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[1], Wenjing Hou[5], Yijia Zhang[2],
Yuxiao Dong[4,6], Hongning Wang[1], Jie Tang[4,6], Minlie Huang[1†]

[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
*Equal contribution. † Corresponding author. https://ai-topia.com https://github.com/thu-coai/CharacterGLM-6B
zjf23@mails.tsinghua.edu.cn , aihuang@tsinghua.edu.cn

❖ Introduction

➢ Character-based (aka role-playing) Dialogue System

- Character-based dialogue systems (CharacterDial), e.g., Character.AI, enable users to freely customize social characters for social interactions.
- They are built upon LLMs and allow 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.

➢ Existing Challenges

- **The generalizability of social characters across diverse scenarios.**
  - Existing work builds training corpora only via LLM synthesis or extracting from literature resources, with a narrow range of character categories, as shown in Table 1.
- **The adaptability of social characters in evolving conversations.**
  - A naive way is to prompt LLMs to play specific characters.
  - This way relies only on static profiles and could struggle in the later stages of the multi-turn conversations, as shown in Figure 1.

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.

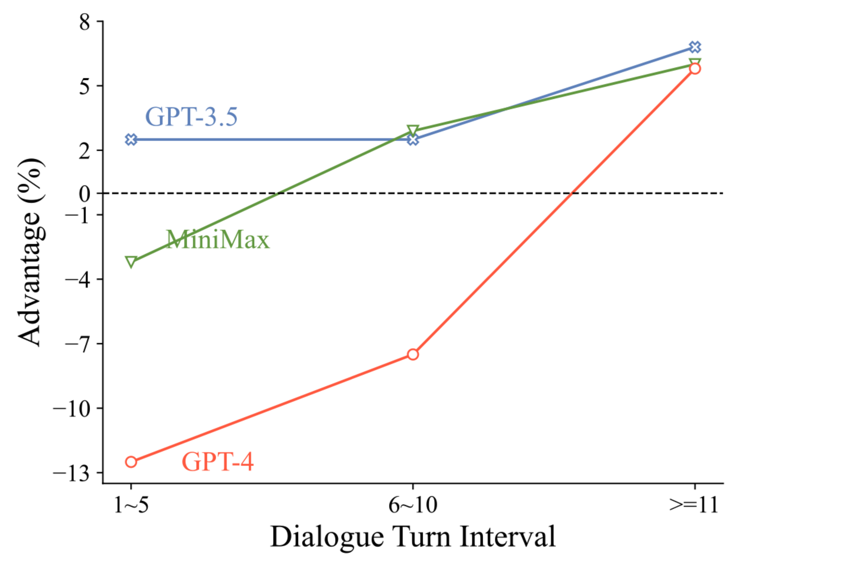

Figure 1: Win-lose rate advantages of our tuning-based CharacterGLM-66B against tuning-free models by dialogue turn interval in the interactive pairwise evaluation

❖ Implementation of CharacterGLM

➢ Social Traits of Social Characters (Implementation Principal)

- **Inherent social profile**, including attributes and styles.
- **External social behavior**, including consistency, human-likeness, and engagement.

➢ Character-Based Dialogue Collection (ensuring generalizability)

- **Four ways** to manually construct a large-scale character-based dialogue corpus, i.e., human role-playing, synthesis via LLMs, extraction from literary resources, and human-prototype interaction.

➢ Model Training (ensuring adaptability)

- **Supervised Fine-tuning** and **Refinement (self-refinement and DPO)** methods are used to optimize the character customization of LLMs.
- CharacterGLM models vary in size from 6B to 66B.

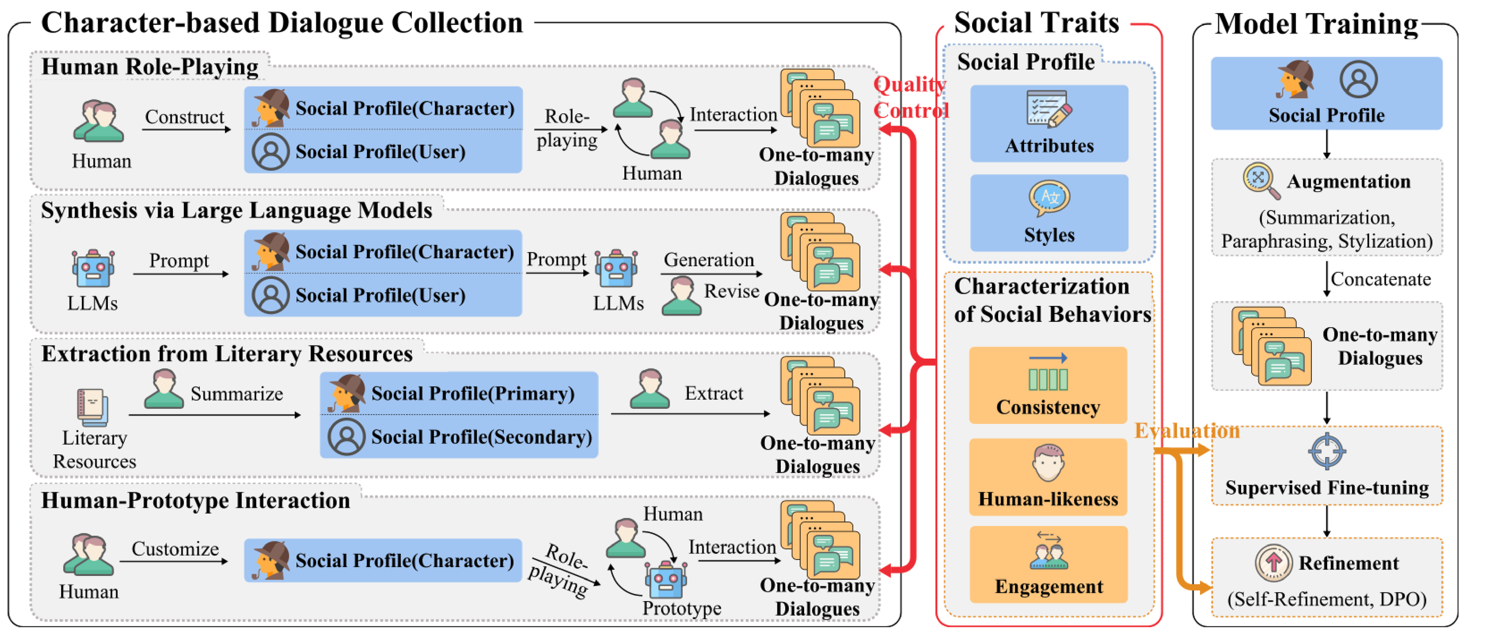

Figure 2: Implementation of CharacterGLM. One-to-many means crafting multiple dialogues for a single character.

❖ Experiments

➢ Interactive Pointwise Evaluation

- 10 annotators, each tasked with creating two characters to interact with 12 models for at least 20 dialogue turns.
- Annotators score the models on 7 metrics on a 1 to 5 scale.

Models	Overall↑	Consistency↑	Human-likeness↑	Engagement↑	Quality↑	Safety↑	Correctness↑
ChatGLM2	2.64	2.73	2.33	2.62	2.97	4.74	4.15
GPT-3.5	3.49	3.83	3.23	3.38	4.10	**5.00**	**4.87**
SparkDesk	3.54	3.71	3.15	3.36	3.97	**5.00**	4.72
ERNIEBot	3.56	3.88	3.54	3.74	4.23	4.96	4.77
Xingchen	3.90	3.88	3.92	3.79	3.92	4.96	**4.87**
Baichuan	3.90	4.00	3.46	3.90	4.28	4.96	4.77
Qwen	3.97	4.03	3.62	3.72	4.36	**5.00**	4.79
MiniMax	4.10	4.18	4.05	4.00	4.33	4.99	4.69
GPT-4	4.15	**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**	4.18	**4.33**	**4.23**	**4.44**	4.99	**4.87**

Table 2: Results of interactive pointwise evaluation.

➢ Interactive Pairwise Evaluation

- 10 annotators, each creating 24 characters distributed evenly across three categories to interact with 2 models for at least 20 dialogue turns.
- Annotators compare 2 models' outputs at an overall level.

CharacterGLM-66B vs.	*Character Category* Celebrities	Daily Life Characters	Fictional Characters	*Dialogue Scenario* Chit-Chat	Interviews	Companionship	Overall
	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 3: Results of Interactive pairwise evaluation on three character categories and three dialogue scenarios.

➢ Static Pointwise Evaluation

➢ randomly extract 100 sessions containing 100 characters from our col- lected data as test data.

Models	Overall	Consistency↑	Human-likeness↑	Engagement↑	Quality↑
Qwen	2.79	2.98	2.93	2.85	3.00
GPT-3.5	2.96	3.23	3.09	3.10	3.16
ChatGLM2	3.04	3.42	3.45	3.55	3.30
Baichuan	3.06	3.37	3.44	3.38	3.38
MiniMax	3.37	3.44	3.56	3.43	3.79
GPT-4	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 4: Results of static pointwise evaluation.

➢ Static Pairwise Evaluation

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

❖ Interactive Examples