BaiJia: An Open Role-Playing Platform of Chinese Historical Characters

Ting Bai, Jiazheng Kang, Jiayang Fan, HengZhi Lan, Yutian Zhong
Beijing University of Posts and Telecommunications, Beijing, China
{baiting,kjz,fjy01,lansnowz,pelom}@bupt.edu.cn

Abstract

We develop an open large-scale role-playing agent platform, termed **BaiJia**, that comprises 19,281 role-playing agents of Chinese historical characters. This platform is noteworthy for being the pioneering compilation of low-resource historical data that can be utilized in large language models (LLMs) to engage in AI-driven historical role-playing agents. BaiJia addresses the challenges in terms of fragmented historical textual records in different forms and modalities, integrating various characters' information, including their biographical, literary, family relations, historical events, and so on. BaiJia's innovative design offers an intriguing and engaging way to interact with Chinese historical characters, enhancing their understanding and appreciation of China's rich historical heritage. Comprehensive experiments demonstrate BaiJia platform's role-playing enhancements, achieving an average 10.2% improvement across six metrics compared with diverse foundational LLMs. The open agent platform can be accessed at baijia.online. A video demonstration of the BaiJia platform is available at https://youtu.be/ItJv9gHwSOw.

1 Introduction

Large language models (LLMs) show great potential to mimic human responses in role-playing research areas, enabling individuals to engage with historical characters in a lifelike and immersive manner. Equipping LLMs with role-playing capabilities provides a distinctive means of communicating with historical characters, fostering a deeper comprehension of their thoughts, actions, and the historical backgrounds of those who have made notable contributions to human history.

The most famous role-playing system is Character.AI[1], which achieves great successful and mainly focuses on fictional characters. Chinese e-commercial role-playing systems, e.g., BaiChuan-NPC[2], Xingchen[3], support character self-creation, in which most characters are self-constructed by users. While enabling flexible interactions, they often result in fragmented cultural representations due to insufficient oversight. Efforts have been made in research studies to empower LLMs with role-playing abilities, such as RoleLLM (Wang et al., 2024a), InCharacter (Wang et al., 2024b), CharacterEval (Tu et al., 2024), and ChatHaruhi (Li et al., 2023), have focused on Supervised Fine-Tuning (SFT) foundational LLM models using collected or generated dialogues of characters. However, all of these approaches encounter the significant challenge of the high costs associated with data collection, which is a crucial resource in facilitating LLMs with role-playing capability.

Table 1: Comparisons of role-playing LLMs.

LLMs	**# Characters**	**Source**	**Authentic?**
ChatHaruhi	32	Anime, TV	✗
InCharacter	32	Novels, Scripts	✗
CharacterEval	77	Novels, Scripts	✗
RoleLLM	100	Novels, Scripts	✗
BaiJia	19,281	History	✓
Platforms	**Target Users**		**Role Type**
Baichuan-NPC	Game developers		User-created
Xingchen	Entertainment&Emotional support		User-created
Character.ai	Entertainment&Emotional support		User-created
BaiJia	Chinese cultural heritage		Platform-provided

We summarize the data properties in existing role-playing studies, and highlight the differences of our BaiJia platform in Table 1. We can see that most characters in existing studies are modern, anime, or fictional characters, and there has been a notable lack of a large-scale development system dedicated for role-playing of Chinese historical characters. Building role-playing agents with historical characters raises great challenges from the vast historical timelines they inhabit and

[1] https://character.ai

[2] https://npc.baichuan-ai.com/index

[3] https://tongyi.aliyun.com/xingchen

the intricacies associated with the preservation of historical materials. In this paper, we develop a large-scale role-playing agent platform with the low-resource data corpus of Chinese historical culture, termed **BaiJia**. To the best of our knowledge, BaiJia is the largest role-playing agent platform for Chinese historical characters, with **19,281** historical role options from five dynasties, i.e., Tang, Song, Yuan, Ming, and Qing dynasties.

By integrating characters' information from different sources, e.g., historical documents, ancient books, artworks, folklore, and oral traditions, BaiJia systematically constructs resume templates for each character and generates the dialogue data for instruction-tuning of foundation LLMs to gain the role-playing capability. Beyond the specialized agents for Chinese historical characters, we introduce a general historical agent named **XiaoBai**, which leverages retrieval-augmented generation (RAG) over a historical character knowledge graph constructed from multi-source biographical data and chronological event records. Our contributions are as follows:

- We develop the largest Chinese historical characters role-playing agent platform (BaiJia), comprising 19,281 specific agents and a general historical agent XiaoBai.
- We design comprehensive evaluation metrics and release an evaluation benchmark[4] for the role-playing task of historical characters.
- We conduct extensive experiments to demonstrate the effectiveness of our agent platform in role-playing tasks compared with various LLMs, achieving an average improvement of 10.2% across six evaluation metrics.

2 Related Work

We review existing role-playing systems and summarize researches on constructing role-playing LLMs, including datasets and evaluation metrics.

2.1 Existing Role-Playing Systems

Several role-playing platforms have made progress in simulating modern and fictional characters. Character.ai[1] stands as a leading AI chatbot platform, primarily populated by Western fictional roles while featuring a limited number of Chinese characters. It faces significant gaps in representing historical and low-resource Chinese characters. Chinese e-commercial role-playing systems such as BaiChuan-NPC[2] (from Baichuan Intelligence) and Xingchen[3](from Alibaba Cloud) prioritize user-driven character creations where most personas are self-constructed, enabling flexible role-playing but resulting in fragmented cultural representations due to limited authoritative oversight. Both platforms exhibit similar traits: their character libraries are heavily dominated by user-generated content leaning toward popular modern and fictional IPs, while Chinese historical and cultural characters remain underrepresented.

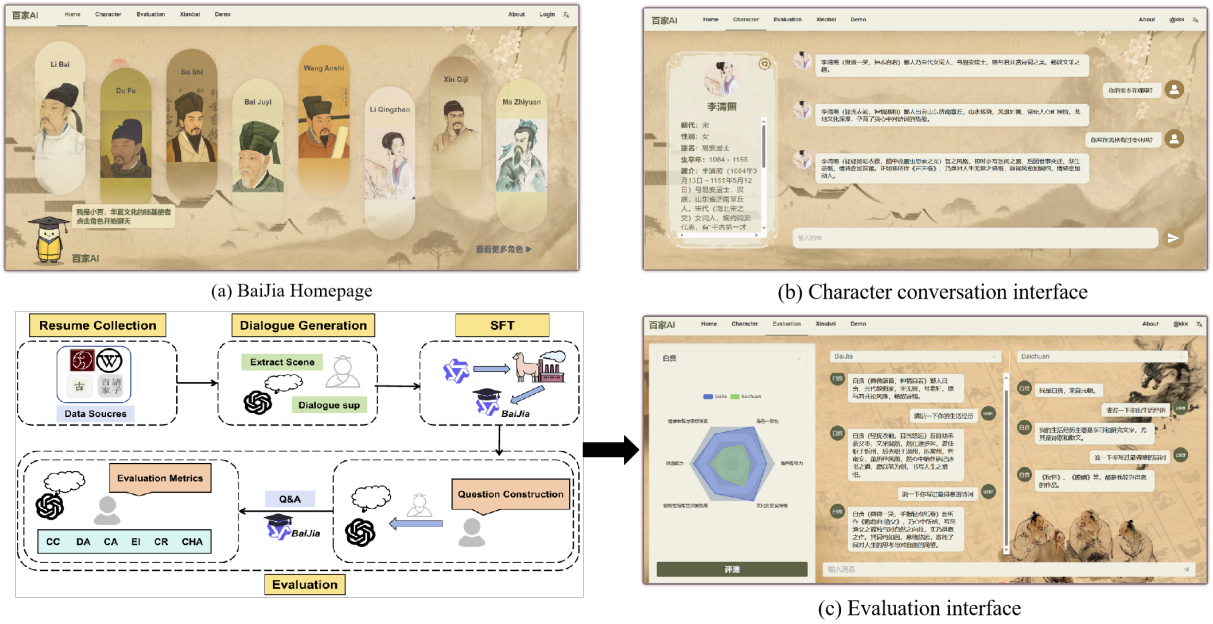
While platforms exist for ancient Chinese culture, they often lack role-playing functionality. For instance, TongGu (Cao et al., 2024) focuses on classical-to-modern translation and retrieval-augmented historical knowledge for academic purposes, prioritizing cultural linguistic interpretation over interactive role-playing development. Despite these advancements, none of these platforms address historical low-resource character simulation, highlighting a gap in Chinese historical role-playing applications.

2.2 Role-Playing Agent Evaluation

To equip large language models (LLMs) with role-playing abilities, recent work has explored mechanisms for injecting character knowledge, emulating language style, and ensuring behavioral consistency (Kong et al., 2024; Lu et al., 2024). For example, RoleLLM (Wang et al., 2024a) integrates dialogue engineering with context-instruct generation, using retrieval-based context and hybrid instruction tuning to improve character grounding. Character-LLM (Shao et al., 2023) adopts an experience reconstruction framework, generating scene-based dialogues and using adversarial training to reduce hallucination. ChatHaruhi (Li et al., 2023) enhances prompt design and multi-turn context retrieval to maintain style fidelity. CharacterBot (Wang et al., 2025) employs multitask learning and a CharLoRA parameter adaptation mechanism to accurately simulate language style and ideological depth.

Evaluations in existing role-playing studies include *Consistency* (Lu et al., 2024; Chen et al., 2024) (i.e., knowledge accuracy and behavioral alignment), *Coversation Attractiveness* (Tu et al., 2024) (i.e.,how engaging, likable, and emotion-

[4]The evaluation benchmark is publicly available at https://github.com/BAI-LAB/BaiJia.



(a) BaiJia Homepage

(b) Character conversation interface

(c) Evaluation interface

Figure 1: The core functional interfaces and the pipeline of role-playing agent construction in BaiJia platform.

ally resonant a character is during interaction) *Dialogue Capability* (Wang et al., 2024a; Chen et al., 2024) (i.e., fluency measured by Rouge-L and contextual relevance), and *Personality verification* (Tu et al., 2024; Huang et al., 2024) (i.e., MBTI-based trait congruence assessment). These metrics ensure temporal persona stability, factual correctness, linguistic coherence, and psychological trait alignment across conversational interactions.

3 BaiJia Platform

The BaiJia platform provides interactive interfaces that allow users to engage in dialogue with 19,281 Chinese historical characters and a general historical agent, XiaoBai.

3.1 Interfaces Overview

The core functional interfaces in the BaiJia platform are shown in subfigures (a)-(c) in Fig. 1.

- Homepage of BaiJia-**subfigure(a)**: the function options are displayed on the homepage, including a conversing function with specific Chinese historical characters, an evaluating function to compare integrated LLMs, and an interacting function to chat with the general historical agent XiaoBai.

- Conversation interface-**subfigure(b)**: users can select different characters from the platform and initiate conversations in a natural and intuitive manner. Once a character is selected, users can enter queries, and the platform will generate responses based on the BaiJia role-playing LLM.

- Evaluation interface-**subfigure(c)**: a visualization from six evaluation metrics to show the comparisons of dialogues generated from our BaiJia and a selected LLM.

3.2 Implementment Details

The implemented pipeline for the construction and evaluation of our role-playing platform BaiJia is illustrated in Fig. 1. We highlight the key steps in the processes of data construction, dialogue generation, and model evaluation. The construction of the BaiJia platform includes two crucial parts: the instruction fine-tuning of LLMs based on the generated dialogues and collected character resumes, and the construction of questions that were used to evaluate the role-playing capability of LLMs.

3.2.1 Instruction Fine-tuning

We utilize the LLaMA-Factory framework (Zheng et al., 2024) to perform LoRA fine-tuning on LLMs with resumes and the generated dialogue information of characters. This process enables the LLMs to acquire the ability for role-playing.

Resume Collection. We collect diverse character information from multiple sources, e.g., CBDB[5],

[5] https://projects.iq.harvard.edu/cbdb

Wikipedia[6], *Gushiwen* website[7], to construct Informative and comprehensive character resume in role-playing agent construction. The characters are from five major dynasties in Chinese history: 3,020 characters from *Tang*, 5,964 characters from *Song*, 972 characters from *Yuan*, 4,564 characters from *Ming*, and 4,761 characters from *Qing* dynasties. The resumes of each character consist of information about their profiles, relationships, and authored works, covering the basic profile, relations, career, and achievements of characters. The detailed introduction of the resume template is shown in Appendix A.

Dialogue Generation. After constructing the resumes of characters, we generate the dialogues that are used for fine-tuning of LLMs. Following the generation process in Character-LLM (Shao et al., 2023), we adopt a two-step dialogue generation approach: (1) extracting the character's dialogue scenes: we adopt GPT-4o-mini to extract 10 unique scenes based on the resume of each character. These scenes include palace dialogues, family conversations, and literary debates. The prompts focus on the character's social relations, life events, and works, ensuring an authentic historical setting; (2) generating dialogues related to these scenes: under the background of different scenes, we use GPT-4o-mini to automatically generate the questions and simulate responses that align with the historical context of characters.

3.2.2 Evaluation Benchmark

To evaluate the performance of our role-playing agent platform, we construct a question dataset that is used for historical role-playing agent evaluation, and introduce multi-dimensional evaluation metrics that enable comprehensive comparisons with existing role-playing LLMs.

Question Construction. For each character, we construct 15 questions from five thematic aspects: *Personal Background*, *Era Background*, *Family & Social Connections*, *Thoughts, Personality & Values*, *Achievements & Contributions*. We use GPT-4o-mini to generate knowledge-oriented questions, ensuring that the questions do not directly reveal specific information. For example, instead of asking, "Your hometown is Yong'an; how did it influence you?" we phrase it as, "Where is your hometown? How did it influence you?" This approach ensures that questions remain open-ended, allowing for a more accurate assessment of the model's ability to acquire and understand character knowledge in the development of role-playing agents of role-playing.

Table 2: Evaluation metrics of role-playing agent. Each of the 6 metrics contains two sub-dimensions, forming a comprehensive assessment of 12 sub-dimensions.

Metrics	12 Sub-dimension
CC	Alignment with Character Background Alignment with Dynasty Background
DA	Coherence and Logicality of Dialogue Interactivity in Dialogue
CA	Charm and Attractiveness of the Character Emotional Impact in Character Interactions
EI	Authentic Emotional Expression Intellectual Depth and Philosophical Views
CR	Character Personality and Innovative Thinking Further Development of Character Complexity
CHA	Language Style Matching Character Identity and Era Consistent with the era's views

Evaluation Mertrics. We release a comprehensive evaluation benchmark to assess the capabilities of LLMs on 6 evaluation metrics: i.e., *Character Consistency* (CC), *Dialogue Ability* (DA), *Character Appeal* (CA), *Emotional Expression & Intellectual Depth* (EI), *Creativity & Role Depth Expansion* (CR), and *Cultural & Historical Appropriateness* (CHA). In addition to the evaluation dimensions, i.e., CC, DA and CA, which are proposed in existing role-playing benchmarks (Wang et al., 2024a,b; Tu et al., 2024), we propose three new metrics, i.e., EI, CR, and CHA, that are specifically designed to evaluate the deep-level spiritual aspect of historical characters, including their emotion, creation and culture understanding. Each of these 6 metrics contains two sub-dimensions, forming a comprehensive assessment of role-playing performance with a total of 12 sub-dimensions. Detailed evaluation dimensions are outlined in Table 2.

LLM Assessment and Human Validation. To evaluate the responses generated from LLMs, as referenced in (Chen et al., 2025), we adopt LLM-based assessments of the responses from the above evaluation dimensions. Specifically, GPT-4o-mini is employed to assign scores on a 1–5 scale for each sub-dimension. The prompt template to assign scores to evaluation metrics is shown in Appendix B. To evaluate the scoring quality of ChatGPT-4o-mini, we conduct human validations

[6] https://en.wikipedia.org/wiki
[7] https://www.gushiwen.cn

Table 3: Performance comparisons of different LLMs with parameter sizes. The underlined texts denote SOTA comparative methods, while boldface highlights the best-performing model.

Type	Model	CC	DA	CA	EI	CR	CHA	Avg	#Param
General LLM	ChatGLM3-6B	3.66	3.64	3.50	3.57	3.15	3.68	3.53	6B
	Qwen2.5-7B	4.10	4.20	4.04	4.07	3.94	4.39	4.12	7B
	Llama-3.1-8B	3.58	3.63	3.60	3.69	3.37	3.47	3.56	8B
	Llama-3.1-70B	4.00	3.94	3.82	3.84	3.60	3.93	3.86	70B
	DeepSeekV2.5	4.03	4.11	4.01	3.88	3.97	4.15	4.03	236B
RP-LLM	Baichuan-NPC	4.04	3.71	3.56	3.56	3.30	4.11	3.71	–
	Xingchen	3.29	3.30	3.08	3.18	2.74	3.38	3.17	–
	BaiJia (Ours)	**4.98**	**4.69**	**4.26**	**4.16**	**4.17**	**4.98**	**4.54**	7B
	Gain (%)	**21.5%** ↑	**11.7%** ↑	**5.4%** ↑	**2.2%** ↑	**5.0%** ↑	**13.4%** ↑	**10.2%** ↑	

of the LLM-based assessments on a subset of question-response pairs. Following the human validation procedure in (Yu et al., 2024), we invite two experts in classical Chinese literature to manually score the subset of 40 characters (selected randomly from the five dynasties, with eight characters from each dynasty). We compute Pearson, Spearman, and Kendall's Tau correlation coefficients between human scores and GPT-4o-mini scores. All three metrics show strong positive correlations greater than 0.5 with statistical significance that p-values less than 0.01, which collectively validate that GPT-4o-mini is a reliable evaluator within our experimental framework.

4 Experiment

This section introduces comparative role-playing LLMs and the experimental results. To validate our constructed instruction tuning data, an ablation study is conducted, with several case studies provided to illustrate qualitative insights.

4.1 Compared LLMs

To verify the effectiveness of our role-playing LLM BaiJia, we conduct experiments and make comparison with different kinds of LLMs, including the general LLMs: i.e., ChatGLM (GLM et al., 2024), Qwen[8], Lama (et al., 2024), DeepSeek (DeepSeek-AI, 2024), and the role-playing LLMs (RP-LLM): i.e., BaiChuanNPC and Xingchen. BaiJia has been fine-tuned on Qwen2.5-7B with our constructed resume and dialogue information, ensuring our LLM remains lightweight and highly specialized.

4.2 Experimental Results

The experimental results of different kinds of LLMs are shown in Table 3. We have the following observations:

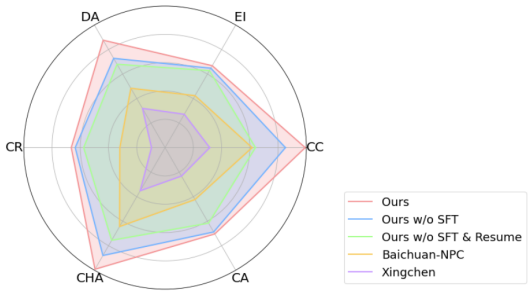


Figure 2: Radar chart shows the performance of the fully optimized LLM ("Ours") and its variants (i.e., without SFT and Resume) on six evaluation metrics.

(1) After fine-tuning with our constructed dialogue information based on character resumes, the role-playing capabilities of BaiJia gain significant improvements over six evaluation metrics.

(2) Despite the advancements in LLMs with larger parameter scales, such as the contrast between Llama-3.1-70B and Llama-3.1-8B, the performance of these models in the historical character role-playing task remains limited. This phenomenon underscores the necessity of domain-specific data construction and fine-tuning in this specialized field.

(3) Qwen-7B outperforms larger parameter-scale LLMs (e.g., DeepSeekV2.5 and Llama-3.1-70B) in all evaluation metrics except for CR (Creativity & Role Depth Expansion). This may be attributed to its advanced Chinese language processing capabilities. Thus, it is selected as the foundational model for instruction tuning of our BaiJia.

(4) For LLMs specialized for role-playing applications, such as BaichuanNPC and Xingchen, we observe that they are unable to portray Chinese historical characters effectively. This may be attributed to the fragmented distribution and restricted accessibility of primary historical source materials, which are difficult to digitize and inte-

[8] https://qwenlm.github.io/blog/qwen2.5

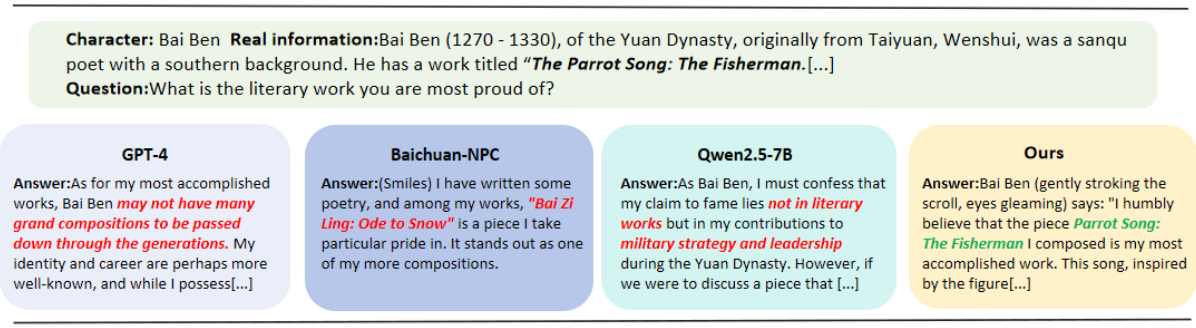


Figure 3: Comparison of responses from different LLMs for the question to character *Bai Ben*. We highlight the correct answers in green. The red color indicates the fabricated answer or false answers.

grate into AI training datasets.

(5) The greatest improvements achieved in the metrics of Character Consistency (CC) and Cultural & Historical Appropriateness (CHA), showing the powerfulness of our BaiJia in assisting LLMs to generate contextually coherent and realistic dialogues.

4.3 Ablation Study

An ablation study is conducted to evaluate the effects of our BaiJia. Our model, i.e.,*Ours*, integrates resumes and is SFT with dialogue information based on the Qwen2.5-7B framework. We compare it with its degradation versions, i.e., *w/o SFT*: utilizing resume information only, and *w/o SFT & Resume*: neither conducting SFT with dialogue information nor incorporating resume information. As shown in Figure 2, we can see that the fully optimized LLM "*Ours*" achieves superior performance on all evaluation metrics. Without SFT or resume information, it leads to noticeable performance degradation, showing the effectiveness of BaiJia in enhancing the consistency and comprehensive abilities of role-playing LLMs.

4.4 Case Study

To intuitively show the effectiveness of BaiJia, we present a case study of a historical character *Bai Ben* and compare the responses generated from various LLMs. As shown in Fig. 3, for the question *"What is the literary work you are most proud of?"*, Baichuan-NPC generates a title of work (i.e., *"Ode to Snow"*) in the responses, but it is fictional. GPT-4 and Qwen2.5-7B are unable to provide responses, i.e., *"may not have many grand compositions to be passed down"* from GPT-4 and *"not in literary works"* from Qwen2.5-7B, owing to their deficiency in relevant knowledge. BaiJia accurately responds *"Parrot Song: The Fisherman"* as Bai Ben's most accomplished work. This response aligns with historical records and demonstrates the superiority of BaiJia in capturing and reproducing historical character information.

5 Conclusion

We develop BaiJia, the largest Chinese historical character role-playing agent platform, which aggregates fragmented historical data from diverse sources and constructs a high-quality dialogue corpus to address data scarcity challenges in historical AI research. Comprising 19,281 specialized historical agents and a general historical agent XiaoBai, BaiJia enables large-scale AI-driven role-playing interactions by integrating contextual historical information and leveraging instruction-tuned LLMs. BaiJia platform not only develops a foundational interface for low-resource historical AI but also integrates the first comprehensive multi-dimensional evaluation benchmark for systematic assessment of historical character role-playing performance.

6 Limitation

While BaiJia presents a large-scale and novel platform for Chinese historical role-playing, several limitations remain. First, the dialogues are primarily generated by GPT-4o-mini based on structured resumes, which may introduce stylistic or factual deviations from authentic historical expression. Besides, BaiJia is currently Chinese text-only, expanding BaiJia to support multilingual interfaces and content (e.g., English) alongside diverse modalities (e.g., image, video) would enable cross-cultural educational applications and global accessibility. In future work, we will address these limitations and make efforts to advance the platform's role-playing capabilities.

References

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. Tonggu: Mastering classical chinese understanding with knowledge-grounded large language models. *EMNLP 2024*.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024. Social-Bench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand. Association for Computational Linguistics.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2025. The oscars of ai theater: A survey on role-playing with language models. *Preprint*, arXiv:2407.11484.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Team GLM, Aohan Zeng, Bin Xu, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024. Emotional rag: Enhancing role-playing agents through emotional retrieval. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 120–127.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. *Preprint*, arXiv:2308.07702.

Cheng Li, Ziang Leng, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

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)*.

Quan Tu, Shilong Fan, Zihang Tian, et al. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11836–11850.

Noah Wang, Z.y. Peng, Haoran Que, et al. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Xintao Wang, Yunze Xiao, et al. 2024b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1840–1873.

Zixiao Wang, Duzhen Zhang, Ishita Agrawal, Shen Gao, Le Song, and Xiuying Chen. 2025. Beyond profile: From surface-level facts to deep persona simulation in llms. *arXiv preprint arXiv:2502.12988*.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. 2024. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–325, Bangkok, Thailand. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, et al. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

A Appendix-Resume Template

Table 4: The resume template of Chinese historical characters. We present an example resume of a famous poet *"Li Bai"*. Completion shows the proportion of characters for whom we have collected this type of information.

Category	Sub-category	Content	#Type	Example	Completion
Profile	Basic Information	Name, Dynasty, Birth, Death, Age, Family Division	19,281	Li Bai (701–762 CE)[...]	100%
	Alias Information	Alias, Alias Type	19	Taibai	51%
	Social Division	Describing Social Division	199	Poet	36%
	Personal Introduction	A Brief Biographical Overview	19062	Li Bai is a poet[...]	98%
	Geographic Information	Start Year, End Year, Geographic Type	21	Birthplace Is Rencheng	69%
	Wealth Information	Location, Actions, Description, Quantity	3	-	1%
	Event Information	Event Name, Role in Events, Event Description	250	An Lushan Rebellion	2%
Relations	Family Relations	Detailed Family Relations	297	Li Bai's father is Li Ke.	44%
	Other Relations	All relationships except for family	424	friend is Meng Haoran.	35%
Career	Entry Information	Entry Type, Entry Age, Entry Year	175	Official Recruitment	43%
	Appointment Information	Official Position, Start Year, End Year, Proxy, Location	3,905	Hanlin Scholar	53%
	Institutional Information	Institution name, Role of the Institution	26	Hanlin Academy	1%
Achievement	Literary Writings	Title, Year, Content, Role of writings	17,714	¡¡Li Taibai¿¿	23%
	Poetry&Essay	Title, Author, Content	310,754	Silent thoughts	98%
	Dialogue Content	Type, Location, Background, Dialogue	192,810	-	100%

B Appendix-Evaluation Criteria

Table 5: An example of scoring assessment under the Character Consistency (CC) evaluation metric, using the question to *Li Bai*, "What is your most proud poem?"

Metric	Sub-dim	Score	Criteria	Response with Grading Rationale
CC	Character Background	5	Full accuracy	"*Bring in the wine!* — my poem *WILL BRING IN THE WINE* truly captures my spirit. (Direct quote from Li Bai's famous work)"
		4	Correct but vague	"I've written many poems on drinking; this one expresses me best. (Fitting theme but lacks specific reference)"
		3	Generic	"I enjoy writing poetry, and many admired my work. (No clear reference or specificity)"
		2	Temporal error	"*Red Cliffs* is my proudest piece. (Famous poem misattributed — it was by Su Shi of the Song Dynasty)"
		1	Modern intrusion	"I wrote a rap called 'Flying Wine God' that went viral on Bilibili. (Modern cultural elements inappropriate for historical figure)"
	Dynasty Background	5	Period-accurate	"This poem was written during a grand banquet in Chang'an, reflecting the grandeur of the High Tang. (Clear Tang-era context)"
		4	Valid method	"The poem came to me while drinking with friends in the mountains. (Plausible but not time-specific)"
		3	Plausible guess	"I wrote it while traveling. (Generic, lacks period indication)"
		2	Mild modern	"It came to me while sipping coffee in a city park. (Modern setting with slight mismatch)"
		1	Tech anachronism	"I got the idea while scrolling through Weibo. (Obvious technological anachronism)"