DOI: 10.54254/2755-2721/2025.LD28270

Application of Large Language Models in Games

Junyou Chen[1], Ziyi Wang[2*], Wenyi Zhang[3]

[1]Zhong Guo High School, Shnaghai, China
[2]College of Cultural Industry Management, Hebei Institute of Communications, Shijiazhuang, China
[3]Singapore Institute of Management Global Education, Singapore, Singapore
**Corresponding Author. Email: wzhang036@mymail.sim.edu.sg*

Abstract. This paper explores the application of large language models (LLMs) in the gaming field. It begins by elaborating on the research background and significance, and reviews the current status of their application and research in the gaming domain. Subsequently, it introduces the characteristics, working principles, and development history of large language models. It then focuses on analyzing the application of large language models in games, taking LLM-controlled non-player characters (NPCs) and dynamic plot generation as examples to dissect their application methods and advantages in games. Finally, it discusses the challenges faced by large language models in gaming applications, such as high resource consumption and unstable interaction with players; at the same time, it looks forward to their impact on the development of the gaming field in the future, believing that with the continuous advancement of technology, they will promote the intelligent transformation of the gaming industry. This will play an important role in the future of the game industry.

Keywords: Large Language Model, Natural Language Interaction, Gaming Application, Game Agent, Dynamic Content Generation.

1. Introduction

Since 2020, large language models represented by the GPT series have made significant breakthroughs. Through in-depth training on massive amounts of data, they have acquired powerful language understanding and generation capabilities [1]. With the rapid iteration of large language models in recent years, they have also been widely applied in the gaming field. Their natural language processing and generation capabilities have brought new challenges and opportunities to game development. At the same time, players can obtain more complex, immersive, and personalized gaming experiences from these games compared to traditional games.

In games, the role of NPCs is usually to shape the narrative atmosphere and drive the plot development. However, in traditional games, NPC dialogue systems mostly rely on limited scripts, templates, or state machines, resulting in rigid interactions, high repetition, and lack of depth [2]. Recent studies have shown that large language models can generate more natural and diverse dialogue content, thereby endowing NPCs with communication abilities closer to humans [2]. In these games, by embedding generative artificial intelligence agents, NPCs can make dynamic decisions and conduct multi-turn dialogues based on the game environment and historical dialogue records, achieving open-ended interactions.

Researchers such as Park demonstrated an experiment of a simulated town in the "Generative Agents" framework. This experimental environment uses large language models to drive NPCs, enabling them to remember past events, plan daily activities, and make reasonable responses based on social relationships [2]. This research demonstrates the potential of large language models in building social ecosystems in virtual worlds. Similarly, researchers such as Akoury conducted a systematic study on players' perception of dialogues generated by large language models. The study found that participating players generally considered this mode more immersive but also pointed out that dialogues occasionally had issues such as inconsistency with settings or context [3].

NPCs driven by large language models can not only enhance players' gaming experience but also reduce the labor costs of game development. Traditional games usually require a lot of manpower and time to create and test dialogue scripts in the game. However, using large language models to automatically generate first drafts can help development teams save this part of the cost, allowing them to focus on optimizing game plots and mechanisms [4]. In addition, using large language models to drive NPCs can also call external systems to obtain real-time data or execute specific functions during conversations with players, thereby expanding the interactive capabilities of the game [4].

Although large language models have been widely used in the gaming field, there are still multiple challenges. Researchers such as Weidinger pointed out that large language models may bring ethical and social risks, including generating harmful content, reinforcing stereotypes, and privacy leaks [5]. Due to the diversity of player groups and frequent interactions with games, the potential impact of these risks is more significant. Therefore, when introducing large language models into games, it is necessary to implement effective mechanisms such as safety filtering and content review to ensure the safety and appropriateness of generated content.

Overall, large language models provide a brand-new way for NPC design in the gaming field, transforming them from passive script-executing programs into active, dynamic, and human-like emotional interactors. With the continuous improvement of model reasoning efficiency, controllability, and multimodal capabilities, it is foreseeable that large language models will be more widely applied in the gaming industry. At the same time, the balance between technological innovation and ethics should be considered to ensure that this technology can truly serve games and create personalized and immersive experiences for players. We will introduce Large Language Models and their role in games.

2. Overview of LLMs

2.1. Definition and characteristics

Large Language Models (LLMs) are deep learning models based on massive text data training, with billions to trillions of parameters, and adopting the Transformer architecture. They can capture language statistical rules through self-supervised learning and complete various natural language processing (NLP) tasks. The core idea is to learn the patterns and structures of natural language through large-scale unsupervised training, simulating human language cognition and generation processes to a certain extent. They have huge parameter scales, and their performance improves with the increase in scale; they can perform multiple tasks with a wide coverage; they obtain general language representation capabilities through pre-training and can adapt to downstream tasks through fine-tuning or prompt engineering; they can remember historical dialogues and provide contextaware answers, with more human-like behavior patterns; their performance will continue to improve with the addition of more massive datasets, showing sustainability [6].

2.2. Working principles and development history

The working principle of LLMs is based on the Transformer architecture, realizing in-depth text understanding and generation through self-supervised learning and probabilistic modeling: first, input text is tokenized, each token is mapped to a unique ID and then converted into a high-dimensional vector representation, while incorporating positional embedding to retain sequence order information; the self-attention mechanism in the Transformer architecture allows the model to dynamically focus on other relevant tokens when processing each token, thereby accurately capturing contextual semantic relationships, and the multi-head self-attention mechanism further enhances the modeling ability for different semantic dimensions; the output of self-attention undergoes nonlinear transformation through a feed-forward neural network, and the output of each sub-layer is transmitted to the next layer through residual connections, while layer normalization is used to stabilize the training process; in the training phase, the model calculates conditional probabilities based on context and continuously optimizes the ability to predict the next token; in the inference phase, it generates coherent text through methods such as greedy search or sampling based on the given context.

Its development history can be divided into three key stages: the Transformer revolution and early exploration stage from 2017 to 2019. Google proposed the Transformer architecture in 2017, which solved the limitations of traditional RNNs and CNNs in processing long sequences through the self-attention mechanism, realizing direct modeling and parallel computing of arbitrary positions in sequences. Subsequently, models such as BERT in 2018, GPT-2 in 2019, and Baidu's ERNIE 1.0 were successively launched, promoting progress in fields such as bidirectional language modeling and Chinese understanding; the large-scale breakthrough stage from 2020 to 2022. Models represented by GPT-3 demonstrated strong in-context learning and few-shot learning capabilities through large-scale expansion of parameter quantities and training data, verifying the feasibility of the large-scale path. Chinese manufacturers also intensively laid out in this stage, focusing on integrating structured knowledge into pre-training, and application scenarios expanded from traditional NLP to broader fields such as code generation and content creation; the multimodal fusion and agent rise stage from 2023 to 2025. Models show characteristics of breakthroughs in multimodal capabilities, improvement in long-context processing, and vigorous development of agent-based applications. The technical architecture shifts from simply expanding parameters to efficiency optimization, with the MoE architecture becoming mainstream [7-10]. At the same time, alignment technologies such as RLHF and Constitutional AI ensure that model behaviors conform to human values, promoting LLMs to develop in a more intelligent and secure direction.

3. Application of LLMs in games

3.1. NPC-related applications

3.1.1. Development history

In the early stage, NPC behaviors were completely pre-set by developers, resulting in low interactivity. In the development stage, with the continuous progress of information technology, NPCs gradually moved towards scripted development. At this time, NPCs could respond more richly to players' behaviors, but this interaction was still rigid, lacking sufficient dynamics. In the current stage, since OpenAI launched the GPT-4 chat-oriented language model in 2023, more and more researchers have begun to try to combine large language models (LLMs) with NPCs, aiming to provide players with more diverse interactive experiences.

3.1.2. Specific application methods

In terms of dialogue generation, LLMs have strong language processing capabilities and are very suitable for simulating real-person dialogue scenarios. NPCs that generate dialogue content with the help of LLMs can greatly enrich their dialogue systems, thereby effectively improving players' experience during gameplay. In dynamic narrate, LLMs have strong context awareness capabilities. When combined with technologies such as MemoryRepository, NPCs can remember historical dialogues and various events in the game, which not only enhances the coherence of the narrative but also improves the authenticity of the characters, allowing players to build long-term relationships with NPCs. In terms of dynamic decision-making and behavior, NPC behaviors in previous games followed fixed decision-making patterns with high predictability, leading to a lack of replayability in the game. The integration of LLMs can increase NPC behavior patterns, such as realizing the bargaining function of merchants in the game, thereby enhancing the replay value of the game and players' immersion. In terms of cooperation and guidance, LLMs can collect relevant information for players and guide them by setting goals to achieve cooperation between players and NPCs, which not only increases the fun of the game and players' experience but also helps new players to a certain extent.

3.2. Other application methods

In the gaming field, large language models (LLMs) are reshaping the player experience with innovative approaches, and their applications penetrate into multiple core dimensions such as player assistants, player roles, and game mechanisms. As intelligent player assistants, LLMs break through the limitations of traditional preset question-and-answer systems and can provide dynamic support to players through natural language interaction. For example, in open-world games, players can directly ask "how to find the hidden ancient ruins" in spoken language, and the assistant will generate personalized guidance based on the player's current progress, mission clues, and even past dialogue habits. It can also real-time parse complex skill tree systems, explain the advantages and disadvantages of different skill point allocation schemes in plain language, and even provide strategic suggestions and emotional encouragement when players encounter setbacks, making the assistance process more humanized.

In terms of player role shaping, LLMs endow non-player characters (NPCs) with unprecedented interaction depth. These characters are no longer limited to fixed lines and behavior patterns but can generate logical and personalized responses based on players' dialogue content, tone, and even character background stories. In martial arts games, the tavern owner may reveal hidden missions because players repeatedly mention "chivalry spirit"; in sci-fi games, alien allies will adjust their trust and dialogue attitude according to players' decision-making positions. This dynamic interaction makes character relationships full of realism and greatly enhances the plot immersion.

In game mechanism innovation, LLMs have become the key to breaking traditional design boundaries. They can real-time generate dynamic plot branches, and each dialogue choice of players may trigger new mission lines or changes in world status. For example, in fantasy RPGs, a player's argument with the elves may change the ecological rules of the entire forest area; at the same time, LLMs can empower procedural content generation, dynamically adjusting level difficulty, monster types, and even narrative styles according to player preferences, allowing each player to obtain a tailored gaming experience and promoting the in-depth transformation of the gaming industry from standardized production to personalized services.

4. Challenges and prospects

4.1. Technical level

4.1.1. Interaction delay

The delay problem in real-time interaction between players and NPCs (with an average of 7 seconds and a maximum of 24 seconds) highlights the core contradiction between the technical characteristics of large models and the needs of real-time scenarios. The challenges are reflected in three aspects: first, model computing and memory bottlenecks. The computing load of the Transformer architecture grows quadratically with the length of the dialogue sequence, making it difficult to balance accuracy and speed; second, hardware and deployment cost limitations. Cloud deployment has high costs and there are differences in network transmission delays; third, the scenario has "zero tolerance" for delays, and existing optimizations are prone to information loss or accuracy degradation. However, the solution path is clear, and future optimizations will be carried out from three aspects: in model algorithms, improving inference efficiency through technologies such as efficient Transformer variants and fusion operators to achieve sub-second delays; adopting a "cloud + edge" hybrid system architecture to reduce transmission delays; using semantic-aware dynamic windows for dialogue history management, combined with pre-computation and caching to optimize responses. The solution to this problem will not only promote the innovation of game NPC interactions but also provide key technical support for real-time scenarios such as the metaverse and virtual humans.

4.1.2. Inconsistent NPC personalities

In the field of LLM-controlled NPCs, there are currently multiple challenges. Technically, LLMs have bottlenecks in understanding complex long-term contexts and maintaining consistent character settings. The memory and computing efficiency issues of the Transformer architecture in processing long sequences can lead to NPC words and deeds deviating from settings. Model structures and training algorithms need to be optimized to capture personality characteristics in multi-turn dialogues. At the data level, it is difficult to obtain and annotate high-quality, large-scale data that conforms to NPC settings, with high costs and difficulty in ensuring diversity and balance, which easily causes the model to be over-fitted or perform poorly, leading to inconsistent personalities. In terms of development costs and efficiency, more computing resources and manpower need to be invested in training optimization and data processing, increasing costs, and complex solutions may also extend the development cycle and increase market competition pressure. In terms of user expectations, players require consistent personalities and rich and real interaction experiences, but it is difficult for NPCs to fit each player's role imagination, and it is not easy to balance universal expectations and personalized needs.

The future is full of prospects. With the development of AI technology, new model architectures or training methods are expected to enhance the ability to process long contexts and maintain character settings. For example, stronger memory mechanisms can help NPCs accurately remember key information, and combining reinforcement learning can optimize interactive behaviors. In terms of data, automatic annotation technology and crowdsourcing platforms can reduce costs and improve efficiency, and synthetic data generated by generative adversarial networks can expand the scale and diversity of training data, supporting the training of models with more stable and personality-consistent performance. Strengthened cooperation between developers and research institutions in the industry, exploration of universal solutions, and formulation of technical standards and best practices will promote technology dissemination and application, improve industry levels, and standardization can also improve players' consistent expectations for NPC experiences in different games, promoting the healthy development of the industry. After solving the problem of inconsistent personalities, games can innovate scenarios where NPCs can have complex emotional and psychological changes, with behaviors evolving in real-time with game events and player interactions, bringing richer and more immersive experiences. NPCs with different personalities can also have complex social relationships and interactions, build a more realistic and vivid game world and expanding the boundaries of gameplay and content.

4.2. Social level

How to ensure that generated game content conforms to moral and legal norms and avoid adverse impacts. When AI has autonomous consciousness, it may have its own goals and values, and if these goals are inconsistent with human interests, they may trigger unpredictable risks. AI algorithms may inadvertently copy or amplify biases in human society. If the training data itself has biases, AI systems may make unfair decisions. When AI has emotional understanding capabilities, it can better understand human needs and intentions, which can meet human desires for certain specific emotional support to a certain extent. However, emotional AI may change the structure of human society. People may over-rely on AI, leading to weak interpersonal relationships and even emotional addiction. These issues currently have no clear answers. False information easily misleads people's understanding and judgment of real events, leading to wrong cognition and damaging social trust and stability. When the public frequently comes into contact with AI-generated false information, they will doubt the authenticity of information, thereby reducing trust in the media, institutions, and even the entire society. AI may generate false information by integrating and analyzing a large amount of personal data, thereby leaking personal privacy. For example, in terms of reputation damage, false descriptions or slander of individuals in false information may damage personal reputation and bring mental stress and social interaction troubles to the parties involved.

5. Conclusion

This paper focuses on the application of large language models (LLMs) in the gaming field. It first elaborates on the research background and significance, reviews the application status, and introduces the definition, characteristics, working principles, and development history of LLMs. It focuses on analyzing their applications in games, taking driving non-player characters (NPCs) and dynamic plot generation as examples to illustrate that LLMs can improve the naturalness of NPC dialogues, realize dynamic decision-making, reduce development costs, and expand interactive capabilities, and mentions relevant experimental cases. At the same time, it points out the technical challenges faced in applications, such as interaction delays and inconsistent NPC personalities, as well as ethical and social risks such as the generation of harmful content and privacy leaks. The conclusion holds that LLMs bring a new way for NPC design in the gaming field, transforming them from passive execution to active and dynamic interaction. With technological optimization, LLMs will be more widely applied in the gaming industry, but it is necessary to balance technological innovation and ethics, and ensure the appropriateness of content through safety mechanisms to create personalized and immersive experiences for players.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Mann, B., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv: 2005.14165 1.3: 3.
[2] Park, J. S., et al. (2023). "Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th annual acm symposium on user interface software and technology.
[3] Akoury, N., Qian, Y., and Mohit, I. (2023). A framework for exploring player perceptions of llm-generated dialogue in commercial video games. Findings of the Association for Computational Linguistics: EMNLP 2023.
[4] Schick, T., et al. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems 36: 68539-68551.
[5] Weidinger, L., et al. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv: 2112.04359.
[6] Wang, Z., Masri, Y., Malarvizhi, S. A., et al. (2025). Optimizing context-based location extraction by tuning open-source LLMs with RAG. International Journal of Digital Earth, 18(1).
[7] Aman, S. S., Kone. T., N'guessan. G. B., et al. (2025). Learning to represent causality in recommender systems driven by large language models (LLMs). Discover Applied Sciences, 7(9): 960-960.
[8] Dennstädt, F., Windisch, P., Filchenko, I., et al. (2025). Consensus Finding Among LLMs to Retrieve Information About Oncological Trials. Studies in health technology and informatics, 329239-243.
[9] Golnari, P., Prantzalos, K., Upadhyaya, D., et al. (2025). Human in the Loop: Embedding Medical Expert Input in Large Language Models for Clinical Applications. Studies in health technology and informatics, 329658-662.
[10] Wu, G., Zheng, L., Xie, H., et al. (2025). Large Language Model Empowered Privacy-Protected Framework for PHI Annotation in Clinical Notes. Studies in health technology and informatics, 329876-880.