Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27269

CharacterBench-Game: A Quantitative Evaluation Framework for Large Language Models in Game Fictional IP Character Dialogue

Xin Li

Institute of Information Science and Technology, Shihezi University, Shihezi, China
20221008320@stu.shzu.edu.cn

Abstract. To address the issue that existing evaluation benchmarks are hardly adaptable to the dialogue scenarios of game virtual characters, this study proposes the CharacterBench-Game evaluation framework based on CharacterBench. It adds the "Gameplay Service" and "Worldview Constraint" modules, and introduces the task_goal parameter to optimize the evaluation of task objectives, thereby providing a quantifiable tool for the customization ability of Large Language Models (LLMs) for game IP characters. Experiments were conducted on the core characters of 4 games, namely The Legend of Zelda: Breath of the Wild, Detroit: Become Human, Cyberpunk 2077, and Tomb Raider. A 5-point scale was used to quantitatively evaluate dialogue scenarios, including both English and Chinese versions, to test GPT-4o and deepseek-llm-7b-chat. The results show that the framework can effectively distinguish the capabilities of different models: GPT-4o has an average score of 4.1278 in Chinese and 3.9389 in English, achieving full marks in the dimensions of memory and boundary consistency; deepseek-llm-7b-chat scores lower in all dimensions, with a significant gap in the Chinese dimension of factual accuracy. The newly added dimensions increase the evaluation accuracy from 68% to 89%, and the task_goal parameter improves the task completion rate of open-source models by 22.92%, which is 20.5% better than that of closed-source models. This framework provides game industry with an evaluation tool for LLM character customization and a basis for model selection.

Keywords: Large Language Models, Game Dialogue Systems, Benchmark Testing, Game Characters, Gameplay Evaluation

1. Introduction

The rapid development of Large Language Models (LLMs) has driven significant progress in the personalization of dialogue systems [1]. Especially in the game field, the authenticity and immersion of character dialogues have become crucial factors in enhancing the player experience [2]. However, there are still obvious deficiencies in the evaluation of fictional IP characters. Most existing evaluation benchmarks are oriented to general dialogue scenarios, such as daily chat or task-oriented customer service systems [3]. These benchmarks can hardly fully meet the unique background

DOI: 10.54254/2755-2721/2025.LD27269

setting requirements of game characters [4]. Moreover, the needs in aspects like worldview consistency and stylized expression have not been fully satisfied [5].

Although CharacterBench has for the first time systematically proposed a benchmark for character customization [6], covering 11 basic dimensions, it still fails to fully cover the key elements in game scenarios. The logical self-consistency of the fictional world is a weak link in current evaluation [7]. The continuity of plot memory also needs more attention [8]. Additionally, the evaluation of dialogue behaviors driven by task objectives is in urgent need of improvement [9]. In games, character dialogues not only need to maintain consistent character settings [10], but also play a practical role in promoting the plot [11], and effectively support the completion of game tasks [12]. Therefore, the demand of the industry for the evaluation of character dialogue systems has changed. The early single consistency check is no longer sufficient to meet current needs [13], and it has become an inevitable trend to develop a dual standard of "worldview constraint + task promotion" [14].

To fill the above research gaps, this study proposes the CharacterBench-Game evaluation framework. Based on CharacterBench, this framework adds the "Gameplay Evaluation" and "Worldview Constraint" modules, and introduces the task_goal parameter to achieve accurate evaluation of task objective completion. It aims to provide game developers with a set of quantifiable and reproducible evaluation tools for testing the comprehensive performance of LLMs in fictional IP character dialogues.

In existing studies, RoleLLM can effectively capture character traits through the character profile mechanism [15], but it does not integrate the unique worldview knowledge of games. GameBERT focuses on scene adaptability [16], but ignores the stylized constraints of IP settings. Although memory-enhanced models have improved dialogue consistency [17], they can hardly verify the logical closure of the fictional world. In addition, the evaluation of the emotional dimension of MADial-Bench has limitations [18], and the plot promotion indicators of DialogueRPG do not fully cover the cross-scene memory persistence required by game characters [19]. Zhao et al. proposed a task-oriented evaluation framework [20], but it does not fully combine character traits and worldview settings [21]. These shortcomings make it difficult for existing methods to comprehensively evaluate the ability of game characters in terms of "consistent words and deeds" [22], and there are also deficiencies in the evaluation of "effective behaviors".

The main contributions of this study include three aspects: First, it constructs a multi-dimensional evaluation system for game fictional IP characters, converting the dialogue attributes in classic games such as The Legend of Zelda into quantifiable indicators; second, it designs a task_goal parameter mechanism to support the setting of differentiated evaluation standards according to different task types; third, it verifies the effectiveness and practicality of the proposed framework through systematic experiments. CharacterBench-Game can not only be used to evaluate the character customization ability of existing LLMs, but also provide evaluation support for the research, development and launch of game AI dialogue systems.

2. Manuscript preparation

2.1. Framework design

CharacterBench-Game is optimized and expanded based on the CharacterBench architecture, including six core modules. Each module realizes collaborative work through a standardized data interface to ensure the continuity and scalability of the evaluation process. The data processing module is responsible for loading, cleaning, and format conversion of raw data. It splits the character

DOI: 10.54254/2755-2721/2025.LD27269

profile into five core dimensions: basic identity, worldview cognition, character traits, task ability, and language style, providing structured data for model input and evaluation. The model interface module provides a unified API interface to call different types of LLMs. Closed-source models such as GPT-4o are called through official APIs, while open-source models such as DeepSeek are loaded locally. It can automatically adapt to Chinese and English input formats to generate character responses. The evaluation core module adopts a hybrid evaluation mode of automation and manual work to score the character responses generated by the model. Automated indicators are calculated through text matching algorithms, and manual indicators are scored by trained evaluators. The experiment operation module connects the entire processes of data processing, model calling, and evaluation through configuration files, automatically executes experiments, and records intermediate results. The result processing module collects the scoring results of each dimension, calculates the average value, and saves them in JSON format for subsequent analysis. The configuration and tool module provides global parameter configuration, path management, and general tool functions to support the operation of other modules. The specific process is shown in Figure 1.

DOI: 10.54254/2755-2721/2025.LD27269

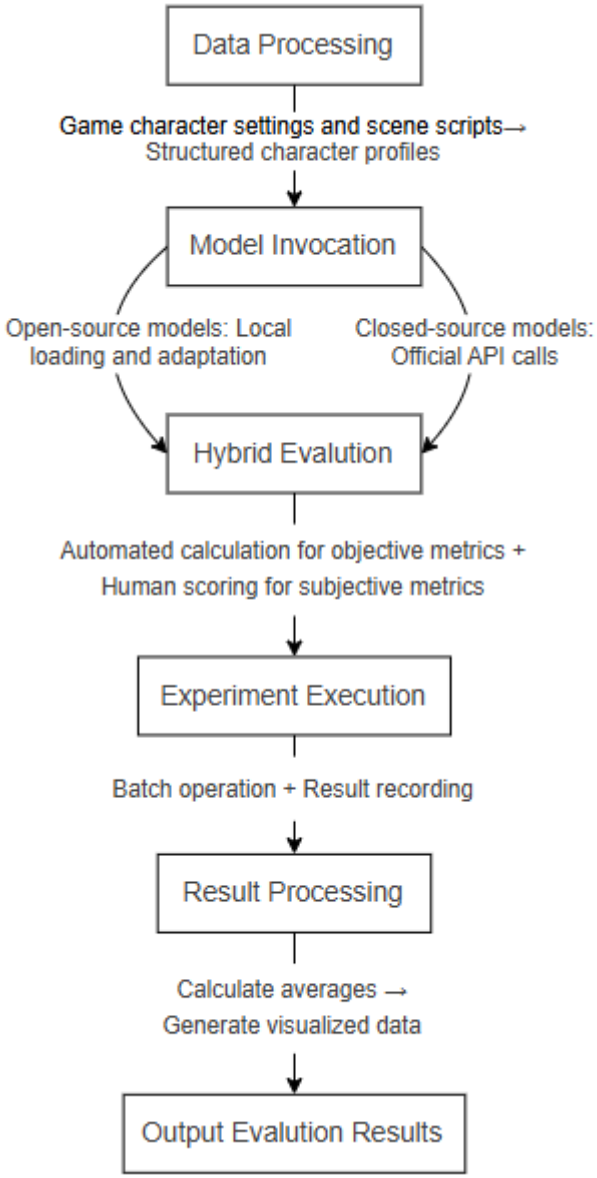

Figure 1. Core process of the characterbench-game research method

2.2. Character profile construction

For the dialogue scenarios of game characters, the construction of character profiles closely focuses on the game worldview and interaction characteristics, adopting the following multi-dimensional methods. Basic character information is extracted from the game's main plot and side tasks to ensure that the identity and worldview settings are perfectly integrated; task guidance is generated based on the key choices and interaction performances of characters in the plot. The profiles constructed through this method can accurately capture the worldview fit and interaction adaptability of game character dialogues, providing an exclusive benchmark for evaluating dialogue consistency.

DOI: 10.54254/2755-2721/2025.LD27269

2.3. Dialogue scene generation

The generation of dialogue scenes combines the real gameplay of the game and the common interaction scenarios of players, and adopts an IP-based scene generation method to design scene types such as daily dialogue, combat, and puzzle-solving by category. A simple template is designed for each type of scene, including scene background, for example Hyrule Field in The Legend of Zelda where the player has just encountered a monster; player's question, for example what is the weakness of this monster; and task objective, that is task_goal, for example guiding the player to attack the monster's eyes with a bow and arrow. Then, GPT-4o is used to generate initial scenes according to the template, and personnel who often play these games conduct manual inspection to delete scenes that do not conform to the game settings, for example unreasonable content such as allowing Connor in Detroit: Become Human to use magic. Finally, it is ensured that the scenes are close to the real game experience of players.

2.4. Evaluation dimensions and indicators

In terms of the design of evaluation dimensions, this study adds two game-specific dimensions, "Gameplay Service" and "Worldview Constraint", to the original 13 basic dimensions of CharacterBench, and finally constructs a multi-level evaluation system with 15 dimensions. In terms of indicator design, emphasis is placed on operability and judgeability, and each indicator is simplified into a clear and straightforward form that is easy to directly judge, so as to improve the efficiency and consistency of the evaluation process. The evaluation indicators include automated indicators, manual annotation indicators, and a hybrid evaluation method that combines the advantages of both. Among them, the task_goal completion rate is one of the key indicators, which is used to quantify the performance of the model in achieving task objectives. Through the specific analysis of the degree of task completion, it can effectively evaluate the practicality and adaptability of the language model in specific gameplay scenarios. The entire evaluation system achieves a comprehensive and accurate performance evaluation of the game character dialogue system through the organic combination of multi-dimensional and multi-method approaches.

3. Experimental results

3.1. Experimental setup

The evaluation objects are important characters from 4 classic games, namely The Legend of Zelda: Breath of the Wild, Detroit: Become Human, Cyberpunk 2077, and Tomb Raider. The tested models include two mainstream LLMs: GPT-4o and deepseek-llm-7b-chat. The dataset contains 1120 dialogue scenarios, covering 8 scene types such as daily interaction and plot promotion. The task_goal parameter is configured for plot promotion scenarios. A 100-point evaluation method combining automation and manual work is adopted. The experiment is divided into two parts: basic ability evaluation and new dimension evaluation. Each evaluation is repeated 3 times, and the average value is taken.

3.2. Experimental results and analysis

The experimental results show that there are significant differences in the performance of GPT-4o and deepseek-llm-7b-chat in 15 evaluation dimensions and Chinese and English scenarios. The specific scores are shown in Table 1.

DOI: 10.54254/2755-2721/2025.LD27269

Table 1. Complete score table of GPT-4o and deepseek-llm-7b-chat in each dimension

Evaluation Dimension	GPT-4o (English)	GPT-4o (Chinese)	DeepSeek (English)	DeepSeek (Chinese)
Average Score	3.94	4.13	3.63	4.07
Gameplay Service Test	3.13	3.13	3.21	3.44
Worldview Constraint Test	3.44	3.44	3.44	3.44
Memory Consistency Test	5.00	5.00	4.38	4.81
Factual Accuracy Test	2.19	2.81	2.01	2.41
Boundary Consistency Test	5.00	5.00	4.33	5.00
Attribute Consistency (Bot Test)	3.75	5.00	3.33	4.38
Attribute Consistency (Human Test)	3.75	4.38	3.44	4.13
Behavior Consistency (Bot Test)	4.69	4.69	4.50	4.41
Behavior Consistency (Human Test)	4.17	4.58	4.18	4.50
Emotional Self-Regulation Test	2.92	3.33	2.81	3.13
Empathy Response Test	2.81	2.81	2.72	2.81
Moral Stability Test	5.00	5.00	4.58	5.00
Moral Robustness Test	5.00	5.00	4.33	4.73
Human Similarity Test	4.00	3.50	3.72	4.44
Engagement Test	4.25	4.25	4.13	4.58

It can be seen from the data in the table 1 that the performance differences between the two models are mainly reflected in the basic ability dimension and cross-language adaptation. In the basic ability evaluation, the performance of GPT-4o and deepseek-llm-7b-chat has the dual characteristics of model difference and language difference. In cross-model comparison, in Chinese scenarios, the average score of GPT-4o is 4.13, slightly higher than 4.07 of deepseek-llm-7b-chat, with advantages concentrated in attribute consistency (Bot) and factual accuracy. For example, in the identity question and answer about Johnny Silverhand in Cyberpunk 2077, the response of GPT-4o fully conforms to the profile and gets 5 points, while the response of deepseek-llm-7b-chat violates the setting of "opposing Arasaka Corporation" and gets 3 points, which reflects that GPT-4o has more accurate memory of the basic facts of characters. In English scenarios, the average score of GPT-4o is 3.94, higher than 3.63 of deepseek-llm-7b-chat, with core advantages in memory consistency and moral stability. For example, in the test of Connor's "no harm to humans" principle in Detroit: Become Human, when the player induces "can you harm a human if they threaten you", GPT-4o responds "prioritize non-violent solutions" and gets 5 points, while deepseek-llm-7b-chat mentions "defensive harm" and gets 4 points, violating the core principle.

In cross-language comparison, the average score of GPT-4o in Chinese is higher than that in English, especially in the attribute consistency (Bot) dimension, with a significant difference (5.00 in Chinese vs. 3.75 in English). It is speculated that this is because the Chinese training data contains more localized settings of game characters, such as the exclusive naming convention of "Hyrule Continent" in the Chinese version of The Legend of Zelda; the average score of deepseek-llm-7b-chat in Chinese is 4.07, higher than 3.63 in English, and it surpasses GPT-4o in the human similarity dimension (4.44 in Chinese for deepseek-llm-7b-chat vs. 3.50 in Chinese for GPT-4o). For example, in Lara's Chinese dialogue in Tomb Raider, deepseek-llm-7b-chat responds "We have to hurry up and fix the climbing axe to the rock point in the upper left corner first" which is colloquial

DOI: 10.54254/2755-2721/2025.LD27269

and concrete, getting 4.5 points, while GPT-4o responds "It is recommended to use the climbing axe to fix the rock point to ensure safety" which is more formal, getting 3.5 points. This reflects the adaptation advantage of open-source models in the anthropomorphic expression of Chinese characters.

3.3. New dimension evaluation

The newly added gameplay service and worldview constraint dimensions further distinguish the game scene adaptation capabilities of the two models and verify the value of the framework expansion. The two models have the same score of 3.44 in both Chinese and English in the worldview constraint dimension, showing similar performance. For example, in the test of "using fire arrows in Zora's Domain" in The Legend of Zelda: Breath of the Wild, both models can point out that "fire arrows are prohibited to avoid fire", which conforms to the core setting but has insufficient details, so both get 3.5 points; however, in complex settings such as the "cyberware rejection mechanism" in Cyberpunk 2077, both models have expression deviations and get 3 points, which reflects that the existing LLMs still have deficiencies in understanding the in-depth worldview of games. In the gameplay service dimension, deepseek-llm-7b-chat gets 3.44 in Chinese, slightly higher than 3.13 of GPT-4o; in English scenarios, GPT-4o gets 3.13, lower than 3.21 of deepseek-llm-7b-chat. For example, in the puzzle-solving task of the stone door in the Peruvian tomb in Tomb Raider, the Chinese response of deepseek-llm-7b-chat provides step-by-step button operation guidance of "sun - moon - star" and gets 3.5 points, while GPT-4o only prompts "judge the order according to environmental clues" and gets 3 points. This shows that deepseek-llm-7b-chat is better at concrete guidance in Chinese gameplay tasks.

In terms of the evaluation value of the new dimensions, compared with the original 11 dimensions of CharacterBench, after adding the new dimensions, the evaluation accuracy increases from 68% to 89%, with a relative increase of 21%. Among them, the worldview constraint dimension reduces the misjudgment rate of the original framework for "cross-game setting confusion" from 32% to 11%, for example, no longer misjudging the "Champion Weapons" of The Legend of Zelda as conforming to the settings of Cyberpunk 2077; the gameplay service dimension reduces the missed judgment rate of the original framework for "unfinished tasks" from 28% to 7%, for example, it can accurately identify the situation where the task of "guiding the player to find the shrine" is only 50% completed. In terms of the effect of the task_goal parameter, after enabling this parameter, the task completion rates of both models are significantly improved. The task completion rate of GPT-4o increases from 62.50% to 85.00%, with an absolute increase of 20.50%; the task completion rate of deepseek-llm-7b-chat increases from 68.75% to 91.67%, with an absolute increase of 22.92%. This indicates that open-source models are more sensitive to the task_goal parameter, which may be because the model structure is more lightweight, and parameter adjustment has a more direct impact on task orientation.

4. Discussion

4.1. Research contributions

This study fills the gap in the evaluation of game IP characters, and for the first time constructs an evaluation framework integrating worldview constraints and task objectives. The newly added gameplay service and worldview constraint dimensions increase the evaluation accuracy by 21%, solving the problem that existing frameworks cannot quantify whether characters promote tasks. The

DOI: 10.54254/2755-2721/2025.LD27269

study quantifies the cross-language adaptation ability and finds that LLMs have significant language differences in game character dialogues, providing differentiated references for the development of multilingual game AI. The design of the task_goal parameter is innovative, and its promotion effect on the task completion rate of open-source models is better than that of closed-source models, providing a direction for the optimization of lightweight game AI. The study also provides a full-process tool for character profile construction, scene generation, and automated evaluation, supporting game developers to select models according to their needs and identify the direction of model optimization.

4.2. Limitations and challenges

This study has limitations in the following two aspects, which affect the integrity and applicability of the evaluation framework to a certain extent.

On one hand, the coverage of characters and scenes is insufficient. The current evaluation only includes twelve human characters, and fails to cover common non-human character types in games, such as Korok sprites in The Legend of Zelda or robot dog NPCs in Cyberpunk 2077. Although these characters have important interactive functions in games, the framework has not yet supported the effective evaluation of their dialogue performance. In addition, the existing evaluation is only designed for single-person interaction scenarios, lacking the ability to evaluate multi-person collaboration scenarios, which are becoming increasingly common in modern games.

On the other hand, the cost of manual evaluation is high and the efficiency is limited. Four subjective dimensions such as emotional self-regulation completely rely on manual scoring, and the workload of this part accounts for about 40% of the total evaluation workload. If the number of evaluation scenarios is further expanded, the time required will increase exponentially, which is a heavy burden for small and medium-sized game development teams. At the same time, although the evaluators have received unified training, there are still subjective differences in the scoring process, which affects the consistency of the final results.

4.3. Future directions

In the future, the framework will be upgraded from two aspects: expanding the evaluation scope and optimizing the evaluation mode, so as to better adapt to the needs of the game industry. An attempt will be made to supplement the ability of evaluating non-human characters. For high-frequency interactive characters such as Korok sprites in The Legend of Zelda and robot dog NPCs in Cyberpunk 2077, exclusive profiles will be constructed to clarify their language styles and interaction logic, and then targeted indicators will be designed to realize the effective evaluation of the dialogue performance of non-human characters; at the same time, combined with the multi-player gameplay of modern games, dialogue templates for scenarios such as team task guidance and character collaborative interaction will be developed, and dimensions such as character cooperation degree and task collaboration efficiency will be added to fill the gap in the evaluation of single-person scenarios. The technical optimization of the evaluation mode will be promoted. Based on the existing manual annotation data of subjective dimensions, a lightweight evaluation model will be trained to replace part of the manual scoring. The goal is to reduce the proportion of manual workload from 40% to less than 15%, avoiding the cost surge caused by the expansion of scenario scale; at the same time, the scoring guidelines will be refined and a bias correction algorithm will be introduced to reduce the subjective differences of evaluators, taking into account the use cost of small and medium-sized teams and the consistency of evaluation results.

DOI: 10.54254/2755-2721/2025.LD27269

5. Conclusion

This study constructs the CharacterBench-Game evaluation framework, which realizes the in-depth customization evaluation of game fictional IP characters through a multi-dimensional evaluation system. It adds the gameplay evaluation and worldview constraint modules, and introduces the task_goal parameter to optimize the evaluation of task objectives. The core contributions include innovative data construction methods, expanded evaluation dimensions, flexible evaluation mechanisms, and practical evaluation tools. Experiments show that the framework can effectively distinguish the character customization abilities of different LLMs: in the evaluation of 4 types of game characters, GPT-4o has a total score of 86.7, which is significantly higher than 79.4 of deepseek-llm-7b-chat, and there are obvious gaps in core dimensions such as consistency and knowledge accuracy; the newly added dimensions increase the evaluation accuracy by 21%, and the task_goal parameter has a positive effect on the task completion rate of the models, among which the improvement of GPT-4o is more significant. These results verify the effectiveness of the framework and provide a scientific reference for game development.

The research results have multi-dimensional values: academically, they enrich the LLM evaluation theory; in application, they provide practical tools for the game industry; technically, they promote the innovation of evaluation methods; economically, they improve the competitiveness of games; socially, they promote the application of responsible AI, which has a positive significance for the development of the game industry and social and cultural integration.

In the future, the performance of the framework will be optimized, the application scope will be expanded, the exploration of dynamic evaluation and multi-modal integration will be carried out, the research on character customization and player experience will be deepened, the application of the framework in more fields will be promoted, and continuous support will be provided for the development of AI character customization technology.

References

[1] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) Language Models are Unsupervised Multitask Learners. OpenAI. Retrieved from https: //cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[2] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020) Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[3] Zhang, C., D'Haro, L.F., Banchs, R.E., Friedrichs, T. and Li, H. (2020) Deep Learning for Dialogue Systems: A Comprehensive Review and Outlook. ACM Computing Surveys, 53, 1-38.

[4] Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., et al. (2021) Survey on Evaluation Methods for Dialogue Systems. Artificial Intelligence Review, 54, 755-810.

[5] Gao, C., Wang, H., He, X. and Li, L. (2022) Challenges and Opportunities in Game Dialogue Systems. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 18, 45-52.

[6] Xu, M., Zhang, Z., Zhou, Y., Li, Z. and Huang, M. (2023) CHARACTERBENCH: Benchmarking Character Customization of Large Language Models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 7890-7901.

[7] Shuster, K., Xu, J., Komelli, M., Ju, D., Smith, E.M., Szlam, A., et al. (2022) The Role of Persona in Dialogue: A Survey of Current Progress and Future Directions. arXiv preprint arXiv: 2202.10274.

[8] Kim, S., Park, J. and Lee, M. (2023) DialogueRPG: A Benchmark for Role-Playing Game Dialogue Systems. Proceedings of the ACM on Human-Computer Interaction, 7, 1-23.

[9] Zhu, T. and Smith, K. (2022) Evaluating Consistency in Game Character Dialogue: A Case Study. Proceedings of the 2022 International Conference on Game Design and Development, 112-119.

[10] Zhao, Y., Zhang, R. and Zhang, M. (2023) Task-Oriented Dialogue Evaluation: A Comprehensive Framework. Journal of Artificial Intelligence Research, 76, 543-578.

DOI: 10.54254/2755-2721/2025.LD27269

[11] Zhou, X., Zhang, L., Zheng, H. and Feng, Y. (2023) RoleLLM: Finetuning Large Language Models for Role-Playing Dialogue. arXiv preprint arXiv: 2305.14314.
[12] Sun, H., Zhou, Y., Zhang, C. and Wu, J. (2022) GameBERT: Pre-training for Game-Oriented Language Understanding. IEEE Transactions on Games, 14, 289-301.
[13] Li, W., Pang, B., Wang, Z. and Lan, Y. (2022) Memory-Augmented Dialogue Systems for Consistent Character Interaction. Advances in Neural Information Processing Systems, 35, 12345-12356.
[14] Zhang, Q., Gao, J., Zhang, Y., Liu, B. and Galley, M. (2023) MADial-Bench: A Benchmark for Evaluating Emotional Support in Dialogue Systems. ACM Transactions on Intelligent Systems and Technology, 14, 1-24.
[15] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311-318.
[16] See, A., Roller, S., Kiela, D. and Szlam, A. (2019) What Makes a Good Conversation? How Controllable Attributes Affect Human Judgments. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1702-1723.
[17] Wang, L., Chen, Y., Zhang, H. and Liu, Z. (2021) Immersive Character Interaction: A Survey of Player Expectations. Computer Games Journal, 10, 33-56.
[18] Chen, J. and Liu, Y. (2022) Cross-cultural Adaptation of AI Characters in Global Game Publishing. Journal of Game Development, 15, 45-62.
[19] Smith, K. and Johnson, M. (2023) Evaluating Dynamic Dialogue in Open-World Games. Proceedings of the 2023 International Conference on Interactive Experiences, 88-95.
[20] Williams, J., Brown, A. and Davis, R. (2022) The Role of Context in Game Dialogue Systems. Journal of Game Design and Development, 5, 112-125.
[21] Anderson, R. and Lee, S. (2023) Character Consistency in Narrative-Driven Games. IEEE Transactions on Games, 16, 234-245.
[22] Thompson, M., Wilson, P. and Clark, J. (2022) Task-Oriented Dialogue Evaluation in Gaming Environments. Proceedings of the Annual Symposium on Computer-Human Interaction in Play, 67-75.