RoleCraft-GLM: Advancing Personalized Role-Playing in Large Language Models

Meiling Tao[1]*, **Xuechen Liang**[2]*, **Tianyu Shi**[3]†, **Lei Yu**[3], **Yiting Xie**[4]
[1]Guangdong University of Technology, Guangzhou, China
[2]East China Jiaotong University, Nanchang, China
[3] University of Toronto, Toronto, Canada
[4]Genfun.ai, Beijing, China

Abstract

This study presents RoleCraft-GLM, an innovative framework aimed at enhancing personalized role-playing with Large Language Models (LLMs). RoleCraft-GLM addresses the key issue of lacking personalized interactions in conversational AI, and offers a solution with detailed and emotionally nuanced character portrayals. We contribute a unique conversational dataset that shifts from conventional celebrity-centric characters to diverse, non-celebrity personas, thus enhancing the realism and complexity of language modeling interactions. Additionally, our approach includes meticulous character development, ensuring dialogues are both realistic and emotionally resonant. The effectiveness of RoleCraft-GLM is validated through various case studies, highlighting its versatility and skill in different scenarios. Our framework excels in generating dialogues that accurately reflect characters' personality traits and emotions, thereby boosting user engagement. In conclusion, RoleCraft-GLM marks a significant leap in personalized AI interactions, and paves the way for more authentic and immersive AI-assisted role-playing experiences by enabling more nuanced and emotionally rich dialogues[1].

1 Introduction

Large Language Models (LLMs) have become a cornerstone in understanding and generating natural language, showcasing remarkable capabilities that often surpass human proficiency in certain language reasoning tasks. Consequently, they have excelled notably excelled as conversational agents, offering high-level responses in various dialogues and significantly influencing human-AI interactions, thus hinting at their potential to reshape numerous aspects of daily life (Bender and Koller, 2020).

Despite these advancements in generative AI (Baidoo-Anu and Ansah, 2023),challenges persist, particularly in meeting the diverse requirements of different user groups.Presently, many AI systems rely on generic models, which may not adequately address the specific needs of varied users. This limitation can negatively influence user experiences and the applicability of AI in certain scenarios (Ackerman et al., 2022). Thus, there is a burgeoning recognition of the importance of personalization within AI, gradually influencing the direction of the field. There is a growing trend toward developing AI systems that are more attuned to individual preferences and needs (Brandtzæg and Følstad, 2018; Lee and Koubek, 2010), emphasizing the need for a deeper understanding of user behaviors and a more tailored approach to AI interactions. This evolving trend suggests a move toward more personalized, user-centric AI models, potentially transforming the generative AI sector to offer more individualized and effective solutions.

Existing open-source LLMs, while trained across broad domains, may not always offer the specialized optimization desired for nuanced role-playing tasks. This suggests that additional customization might be beneficial to more effectively meet the specific requirements of role-playing scenarios. On the other hand, advanced LLMs such as GPT-4 (OpenAI, 2023) showcase enhanced abilities in role-playing due to their extensive training and sophisticated algorithms. However, this closed-source model may introduce practical limitations. These include higher costs associated with API usage, limited scope for fine-tuning to tailor the models for specific role-playing contexts, and restrictions in context window sizes which can impact the fluidity and depth of generated dialogues in complex role-playing scenarios.

In this paper, as illustrated in the figure 1, we

* Equal contribution. This work is done during internship at Genfun.ai.

†Corresponding author: ty.shi@mail.utoronto.ca

[1]Access models, demos at https://github.com/tml2002/RoleCraft

Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), pages 1–9
March 22, 2024

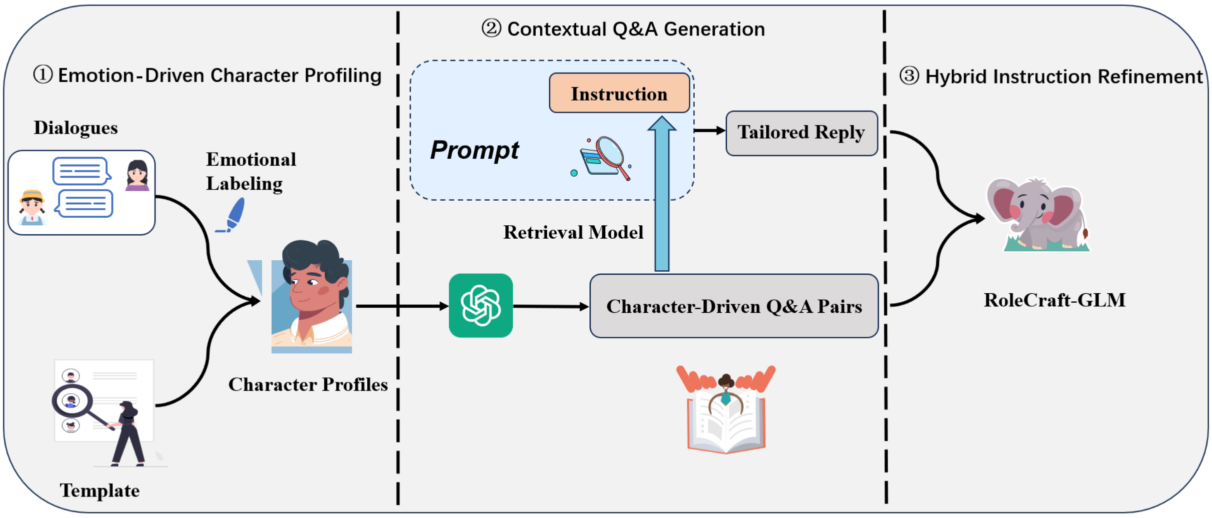

Figure 1: Overview of the RoleCraft-GLM framework: (1)Emotionally annotated dialog datasets play a key role in creating role profiles that reflect specific emotional traits. (2)The generation of Q&A pairs, based on context and known character traits, ensures that dialogues are consistent with the character profiles. (3) A hybrid approach of generic and character-specific instructions is used to train the GLM for various dialog scenarios.

introduce the RoleCraft-GLM framework designed to enhance personalized role-playing experiences with LLMs. Moving beyond traditional celebrity-focused characters, we focus on diverse, non-celebrity personas, each with unique emotional annotations. This approach aims to enrich realism and emotional depth in language interactions. We compiled a novel dataset encompassing a wide array of real-world dialogues, with careful consideration for personal privacy and copyright laws. Our data analysis highlights the potential benefits of integrating emotional labels in dialogue datasets for improved natural language processing. We conducted comparative experiments using models like ChatGLM3, fine-tuned with the Low-Rank Adaptation (LoRA) method, to assess RoleCraft-GLM's effectiveness in producing nuanced and character-consistent dialogues.

The main contributions of our work are as follows:

- We introduce a groundbreaking dataset that centers on non-celebrity characters, each characterized by unique emotional annotations. This innovative shift away from the traditional focus on celebrity-centric characters greatly enhances the realism and intricacy of everyday interactions within language modeling.

- We have developed a novel training strategy that considers more detailed personal role data. This strategy includes a fine-grained approach to character portrayal, emphasizing emotional depth, and fostering contextual awareness in dialogue generation.

- Our framework has shown promising performance, as evidenced by multi-dimensional evaluations, in comparison to current state-of-the-art models. These evaluations rigorously assess aspects such as dialogue authenticity, emotional accuracy, and contextual relevance, highlighting our framework's advanced capabilities in these critical areas.

2 Related Work

2.1 Personalization of LLMs

The recent strides in LLMs, particularly in understanding user context and preferences, have significantly propelled the personalization aspect of AI interactions.Consequently, technologies like effective prompt design and feedback mechanisms, including similarity-based retrievers, enhance AI's ability to learn from past interactions and adapt to user needs more precisely. These methods enable AI systems to proactively identify and rectify errors, thus enhancing their performance over time (Dalvi et al., 2022; Madaan et al., 2022).Furthermore,this evolution in LLMs lays the groundwork for AI applications that are not only more personalized and intuitive but also deeply aligned with user-centric principles.The evolution of personalization in AI interactions highlights a crucial shift towards understanding and catering to individual user preferences. Projects like Snap's MyAI[2] and Meta's Abraham Lincoln bot[3] serve as prime examples, demonstrating how AI can be customized for enhanced user engagement and specific functional needs., illustrating AI's potential for enhanced user engagement and functionality through customization (Bender et al., 2021). These initiatives emphasize the importance of tailoring AI systems to meet specific user needs. Against this backdrop, our work in RoleCraft-GLM aims to build upon these advancements to offer nuanced, emotionally rich AI interactions that align with individual user contexts and needs.

2.2 Role-Playing

The evolution of role-playing in AI, marked by the transition from basic text-based interactions to intricate character simulations (Park et al., 2023), reflects the strides made in natural language processing and AI technologies. Initially, AI role-playing systems offered only fundamental exchanges, limited in their ability to craft dialogues with emotional depth and contextual relevance. With the emergence of advanced models such as GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), and ChatGLM, there was a notable enhancement in AI's capability for engaging in more sophisticated, context-aware conversations. Yet, these improvements also underscored a significant gap in personalization for role-playing applications. Predominantly, LLMs trained on wide-ranging, generic datasets fell short in handling scenarios that demanded a deeper understanding of nuanced emotional nuances (Radford et al., 2018) and specific character traits. To address these shortcomings, we have meticulously developed the RoleCraft-GLM framework. It stands out with its unique dataset, focusing on diverse, non-celebrity personas enriched with detailed emotional annotations (Bender and Koller, 2020). This dataset is key to overcoming previous limitations, facilitating a new level of personalization and emotional intricacy in AI role-playing interactions.

3 Methodology

As shown in Figure 1, the RoleCraft-GLM framework, rooted in 'Role' and 'Craft', represents our approach to enhancing AI role-playing. 'Role' emphasizes creating distinct, multi-dimensional characters, each with unique personality traits and emotional depths. 'Craft' involves the intricate process of constructing dialogues that genuinely reflect these character traits, thereby adding depth and realism to conversations.

Building on this foundation, the RoleCraft-GLM framework is underpinned by key principles to elevate the authenticity of role-playing (Wang et al., 2023). The first principle, 'Fine-Grained Character Portrayal', is pivotal in endowing each character with detailed and nuanced traits and backgrounds, integral to the 'Role' aspect of RoleCraft. This approach is focused on creating characters that are reflective of real-life individuals in their personality, and behaviors (Kim et al., 2023), setting the stage for realistic and compelling character portrayals. Progressing to the second principle, 'Mastery of Emotion and Style', we concentrate on the emotional expressions and speaking styles of characters (Li et al., 2023). This principle, key to the 'Craft' element of RoleCraft, enriches dialogues with diverse emotions and distinctive speech patterns, effectively capturing the unique emotional states and communication styles of each character. Furthermore, the 'Accurate Application of Character Knowledge' principle emphasizes incorporating each character's background and experiences into the dialogue generation process (Shao et al., 2023). This ensures that the dialogues are in harmony with the characters' personas, encompassing their unique experiences and insights. Concluding with the 'Context-Aware Dialogue Generation' principle, our system is designed to dynamically tailor dialogues based on the prevailing context (Zhanga et al., 2023). This is crucial for maintaining a seamless and logically consistent conversation flow, essential for immersive and credible role-playing experiences.

3.1 RoleCraft-GLM Framework

Our methodology, guided by key design principles, uniquely advances the capabilities of LLMs in role-playing. Setting ourselves apart from approaches such as RoleGLM (Wang et al., 2023), we focus on an innovative integration of fine-grained character portrayal, profound emotional depth, and heightened contextual awareness in dialogue generation. This approach differentiates our work from existing models and addresses challenges in a novel way, enhancing how LLMs can be utilized for creating

[2]https://www.personal.ai/
[3]https://ai.meta.com/

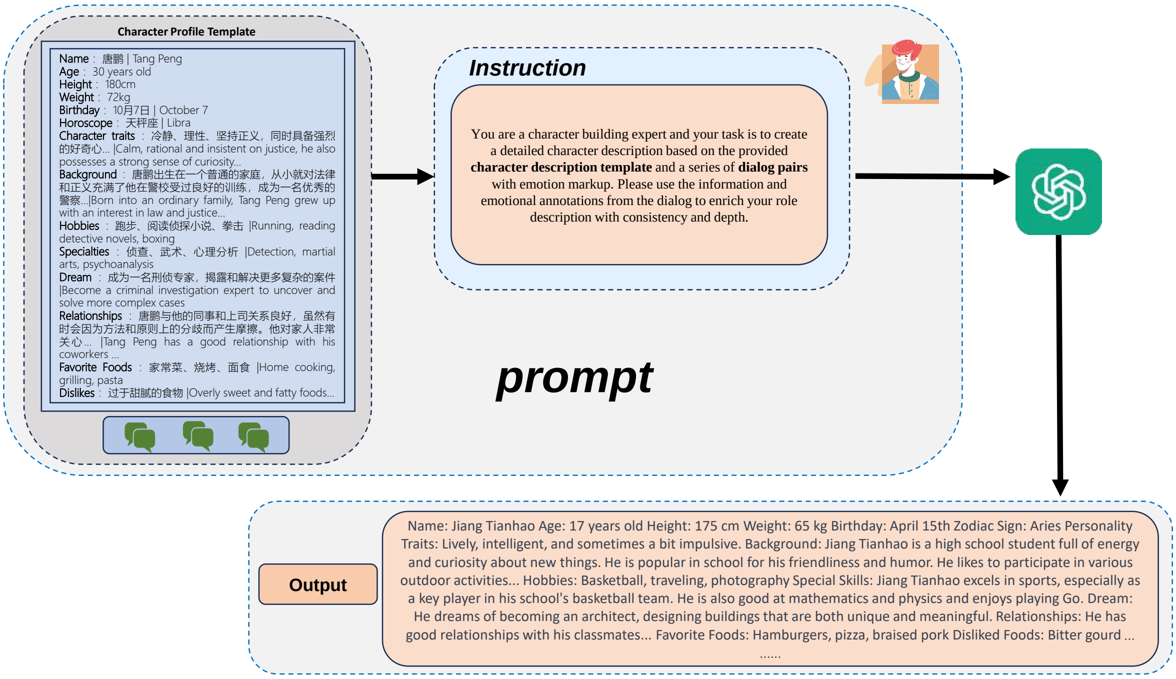

Figure 2: Here is an example of generating a detailed character description. Utilizing a character description template along with an emotionally annotated dialogue dataset facilitates the generation of detailed character descriptions based on prompts.(The instruction and output have been translated into English.)

more realistic and engaging role-playing scenario.

Emotion-Driven Character Profiling

To address the challenges of limited emotional diversity and unconvincing character portrayals in dialogues generated by LLMs, we've adopted a detailed emotion classification strategy. This approach involves meticulously annotating emotions within the dialogue dataset, thereby steering the GPT-4 to craft character profiles that mirror these identified emotions. Consider a character who displays a spectrum of emotions from joy to disappointment. Marking these diverse emotional states allows for a natural and fluid transition in their dialogues within a single scene, effectively capturing the complexity and dynamism of human emotions. This approach challenges the LLMs to accurately depict these emotional shifts, ensuring that the dialogues genuinely represent the intricate and ever-evolving nature of human emotions, thus enhancing the overall user interaction experience.

Contextual Q&A Generation

To address the challenge of context-irrelevant responses, which is a common issue in dialogue systems where interactions often lack relevance to the ongoing scenario or character specifics, we employ GPT-4 to generate contextually coherent Q&A pairs. For example, when a character faces a dilemma, the system is designed to produce queries and responses that align with the character's established traits, such as indecisiveness and anxiety, thereby maintaining the authenticity of the dialogue in relation to the character's profile.

Hybrid Instruction-Based GLM Refinement

Our methodology employs a hybrid training approach that seamlessly integrates general instructions with character-specific Q&A pairs. This strategy is carefully crafted to strike a balance between the flexibility required for dynamic dialogue generation and the need to uphold character integrity. In practical terms, this means equipping the LLMs to adeptly navigate a spectrum of conversational scenarios. These range from broad, general interactions to more intricate exchanges that demand responses finely tuned to the unique profiles of individual characters. By training the LLM with this diverse mix of inputs, the model becomes proficient in handling various situational dialogues, accurately reflecting each character's distinct attributes and the specific subtleties of the conversation. As a result, this hybrid training method fosters the creation of dialogues that are both contextually adaptive and consistent with the characters' distinct personalities.

3.2 Semantic-Enhanced Retrieval Optimization

Addressing the issue of inaccurate and semantically irrelevant information retrieval in dialogues, we have adopted the BGE[4] retrieval method. BGE is an efficient Chinese and English semantic vector model that ensures the accuracy of responses, especially when dealing with sensitive topics, and remains semantically sensitive to the context, significantly enhancing the quality of interaction. (Xiao et al., 2023). This familiarity allows models to generate dialogue based on a wealth of pre-existing knowledge. In contrast, modern datasets prioritize the nuanced portrayal of personal and everyday characters. These datasets are derived from diverse sources, including real chat logs, customer service interactions, and fictional narratives from less mainstream media. Such characters might include a typical office worker dealing with daily stressors or a mother showing love and responsibility in a family setting. The dialogues here involve specific, real-life scenarios, such as office interactions or typical family conversations, which lack the broad pre-existing knowledge base associated with public figures.

4 Experiments

4.1 Dataset

In the evaluation of machine learning models, the role of datasets is paramount, particularly in language processing and character portrayal. Traditional datasets predominantly highlight eminent figures, such as the legendary Sun Wukong, whose familiar attributes and stories are widely acknowledged, facilitating model development (Sabadoš, 2021). However, these datasets often neglect the finer details and emotional complexity of lesser-known or everyday characters, leading to a representation gap (Rolf et al., 2021). Our unique dataset bridges this gap by focusing on the rich, nuanced depiction of ordinary individuals. It involves an in-depth exploration and portrayal of each character's distinct personality traits and emotional depths, delving into aspects usually overshadowed in dominant narratives.

In constructing our dataset, we designed 20 unique and personalized Chinese characters to mirror a wide spectrum of real-world dialogues. These characters ranged from everyday individuals to fictional ones inspired by scripts. Our diverse data sources included social media interactions, film and television scripts, and customer service dialogues. We emphasized personal privacy and copyright law compliance, ensuring all data was cleansed and anonymized.

We filtered out redundant data and multi-party conversations to reshape the original data into contextually relevant dialogues. For example, script-based dialogues were restructured to better depict character interactions and emotional dynamics. Table 1 provides basic statistics for RoleCraft-GLM. Our final dataset comprises 27,259 multi-turn dialogues, distinctly different from datasets like Reddit comments (Al-Rfou et al., 2016), Sina Weibo (Shang et al., 2015), and Twitter datasets (Ritter et al., 2011), which mostly capture less structured, multi-participant interactions.

In addition, we annotated each dataset entry with emotion labels to capture characters' distinct emotional traits, adding an emotional layer to model training. We used Ekman's "Six Basic Emotions Theory" (Ekman, 1992) to label utterances and included additional emotions like neutral, excited, and depressed, totaling ten categories. The use of emotion labels in dialogue datasets has been proven to enhance natural language processing by improving response retrieval and emotional relevance (Zhou et al., 2017). These labels also enrich conversational analysis and aid in building natural dialogue systems (Bothe et al., 2019).

4.2 Experiment Settings

In our research, we anticipate that fine-tuning our model using a specifically designed dataset for role-playing will result in superior performance in character portrayal compared to baseline models. This expectation is based on the customized nature of the dataset, which includes detailed emotional annotations and context-specific scenarios that are essential for nuanced character interactions. Through this specialized training, we expect our model to accurately capture and express the intricacies of character-specific language styles and emotional responses, surpassing baseline models that may lack such targeted training. Our experiments aim to validate this hypothesis and showcase the advanced capabilities of our model in role-playing tasks.

[4]https://github.com/FlagOpen/FlagEmbedding

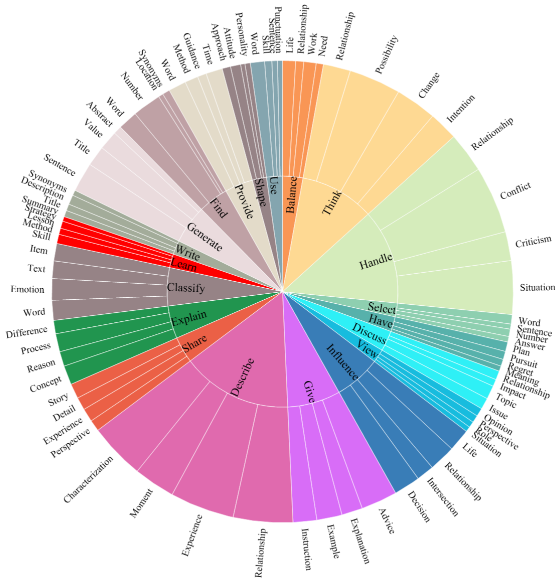

Figure 3: Verb-noun structure of Instructions. The inner circle representing the top 20 verbs and the outer circle listing the direct noun objects.

Category	Value
# Total Dialogues	2,7259
Avg.round of dialogues	14.64
# Characters	20
Character Personality Traits	45
Avg.length of profile	394.05
# Instructions	3,9422
Character-specific instructions	9842
General instructions	2,9580
Avg. instruction length	28.93
# Response	15,7742
Character-specific response	9842
General response	14,7900
Avg.response length	33.86

Table 1: Statistics of datasets

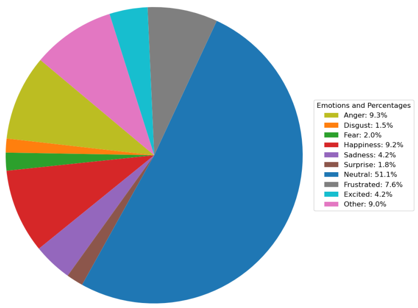

Figure 4: Emotion distributions in dialogues

Figure 5: The word cloud represents a visual distribution of personality traits for Chinese characters within the dataset, with larger words indicating a higher frequency of associated traits.

4.2.1 Baselines

We assessed the ChatGLM3 model[5], enhancing its performance on specific datasets using the Low-Rank Adaptation (LoRA) fine-tuning method (Hu et al., 2021). LoRA's precision in fine-tuning, essential for handling personalized and emotionally rich content, maintains the model's core capabilities while adapting to new data features. We benchmarked our RoleCraft-GLM's performance against industry standards such as GPT-3.5 and GPT-4, and leading Chinese dialogue generation technologies like ChatGLM2[6] and ChatGLM3, along with Baichuan2 (Yang et al., 2023) and Qwen (Bai et al., 2023). In our comparative experiments, we evaluated RoleGLM, which was fine-tuned using LoRA on specific datasets, as a benchmark. To match RoleGLM's setup, we also focused on a similar number of Chinese roles for consistency in our training approach. By selecting ChatGLM2 over ChatGLM3, we aimed to closely compare performances under equivalent conditions. This extensive evaluation underlined our model's distinctive advancements in dialogue generation.

4.2.2 Evaluation criteria

(1) Rouge-L Score: A commonly used metric (Lin, 2004) for evaluating natural language generation, measuring the overlap between model-generated text and real (ground truth) text. We focused on average score (Avg), general instruction response (RAW), role-playing speaking style (CUS), and specific role knowledge (SPE).

[5]https://github.com/THUDM/ChatGLM3
[6]https://github.com/THUDM/ChatGLM2-6B

Case	周小北，听说有两个泰国同学给你起了绰号，这让你感觉如何呢？ | Zhou Xiaobei, I heard that two of your Thai classmates gave you a nickname, how does that make you feel?
GPT-4	我对这个昵称其实感到既好笑又有点受宠若惊。虽然绰号听起来可能有点滑稽，但我知道我的泰国同学并没有恶意，他们只是在轻松地与我交流。 I actually felt both amused and a little flattered by the nickname. While the nickname may sound a bit funny, I know that my Thai classmates didn't mean any harm, they were just interacting with me in a relaxed manner.
Baichua2	听到有两个泰国同学给我起了绰号，我会感到好奇和高兴。这对我来说是一种友好的社交互动，我会以开放和幽默的态度来接受它。 I would be curious and happy to hear that two of my Thai classmates had given me a nickname. This is a friendly social interaction for me and I will accept it with openness and humor.
Qwen	我开始有点生气，因为他们是在取笑我，但后来我意识到或许我需要更了解文化差异，逐渐学习怎么应对。 I was a little angry at first because they were making fun of me, but then I realized that maybe I needed to be more aware of cultural differences and gradually learn how to deal with them.
Chatglm3	哦，那两个泰国同学给我起绰号吗？我觉得挺有趣的，我喜欢新奇和幽默，所以绰号对我来说是友情和创意的一部分，很开心！ Oh, those two Thai classmates gave me a nickname? I think it's funny, I like novelty and humor, so nicknames are part of friendship and creativity for me, it's fun!
Ours	哈哈，泰国同学给我起绰号？太有趣了！我好奇他们怎么想的，可能是因为我的奇思妙想吧！这个绰号我可得好好珍藏！ Haha, Thai classmates giving me nicknames? That's so funny! I wonder what they thought, probably because of my whimsy! That's a nickname I'm going to treasure!

Figure 6: A case of generated responses from our model and baseline models to a character-specific introduction.

Table 2: Rouge-L Evaluation

Model	Avg	RAW	CUS	SPE
GPT-3.5	0.4532	0.5569	0.5197	0.2831
GPT-4	0.4633	**0.5661**	**0.5264**	0.2973
ChatGLM2	0.4054	0.5104	0.4063	0.2996
ChatGLM3	0.4161	0.5218	0.4159	0.3108
Baichuan2	0.4273	0.5308	0.4576	0.2934
Qwen	0.4264	0.5297	0.4617	0.2879
Ours	**0.4704**	0.5385	0.5154	**0.3573**

Model	Avg. Ranking
ChatGLM3	1.86
Baichuan2	2.95
Qwen	3.38
Ours	**1.42**

Table 3: GPT-4 Evaluation

(2) Average Ranking Using GPT Scoring: In our work, we used the GPT-4 score (Fu et al., 2023) to evaluate the average rank of models on different dialogue-generating tasks, focusing on two main criteria for scoring: first, the distinctiveness and accuracy of the character's speaking style in matching their profile, and second, the richness of character-related knowledge and memory incorporated into the dialogues.

(3) Comparison of Emotionally Annotated and Non-Annotated Models: We compared models with and without emotional annotations in specific role knowledge memory (SPE) to evaluate the role of emotional annotation in enhancing model performance.

4.3 Performance Analysis

Results from Tables 2 and 3 clearly demonstrate our model's exceptional performance across multiple key performance indicators, particularly in specific role knowledge memory (SPE). Our model significantly outperformed GPT-4 and other models with a score of 0.3573 in this dimension, highlighting its superior ability in understanding and generating complex dialogues involving specific roles. However, GPT-4 leads in general instruction response accuracy (RAW) with a score of 0.5661, reflecting its strong capability in interpreting and responding to general instructions. Our model, while slightly less efficient in this dimension, still maintains a high score, proving its effectiveness in handling everyday dialogues. These findings underscore our method's significant effectiveness in deepening role understanding and enhancing dialogue generation quality. In particular, in emotion-driven role portrayal, our model demonstrated precise capture of each role's emotional traits and personality, surpassing traditional role-playing models in expressing role-specific knowledge. Moreover, our model not only excelled in generating role-specific dialogue content but also showed remarkable ability in maintaining natural flow and contextual consistency.

Table 4: Rouge-L Evaluation

model	Avg	RAW	CUS	SPE
RoleGLM	0.4570	0.5255	0.5049	0.3406
Ours	0.4641	0.5251	0.5128	0.3544

The Rouge-L assessment results from Table 4 indicate our model's performance improvement in three dimensions compared to RoleGLM. This suggests that our dataset enabled the model to more accurately capture and reflect the everyday characteristics and emotional states of roles, thereby enhancing the dialogue's realism and personalization. Additionally, the inclusion of emotional annotation

Method	SPE
RoleCraft-GLM(w/o emo)	0.3362
RoleCraft-GLM(w emo)	0.3573

Table 5: Comparing Emotion-Annotated and Non-Annotated Data

further enhanced the model's ability to understand and generate nuanced emotional dialogues, especially evident in the high SPE scores, indicating our method's outstanding performance in specific role knowledge and memory.

In our ablation experiments (see Table 5), we compared two versions of the RoleCraft-GLM model: one incorporating emotional annotations and the other without them. The primary objective was to assess the specific impact of emotional annotation on improving the model's scores in specific role knowledge memory (SPE). The results showed that the RoleCraft-GLM model with emotional annotation scored higher in SPE than the version without it. This difference emphasized the importance of emotional annotation in improving the model's understanding and generation of dialogues for non-celebrity roles that are closer to everyday life and lack extensive public knowledge or prior information. In these cases, emotional annotation not only provided the model with key information to deeply understand the roles' emotional states and personality traits but also ensured that the generated dialogues were closer to the roles' true feelings and personalized expressions.

5 Conclusions

In this paper, we present RoleCraft-GLM, our innovative framework aimed at enhancing personalized role-playing experiences. Central to this framework is a unique dataset featuring 20 diverse, non-celebrity Chinese characters, each with distinct emotional annotations. This shift from traditional celebrity-focused characters to more authentic, everyday personas marks a significant advancement in language modeling. RoleCraft-GLM's dataset emphasizes real-life scenarios and emotional depth, setting new standards in natural language processing. Our evaluations demonstrate that RoleCraft-GLM excels in creating dialogues that are not only contextually rich but also emotionally nuanced, outperforming conventional models. In the future, we hope to to develop behavioral agents that excel in personalization and interactivity, skillfully tailored to individual user preferences, thereby elevating the level of user engagement.

Acknowledgment

We would like to thank Genfun.ai for providing computation resources and APIs. We would also thank Minkun Yu and Zekun Wang for helpful suggestions.

References

Samuel Ackerman, Ateret Anaby-Tavor, E. Farchi, Esther Goldbraich, George Kour, Ella Ravinovich, O. Raz, Saritha Route, Marcel Zalmanovici, and Naama Zwerdling. 2022. High-quality conversational systems. *ArXiv*, abs/2204.13043.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*.

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. *arXiv preprint arXiv:2309.16609*.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Chandrakant Bothe, C. Weber, S. Magg, and S. Wermter. 2019. Enriching existing conversational emotion datasets with dialogue acts using neural annotators. *ArXiv*, abs/1912.00819.

P. Brandtzæg and A. Følstad. 2018. Chatbots: changing user needs and motivations. *Interactions*, 25:38–43.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9465–9480.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Sangwon Lee and Richard J Koubek. 2010. Understanding user preferences based on usability and aesthetics before and after actual use. *Interacting with Computers*, 22(6):530–543.

Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. 2023. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*.

OpenAI. 2023. Gpt-4 technical report.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alan Ritter, Colin Cherry, and Bill Dolan. 2011. Data-driven response generation in social media. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Esther Rolf, Theodora Worledge, B. Recht, and M.I. Jordan. 2021. Representation matters: Assessing the importance of subgroup allocations in training data. *ArXiv*, abs/2103.03399.

Noemi Sabadoš. 2021. Automatsko generisanje skupa podataka za treniranje modela za automatsko prepoznavanje osobe na slici. 36:536–539.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Hongbo Zhanga, Chen Tang, Tyler Loakmana, Chenghua Lina, and Stefan Goetze. 2023. Cadge: Context-aware dialogue generation enhanced with graph-structured knowledge aggregation. *arXiv preprint arXiv:2305.06294*.

Zhi-Min Zhou, Man Lan, Yuanbin Wu, and Jun Lang. 2017. Single turn chinese emotional conversation generation based on information retrieval and question answering. *2017 International Conference on Asian Language Processing (IALP)*, pages 103–106.