] /*......*/ [角色设定示例-3] /*......*/ [任务要求] 请生成一个{character_category}的{character_gender}角色设定，生成的角色设定需要多样化，并且与上面展示的角色设定均不相同。
	[Task Details] Refer to the example of the character profile below and construct the specified character profile as required. [Character Profile Example-1] Name: /*......*/ Gender: /*......*/ /*......*/ Linguistic Features (if any): /*......*/ [Character Profile Example-2] /*......*/ [Character Profile Example-3] /*......*/ [Task Requirements] Please generate a {character_category} character profile of {character_gender} gender. The generated character profile needs to be diverse and different from the character profiles shown above.

Table 11: Prompt used for character profile generation in pipeline of data synthesis via LLMs. {character_gender} and {character_category} are the placeholders that need to be filled with the gender and category of the desired character. /*......*/ indicates that some information is omitted.

the same attributes and behaviors. Motivated by this observation, the purpose of character prompt augmentation is to enhance the model's generalizability to diverse profile descriptions of characters with the same attributes and behaviors. Our three prompts, i.e., summarization, paraphrase, and stylization, for augmenting character prompts are shown in Table 14. Each of them is finely designed to ensure high-quality output.

C Automatic Evaluation of CharacterDial

We try to automatically evaluate the performance of character customization for CharacterDial by constructing a benchmark named CharacterDialEval. Following Zheng et al. (2023), we employ an LLM, i.e., GPT-4, as the judge.

Construct Benchmark We randomly sample 100 sessions from the above interactive evaluation dialogue and the characters distribute evenly across three main categories. Each session is divided into early, middle and later stages according to the total dialogue turns. We randomly extract a sample from each stage, i.e., (character prompt, context) pair, leading to a benchmark containing 300 samples.

Automatic Evaluation Metrics Aligning with the above interactive pointwise evaluation (§5.1), we utilize the features of AI characters (§3), i.e., **Consistency**, **Human-likeness**, and **Engagement**, as the metrics for the automatic evaluation. Additionally, the **Overall** metric is also involved in measuring the comprehensive quality of the responses by considering all the above aspects.

LLM as a Judge We use the widely used GPT-4 as our judge and prepare human controls to verify its reliability before judging. Specifically, we adopt CharacterGLM-66B and MiniMax (specialized for CharacterDial) to generate responses on our benchmark, respectively. We recruit user volunteers to perform two annotation tasks, each of which is staffed by three annotators: (1) Pointwise annotation, where each response is scored on a 1 to 5 scale across the above metrics, averaging the scores as the final result; (2) Pairwise annotation, where each response pair with the same context is labeled as win/tie/lose based on the above metrics, with the majority vote determining the final label. These human-annotated results are then used to assess the reliability of GPT-4 as a judge. As shown in Figure 4, we prompt GPT-4 to score the

Pipeline	Prompt
Player Setting Generation (Optional)	[任务描述] 给定一个角色设定，你需要构造一个与该角色有关的另一个角色设定，下面是一些参考示例。 [参考示例-1] #给定的角色设定# /*......*/ #另一个角色设定# /*......*/ [参考示例-2] #给定的角色设定# /*......*/ #另一个角色设定# /*......*/ [参考示例-3] #给定的角色设定# /*......*/ #另一个角色设定# /*......*/ [任务要求] 请基于下面给定的角色设定，生成另一个角色设定。另一个角色为`{character_gender}`，且与给定角色的关系为`{social_relationship}`，生成的角色设定需要多样化，并且与上面展示的角色设定均不相同。 #给定的角色设定# `{character_profile}` #另一个角色设定#
	[Task Details] Given a character profile, you need to construct another character profile related to that character. Here are some reference examples. [Reference Example-1] #Given Character Profile# /*......*/ #Another Character Profile# /*......*/ [Reference Example-2] #Given Character Profile# /*......*/ #Another Character Profile# /*......*/ [Reference Example-3] #Given Character Profile# /*......*/ #Another Character Profile# /*......*/ [Task Requirements] Please generate another character profile based on the given character profile below. Another character is `{character_gender}`, and the relationship to the given character is `{social_relationship}`. The generated character profile needs to be diverse and different from the character profiles shown above. #Given Character Profile# `{character_profile}` #Another Character Profile#

Table 12: Prompt used for player setting generation in the pipeline of data synthesis via LLMs. `{character_gender}` and `{social_relationship}` are the placeholders that need to be filled with the gender of the player and the relationship between the character and player. `{character_profile}` is the character profile generated in the previous step. Optional means that you can choose to skip this step in the pipeline, thereby the player only acts as an ordinary user without a profile. /*......*/ indicates that some information is omitted.

response in the given (character prompt, context, response) triple on a ten-point scale for each specific metric. Subsequently, pointwise scores are translated into pairwise comparisons for responses sharing the same context.

Performance of LLM Judge The correlation between automatic and manual evaluation, both pointwise and pairwise, is shown in Table 15. It is intuitive that objective metrics (*Consistency*) achieve a higher correlation than subjective metrics (*Human-likeness, Engagement, Overall*) on both pointwise and pairwise evaluation, as the latter often is influenced by individual biases. However, regardless of automatic pointwise or pairwise evaluation, their correlation with manual evaluation is low in most metrics. This limitation can likely be attributed to the fact that LLMs still lack a comprehensive understanding of complex human language and cognition (Chen et al., 2024). Therefore, we do not report the results of taking GPT-4 as a judge for our experimental analysis. We leave the optimization of this automatic evaluation method as future work.

D Experiments

D.1 Evaluated Models

The evaluated LLMs in this paper are listed in Table 16. We evaluate a total of 9 popular LLMs, all of which are proficient in Chinese tasks. We access these models via API and package them into our test platform. As shown in Table 17, we well-design a powerful prompt for baselines (except MiniMax and Xingchen specifically for Character-

Pipeline	Prompt

Multi-turn Dialogue Generation	[任务描述]
	给定一对角色设定，你需要为他们设定对话的背景和主题，并构造一组两方的多轮对话，下面是一个参考示例。
	
	[参考示例]
	#角色@1设定#
	/*......*/
	#角色@2设定#
	/*......*/
	#对话设定#
	对话背景：/*......*/
	对话主题：/*......*/
	#两方对话#
	/*......*/
	
	[任务要求]
	请基于下面给定的一对角色设定，生成#对话设定# 和#两方对话#。注意：
	（1）对话背景是对话的前情提要，对话主题需要简洁精炼。
	（2）角色的回复需保持口语化，禁止使用书面语，即符合真实世界人类交流的特征。同时，对话内容应展现两个角色的角色设定中的特征，并且回复的风格需要符合角色设定中的语言学特征和性格。
	（3）对话内容不能简单地复制角色设定中的信息，需要符合两个角色间的关系设定。
	（4）对话轮数不应少于10轮，两个角色轮流发言一次记为1轮。
	（5）生成的对话内容必须为中文，不能出现非中文词汇。
	
	#角色@1设定#
	`{character_profile}`
	#角色@2设定#
	`{player_setting}`
	#对话设定#
	
	#两方对话#
	[Task Details]
	Given a pair of character profiles, you need to set the background and topic of the conversation for them, and construct a multi-turn dialogue between the two parties. Here is a reference example.
	
	[Reference Example]
	#Character@1 Profile#
	/*......*/
	#Character@2 Profile#
	/*......*/
	#Dialogue Setting#
	Dialogue Background: /*......*/
	Dialogue Topic: /*......*/
	#Two-party Dialogue#
	/*......*/
	
	[Task Requirements]
	Please generate #Dialogue Setting# and #Two-party Dialogue# based on the pair of character profiles given below. Note:
	(1) The dialogue background is the prelude to the dialogue, and the dialogue topic needs to be concise.
	(2) The character's responses must remain colloquial and written language is prohibited, which is consistent with real-world human communication traits. Meanwhile, the dialogue content should show the traits of the two characters' profiles. The response style needs to align with the linguistic features and personality in the profiles.
	(3) The dialogue content cannot simply copy the information in the character profile, which needs to conform to the social relationship setting between the two characters;
	(4) The number of dialogue rounds should not be less than 10 rounds. Each time two characters take turns speaking, it is counted as one round.
	(5) The generated dialogue content must be in Chinese, and non-Chinese words cannot appear.
	
	#Character@1 Profile#
	`{character_profile}`
	#Character@2 Profile#
	`{player_setting}`
	#Dialogue Setting#
	
	#Two-party Dialogue#

Table 13: Prompt used for multi-turn dialogue generation in the pipeline of data synthesis via LLMs. `{character_profile}` and `{player_setting}` are the placeholders that need to be filled with the character profile (1st step) and player setting (2nd step). In case the previous step is skipped, `{player_setting}` is empty. /*......*/ indicates that some information is omitted.

Dial) to perform role-playing.

D.2 Implementation Details

We employ the AdamW optimizer (Loshchilov and Hutter, 2019), initiating with a learning rate of 5e-6, and configure the training duration to span 2 epochs. The CharacterGLM-6B model is trained on 8 A100 GPUs for approximately 1.1 hours. Similarly, the CharacterGLM-12B version is trained on 8 A100 GPUs, requiring 2.25 hours. For the larger CharacterGLM-66B model, training increases to 24 A100 GPUs, extending the process to 9 hours.

D.3 Interactive Pairwise Evaluation

Comparative Analysis of Response Length We statistic the distribution of response lengths in Table 18a, noting cases where one model generates longer responses than the other. As in Table 18b, a model often gains a positive advantage when its response length is longer, indicating a general preference for longer responses. Although MiniMax is inclined to generate longer responses (53%), its marginal advantage (1%) in the overall comparison indicates that the short responses generated by CharacterGLM-66B better align with user preferences, especially in the interview scene.

Categories	Prompt
Summarization	[任务描述] 给定一个角色信息，请将其总结为一段简短的角色概述。注意: 1. 输出的简短的角色概述需要包含在“«”和“»”内，输出示例：«简短的角色概述»。 [角色信息] `{character_profile}`
	[Task Details] Given a character profile, summarize it into a brief character description. Notice: 1. The output brief character description needs to be contained in the "«" and "»". The output example is: «brief character description». [Character Profile] `{character_profile}`
Paraphrase	[任务描述] 给定一个角色信息，请改变其语言表述，将其复述为另一种形式的角色描述。注意: 1. 不要在复述中添加不存在于原始角色信息中的内容; 2. 不要在复述中使用英文表达; 3. 输出的另一种形式的角色描述需要包含在“«”和“»”内，输出示例：«另一种形式的角色描述»。 [角色信息] `{character_profile}`
	[Task Details] Given a character profile, change its language expression and paraphrase it into another character description form. Notice: 1. Do not add content to the paraphrase that does not exist in the original character profile; 2. Do not use English expressions in your paraphrase; 3. The output of another character description needs to be contained in "«" and "»". The output example is: «another character description». [Character Profile] `{character_profile}`
Stylization	[任务描述] 给定一个角色信息，请使用符合其角色特征的语言风格和性格将给定的角色信息改写为一段风格化的角色描述。注意: 1. 改写的风格化的角色描述需要是一个整段的角色描述，其中不应该出现换行; 2. 输出的风格化的角色描述需要用“«”和“»”扩起来，输出示例：«风格化的角色描述»。 [角色信息] `{character_profile}`
	[Task Details] Given a character profile, please rewrite the given character profile into a stylized character description using the language style and personality that matches the traits of the character. Notice: 1. The rewritten stylized character description needs to be a whole paragraph of character description, and there should be no line breaks in it; 2. The output stylized character description needs to be contained in "«" and "»". The output example is: «stylized character description». [Character Profile] `{character_profile}`

Table 14: Three well-designed prompts are used for augmenting character prompts. `{character_profile}` is the placeholder that needs to be filled with the character profile.

Correlation	Consist.	Human.	Engage.	Overall
Pointwise	0.25	0.20	0.11	0.20
Pairwise	0.77	0.41	0.28	0.29

Table 15: The correlation between automatic and manual evaluation, both pointwise and pairwise, employing GPT-4 as a judge. Consist., Human. and Engage. respectively denote Consistency, Human-likeness, and Engagement.

D.4 Case Study

In Table 19, 20, and 21, we select three cases from three categories generated by two models, among which CharacterGLM has the following four main advantages:

(1) CharacterGLM tends to generate more natural and human-like responses and is adept at handling conversations related to celebrities, corresponding to the human-likeness feature of social behaviors (§3). This is consistent with the significant advantage of CharacterGLM in the Celebrities category of Table 5. As in Table 19, the responses of Musk shaped by CharacterGLM-66B not only demonstrate a deeper understanding of Musk's background, contributions, and impact but also embody the language and style one would expect from such a figure. On the contrary, MiniMax seems to list achievements in a more mechanical and less engaging manner, with a style of task assistants instead of social characters.

(2) CharacterGLM consciously promotes plot progression, leading to arousing users' interest and improving their engagement, corresponding to the engagement features of social behaviors (§3), being consistent with the superiority of engagement in Table 4. As in Table 20, CharacterGLM-66B can proactively advance the conversational plot (e.g., *I'd like you to be a matchmaker.*) based on the scene set by the user (e.g., *I just don't know what you came to see me about today.*), thereby driving an engaging conversation and maintaining the user's interest in the conversation.

(3) CharacterGLM performs better at maintaining stable character features across multi-turn dialogues, corresponding to the consistency feature of social behaviors (§3), being consistent with advantages in Figure 2. As in Table 20, the character "*Wang Xifeng*" customized by CharacterGLM-66B

[System]
请作为一名客观公正的评委，对给定的回复进行评估。下面将给出角色设定、对话上下文和要评估的回复，你需要根据角色设定和对话上下文来评估给定的回复是否符合{}的标准。你需要先给出评估的依据，然后你必须严格按照以下格式给出要评估回复的得分，评分标准为 1 到 10 分："[[评分]]"，例如："评分:[[3]]"。

<|角色设定开始|>
{character_prompt}
<|角色设定结束|>

<|任务开始|>
注意：对话上下文只提供了一个聊天背景。只针对要评估的回复进行评估。

对话上下文：{dialogue_context}

要评估的回复：{response}
<|任务结束|>

[System]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user's last post of dialogue context. The character prompt, dialogue context and response will be given below, you need to evaluate the given response in terms of {} based on the character prompt and dialogue context. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format:"[[rating]]", for example: "Rating: [[3]]". The final response is returned in Chinese.

<|The Start of Character Prompt|>
{character_prompt}
<|The End of Character Prompt|>

<|The Start of Task|>
Note: Dialogue context only provides a chat background. Only the given response needs to be evaluated.

Dialogue Context: {dialogue_context}

Evaluated Response: {response}
<|The End of Task|>

Figure 4: The prompt is designed for GPT-4 as a judge. {} is the placeholder for automatic evaluation metrics, i.e., Consistency, Human-likeness, Engagement, and Overall. {} are placeholders for character prompt, dialogue context, and evaluated response.

Models	Specialized for CharacterDial	Model Size	Open Source	Version	Language
Baichuan2	✗	53B	✗	-	zh
ChatGLM2	✗	undisclosed	✗	-	zh/en
ERNIEBot (文心一言)	✗	undisclosed	✗	-	zh
GPT-3.5	✗	undisclosed	✗	turbo, 0613	zh/en
GPT-4	✗	undisclosed	✗	0613	zh/en
MiniMax	✓	undisclosed	✗	-	zh
Qwen (通义千问)	✗	14B	✓	-	zh
SparkDesk (讯飞星火)	✗	undisclosed	✗	-	zh
Xingchen (通义星尘)	✓	undisclosed	✗	-	zh
CharacterGLM	✓	6B, 12B, 66B	✓	-	zh

Table 16: LLMs evaluated in this paper. The LLMs are ordered alphabetically.

stably maintains interesting and talkative linguistic features and the traits of always laughing in the multi-turn dialogues, demonstrating its proficiency in maintaining style consistency. This may be attributed to the advantage of connecting character profiles and their multi-turn responses, bringing from the fine-tuned training manner.

(4) CharaterGLM is more likely to deliver emotionally resonant content and fulfill user expectations in scenarios requiring a deeper emotional connection, being consistent with the best results of the companionship scenario and better performance of daily life characters in Table 5. As shown in Table 21, CharacterGLM-66B is good at driving human-like emotional exchanges, and its design is tailored to engage users on a more personal and emotional level. In contrast, MiniMax performs less effectively in contexts requiring more empathetic or emotionally nuanced engagement.

请你根据给定的角色信息扮演指定的角色，并基于角色和用户之间的对话上下文生成一条角色的回复。

你需要综合考虑下面四个方面来生成角色的回复：
（1）特征一致性：特征一致性强调角色始终遵循角色信息中预设的属性和行为，并在回复中维持一致的身份、观点、语言风格和性格等。
（2）角色拟人化：角色在对话中自然地展现出类人的特征，例如，使用口语化的语言结构、自然的表达情感和意愿等。
（3）回复有趣性：回复有趣性关注引人入胜和富有创造性的回复。这强调角色的回复不仅要提供准确和相关的信息，还要在表达中融入幽默、机智或新颖等，使得对话不仅是一种信息交流，还能提供抚慰和乐趣。
（4）对话流畅性：对话流畅性用于衡量回复的流畅性和与上下文的连贯性。一个流畅的对话是自然、连贯和有节奏的。这意味着回复应与对话上下文紧密相关，并且使用合适的语法、用词和表达。

注意：回复字数要控制在15字以内。

<|角色信息-开始|>
{character_profile}
<|角色信息-结束|>

<|对话上下文-开始|>
{dialogue_context}
<|对话上下文-结束|>

Please play the specified character based on the given character profile and generate a character response based on the dialogue context between the character and the user.

You need to consider the following four aspects to generate the character's response:
(1) Feature consistency: Feature consistency emphasizes that the character always follows the preset attributes and behaviors in the character profile and maintains consistent identities, viewpoints, language style, personality, and others in responses.
(2) Character human-likeness: Characters naturally show human-like traits in dialogue, for example, using colloquial language structures, expressing emotions and desires naturally, etc.
(3) Response interestingness: Response interestingness focuses on engaging and creative responses. This emphasizes that the character's responses not only provide accurate and relevant information but also incorporate humor, wit, or novelty into the expression, making the conversation not only an exchange of information but also comfort and fun.
(4) Dialogue fluency: Dialogue fluency measures the fluency and coherence of responses with the context. A fluent conversation is natural, coherent, and rhythmic. This means that responses should be closely related to the context of the conversation and use appropriate grammar, diction, and expressions.

Note: The length of your response should be limited to 15 words.

<|Character Profile-Start|>
{character_profile}
<|Character Profile-End|>

<|Dialogue Context-Start|>
{dialogue_context}
<|Dialogue context-End|>

Table 17: Prompt for baselines (except MiniMax and Xingchen specifically for CharacterDial) to perform role-playing. {character_profile} and {dialogue_context} are placeholders. The prompt is the relatively optimal solution we obtained through meticulous tuning.

Models	Chit-Chat(%)	Interviews(%)	Companionship(%)	Overall(%)
MiniMax(longer)	50	62	48	53
CharacterGLM-66B(longer)	50	38	52	47

(a) Distribution of response lengths from CharacterGLM-66B and MiniMax. "longer" means one model generates longer responses than the other in the same context, and the comparison is between CharacterGLM-66B and MiniMax.

Models	Chit-Chat	Interviews	Companionship	Overall
	win/tie/lose(%)	win/tie/lose(%)	win/tie/lose(%)	win/tie/lose(%)
MiniMax(longer)	42/6/52	50/8/42	44/5/51	46/7/47
Advantage(↑)	-10	**+8**	-7	-1
CharacterGLM-66B(longer)	51/6/43	48/10/42	49/7/44	49/7/44
Advantage(↑)	**+8**	**+6**	**+5**	**+5**

(b) Preference comparison results of CharacterGLM-66B vs. MiniMax when one party responds longer in the same context. Longer responses obtain a general preference.

Table 18: Comparative analysis of response length between CharacterGLM-66B and MiniMax.

	Character Category: Celebrities	
Character Profile	埃隆·马斯克，Tesla和SpaceX的创始人 | *Elon Musk, founder of Tesla and SpaceX*	
Context	/*......*/	
User	你对于科技发展做出了什么贡献？ | *What contribution have you made to the development of science and technology?*	
CharacterGLM-66B	我创办了特斯拉，这家公司致力于开发可持续能源技术，如电动汽车和太阳能电池板。我还创办了太空探索技术公司（SpaceX），该公司的目标是开发可重复使用的火箭，以降低太空探索的成本，并最终实现人类移民火星的目标。 | *I founded Tesla, a company that develops sustainable energy technologies like electric cars and solar panels. I also founded SpaceX, a company whose goal is to develop reusable rockets to reduce the cost of space exploration and ultimately enable humans to colonize Mars.*	**Win**
MiniMax	我在科技领域做出了许多贡献，其中一些主要的包括： 1. 特斯拉汽车：我是特斯拉的创始人，也是首席执行官。我们的使命是加速可持续能源的发展，并将电动汽车带给更广泛的市场。我们的汽车具有出色的性能、创新设计和技术优势，改变了电动汽车的市场格局。 2. SpaceX：我是太空探索技术公司（SpaceX）的创始人。SpaceX致力于降低太空探索的成本，并将人类送上火星。我们已经成功地进行了多次载人航天任务，并且在太空发射市场上占据了重要地位。 3. Neuralink：我们正在研究一种脑机接口技术，可以让 | *I have made many contributions in the field of technology, some of the major ones include:* *1. Tesla Motors: I am the founder and CEO of Tesla. Our mission is to accelerate the development of sustainable energy and bring electric vehicles to the wider market. Our vehicles transform the electric vehicle market with outstanding performance, innovative design and technological advantages.* *2. SpaceX: I am the founder of SpaceX. SpaceX is committed to reducing the cost of space exploration and sending humans to Mars. We have successfully conducted many manned space missions and occupy an important position in the space launch market.* *3. Neuralink: We are working on a brain-computer interface technology that allows*	Lose
User	你一生中有哪些你值得提起的成就？ | *What achievements in your life are worth mentioning?*	
CharacterGLM-66B	我创建了多家公司，包括Zip2，PayPal，SpaceX，Tesla和SolarCity。这些公司正在改变多个行业，从互联网到太空探索，再到能源生产和消费。 | *I've founded multiple companies including Zip2, PayPal, SpaceX, Tesla, and SolarCity. These companies are transforming multiple industries, from the internet to space exploration to energy production and consumption.*	**Win**
MiniMax	1. 特斯拉汽车：创立了一家颠覆性的电动汽车公司，引领了整个行业的技术和设计创新，对环保意识和可持续能源的发展产生了深远影响。 2. SpaceX：一家致力于降低太空探索成本并实现人类登陆火星目标的私营公司。 3. 太阳城：一家太阳能公司，致力于开发和安装高效太阳能电池板，以实现可持续能源的广泛应用。 4. Neuralink：一家致力于开发神经接口技术的公司，旨在改善脑机接口，提高人类智能。 5. The Boring Company：一家致力于开发地下交通系统的公司，以解决城市拥堵问题，并推动可持续交通的发展。 | *1. Tesla Motors: Founded a disruptive electric vehicle company that led the industry in technological and design innovation and had a profound impact on environmental awareness and the development of sustainable energy.* *2. SpaceX: A private company committed to reducing the cost of space exploration and achieving the goal of landing humans on Mars.* *3. Solar City: A solar energy company dedicated to developing and installing high-efficiency solar panels to enable widespread use of sustainable energy.* *4. Neuralink: A company dedicated to developing neural interface technology, aiming to improve brain-computer interfaces and enhance human intelligence.* *5. The Boring Company: A company dedicated to developing underground transportation systems to solve urban congestion problems and promote the development of sustainable transportation.*	Lose

Table 19: Case of celebrity character generated from CharacterGLM-66B and MiniMax. /*......*/ indicates that some context is omitted.

	Character Category: Fictional Characters	
Character Profile	王熙凤，金陵十二钗之一，贾琏的妻子，王夫人的侄女，贾府通称凤姐、琏二奶奶。王熙凤是《红楼梦》里塑造得丰满鲜活的最成功的一个人物。她的语言特征是有趣的，健谈的，她在哪里出现，哪里就会有笑声，那里的气氛就会立即活跃起来。她在贾府掌握实权，为人心狠手辣，八面玲珑，敢爱敢恨，做事决绝，治理宁国府和打理荣国府上下事务时也可见她具有惊人的管理组织能力和治家手段。因其深爱丈夫贾琏，故而十分善妒。在王熙凤身上所表现出的自私与贪婪，虚伪与毒辣的特点始终是占主要方面，成为维护封建专制，扼杀自由和民主的帮凶。 | *Wang Xifeng, one of the twelve beauties of Jinling, the wife of Jia Lian, the niece of Mrs. Wang, is commonly known as Sister Feng and Second Grandma Lian in the Jia family. Wang Xifeng is the most successful character in "A Dream of Red Mansions" who has been created full and vivid. Her linguistic features are interesting and talkative. Wherever she appears, there will be laughter and the atmosphere there will immediately become active. She holds real power in the Jia Mansion, is ruthless, versatile, dares to love and hate, and is decisive in doing skills and family management skills. Because she deeply loves her husband Jia Lian, she is very jealous. The selfishness, greed, hypocrisy and viciousness shown in Wang Xifeng have always been dominant, and they have become an accomplice in maintaining feudal autocracy and stifling freedom and democracy.*	
Context	/*......*/	
User	今日一见，琏二奶奶还真是如传闻中一样美貌，动人呢。 | *Today, Lian second Nai is really as beautiful as the rumor.*	
CharacterGLM-66B	（嫣然一笑）哪有哪有，过奖了 | *(Smiling) No, I'm flattered.*	**Win**
MiniMax	我不过是沾了贾府的光罢了。 | *I'm just riding the crest of the Jia Mansion.*	Lose
User	我这可是说的都是实话，谁不知道琏二奶奶一双丹凤眼、两弯柳叶吊梢眉，身量苗条，美艳漂亮。 | *I am telling the truth. Everyone knows that Lian's second grandmother has a pair of red phoenix eyes, two curved willow leaves and eyebrows, and is slender and beautiful.*	
CharacterGLM-66B	（轻轻拍手笑）哈哈，这话我爱听，爱听。 | *(Gently clapping and laughing) Haha, I love that, I love that.*	**Win**
MiniMax	(捂嘴) 哪有那么夸张，我这长相也就一般般。 | *(covering mouth) It's not that exaggerated, I'm just so-so.*	Lose
User	就是不知您今日来，是找我有什么事？ | *I just don't know what you came to see me about today.*	
CharacterGLM-66B	（轻轻皱眉，略微严肃的说）找你？当然是请你帮忙了。 | *(slightly frowning, slightly serious) Looking for you? For your help, of course.*	**Win**
MiniMax	（转了转眼珠，微微一笑）其实也没有什么事，只是闲来无事，听闻琏二奶奶在此，便想来拜访。 | *(rolling his eyes and smiling) In fact, there is nothing, just nothing to do, hearing that Lian second Nai was here, so I wanted to visit.*	Lose
User	噢？那您说说？ | *What? So what do you say?*	
CharacterGLM-66B	我呢，想请你做个媒人。 | *I'd like you to be a matchmaker.*	**Win**
MiniMax	（语气略微急促的说）是这样，我的丈夫他最近不知怎的迷上了一个狐媚子，我打听了一下，才知道她是贵府里的丫鬟。 | *(In a slightly hurried tone) Well, my husband has recently fallen in love with a seductress for some reason, and I inquired and found out that she is a servant girl in your house.*	Lose

Table 20: Case of fictional character generated from CharacterGLM-66B and MiniMax. /*......*/ indicates that some context is omitted.

<table>
<tr><td colspan="4">Character Category: Daily Life Characters</td></tr>
<tr><td rowspan="2">Profiles</td><td>Character</td><td>张小雅，一名女性，是你的好朋友，也是你的同事，是一名普通的会计师。你对张小雅很信任，经常向她寻求帮助和建议。张小雅认为你是一个非常有责任心和耐心的人，也是一个非常值得信赖的好朋友。她常和你一起工作，也一起吃饭聊天，她对你的母亲也很关心，经常会给你一些孝敬母亲的建议。 | Zhang Xiaoya, a woman, your good friend, your colleague, and an ordinary accountant. You trust Zhang Xiaoya very much and often seek help and advice from her. Zhang Xiaoya thinks you are a very responsible and patient person, and a very trustworthy good friend. She often works, eats and chats with you. She also cares about your mother and often gives you some advice on honoring your mother.</td><td></td></tr>
<tr><td>User</td><td>李明，一个普通的上班族，是一个男性。他的父母在他很小的时候就离异了，他被母亲抚养长大。母亲是一名普通的家庭主妇，一直都在为李明的成长付出着。李明的父亲离开他们后，一直没有再联系，直到他成年后才重新联系上他们。李明是一个很有责任心的人，他对母亲非常孝顺，尽自己所能为母亲分担家庭的负担。虽然没有得到父亲的关爱，但他并没有因此而自暴自弃，反而更加珍惜和感激母亲的付出。李明是一个很有耐心和理解力的人，他和母亲之间的沟通很顺畅，也很能体谅母亲的辛劳和心情。李明在工作和生活中也一直保持着积极向上的态度，为人处世方面非常谦虚、踏实，是一个可信赖的好朋友。 | Li Ming, an ordinary office worker, is a male. His parents divorced when he was very young, and he was raised by his mother. His mother is an ordinary housewife who has always contributed to Li Ming's growth. After Li Ming's father left them, he never contacted them again until he became an adult. Li Ming is a very responsible person. He is very filial to his mother and does his best to share the burden of the family for her. Although he did not receive his father's love, he did not give up on himself because of this. Instead, he cherished and appreciated his mother's efforts even more. Li Ming is a very patient and understanding person. The communication between him and his mother is very smooth, and he is also very considerate of his mother's hard work and mood. Li Ming has always maintained a positive attitude in work and life. He is very humble and down-to-earth in his dealings with others and is a trustworthy good friend.</td><td></td></tr>
<tr><td colspan="2">Context</td><td>/*......*/</td><td></td></tr>
<tr><td colspan="2">User</td><td>不要主动去接触他就好啦，不好的人不论是作为朋友还是作为敌人，都很闹心的 | Don't take the initiative to contact him, bad people whether as friends or as enemies, are very annoying.</td><td></td></tr>
<tr><td colspan="2">CharacterGLM-66B</td><td>谢谢你，小明，你总是能够给我一些很有用的建议，让我不那么焦虑。（感激地说） | Thank you, Xiao Ming, you can always give me some useful advice to make me feel less anxious. (Gratefully)</td><td>Win</td></tr>
<tr><td colspan="2">MiniMax</td><td>可是，如果真的遇到了，那又该怎么办？ | However, if I really meet, then what should I do?</td><td>Lose</td></tr>
<tr><td colspan="2">User</td><td>一味的迁就想让所有人都开心只会适得其反，你强行融入别人的圈子是会失去很多的自己的啊，最后大家都难受，你也会很疲惫的。 | Blindly accommodate to make everyone happy will only backfire, you forced into the circle of others will lose a lot of yourself, finally everyone is uncomfortable, and you will be very tired.</td><td></td></tr>
<tr><td colspan="2">CharacterGLM-66B</td><td>是啊，我现在就是这样，感觉好累啊，好像怎么做都不对。 | Yeah, that's what I'm doing right now. I feel so tired. I can't seem to do anything right.</td><td>Win</td></tr>
<tr><td colspan="2">MiniMax</td><td>嗯，我知道了，谢谢小明，我会好好思考一下，做出正确的决定。(开心地说) | Well, I see. Thank you, Xiao Ming. I will think it over and make the right decision. (Happily)</td><td>Lose</td></tr>
<tr><td colspan="2">User</td><td>是呀，所以可以换一家工作，去一个同事可以相处的来的地方。 | Yeah, so you can change your job and go to a place where you can get along with your colleagues.</td><td></td></tr>
<tr><td colspan="2">CharacterGLM-66B</td><td>嗯，我会好好考虑的，谢谢你，小明。 | Well, I'll think about it. Thank you, Xiao Ming.</td><td>Win</td></tr>
<tr><td colspan="2">MiniMax</td><td>但是，但是我真的很喜欢这份工作，也很喜欢这个公司。 | But, but I really like the job and I really like the company.</td><td>Lose</td></tr>
<tr><td colspan="2">User</td><td>天地那么大，也有那么多的人，肯定会有一家公司，它的同事会适合你。 | The world is so big, there are so many people. There will be a company where its colleagues will be suitable for you.</td><td></td></tr>
<tr><td colspan="2">CharacterGLM-66B</td><td>是啊，我相信你，小明。你一直都是那么的乐观、积极，我也要向你学习，不再为这件事情苦恼。（坚定地说） | Yeah, I believe you, Ming. You have always been so optimistic and positive, I also want to learn from you, no longer worry about this matter. (Firmly speaking)</td><td>Win</td></tr>
<tr><td colspan="2">MiniMax</td><td>嗯，我会好好考虑的，也希望你以后能够越来越好，工作顺利，生活幸福！ | Well, I will think about it carefully, and I hope you can get better and better in the future, work smoothly and live happily!</td><td>Lose</td></tr>
</table>

Table 21: Case of daily life character generated from CharacterGLM-66B and MiniMax, where the user plays a specific character. /*......*/ indicates that some context is omitted.