Optimization Strategies for Role-Playing Games Based on Large Language Models

Bowei Chen[a]
Department of Computer, North China Electric Power University, Baoding, Hebei Province, China

Keywords: Large Language Models, Roel-Playing Games, Optimization Strategy.

Abstract: With the booming development of the video game industry around the world, it has become an important industry that cannot be ignored. The evolution of video games and the close combination of computer technology have promoted the continuous advancement of game innovation. In recent years, with the rapid development of large language models (LLMs), its application potential in various fields has gradually emerged. Especially in the game industry, the combination of LLMs and game mechanisms has triggered extensive discussions. In this context, this paper selects role-playing games (RPGs) as the research object to explore the application of LLMs in this type of game and the possible impact. RPGs emphasize the interaction and emotional experience between players and virtual characters, and they show a high degree of complexity and innovation needs in terms of storylines, character development, and dialogue systems. This paper aims to investigate the practical application of LLMs in RPGs, analyze their potential in improving game experience, enhancing character interaction, and enriching plot development, and deeply explore the advantages of their application, in order to provide a reference for research and practice in this field.

1 INTRODUCTION

The symbiotic relationship between artificial intelligence and gaming traces its origins to the mid-20th century, with seminal developments occurring during computing's formative era. As early as 1952, pioneering research initiatives in heuristic programming were launched, culminating in 1956 with the emergence of the first self-learning algorithm capable of mastering checkers - a technological milestone that effectively constituted the embryonic phase of intelligent gaming systems. Throughout the subsequent three decades, researchers persistently endeavored to integrate artificial intelligence (AI) technologies into game design architectures. By 1990, AI-assisted game design began to emerge, though constrained by the technological limitations of the era, with applications primarily restricted to rudimentary game components and fundamental operational systems. The 21st century has witnessed accelerated advancements in AI driving progressively sophisticated implementations across gaming ecosystems, with contemporary applications now permeating core gameplay mechanics, player modeling systems, and real-time environmental generation pipelines. By 2020, AI had been systematically implemented in procedural content generation (PCG) for gaming applications, encompassing level design optimization, character animation automation, and narrative scripting. This technological integration substantially enhanced developmental efficiency across the industry. The emergence of LLMs particularly demonstrated transformative potential in dynamic dialogue generation and contextual quest synthesis, thereby introducing novel possibilities for innovative game design paradigms. Xu et al. conducted a comprehensive investigation on the application of LLMs in gaming agents, examining their functional implementations in digital games. The study confirmed LLMs' substantial potential within gaming contexts while identifying significant developmental opportunities requiring further exploration(Xu et al., 2024). At the 2024 Game Developers Conference (GDC), Ubisoft unveiled its experimental AI framework NEO NPC, designed to implement LLM-powered conversational agents within gaming environments. Ubisoft's technical specifications

[a] https://orcid.org/0009-0002-2623-3745

Chen, B.
Optimization Strategies for Role-Playing Games Based on Large Language Models.
DOI: 10.5220/0013703100004670

In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 632-637
ISBN: 978-989-758-765-8

indicate this architecture enables dynamic dialogue generation constrained by predefined narrative parameters (backstory, setting, character personas) while maintaining strict alignment with designers' creative directives through parameterized response filtering mechanisms. In essence, game developers seek to leverage LLMs to bring non-player characters (NPCs) to life. The integration of the LLM-integrated system's deployment marks a paradigm shift in artificial intelligence applications within interactive entertainment domains and reveals substantial unexplored potential in emergent gameplay design.

Since there are many types of games, and in each type, LLMs play different roles and undertake various tasks, this paper collects research papers related to RPGs to understand the roles that LLMs can play in RPGs. It explores how LLMs can assist in game development, optimize players' experiences, and whether there is a potential for better integration between LLMs and RPGs in the future, resulting in even better outcomes as LLMs continue to evolve.

2 RPG OVERVIEW

RPG is a type of interactive video game in which players assume the roles of virtual characters, typically tasked with completing missions, interacting with other characters, and exploring the game world. The core features of RPGs lie in character development and storyline progression. Players continuously enhance their character's abilities, equipment, and skills by engaging in various activities within the game. In this process, the growth of the character is closely tied to the progression of the story, providing players with an immersive gaming experience.

2.1 An Engaging Storyline

First is the plot with depth. A great RPG requires a complex and engaging main storyline that draws players in. It often includes elements such as character development, emotional conflicts, and moral choices, allowing players to not only "play" the game but also "experience" the story.

The second is diverse endings. Based on the player's choices, the game's storyline may have multiple endings, which increases replay value. The player's decisions have a tangible impact on the world, sparking an interest in exploring different endings.

2.2 Vivid Character Development

First is rich character design. Each character has a distinct personality, backstory, and motivation. A great RPG reveals the characters' pasts, goals, and emotions through interactions with NPCs, allowing players to become more immersed in the game.

The second is the character development system. In addition to numerical improvements, characters should also develop in terms of skills, personality, and even appearance. The player's choices can shape the character's development path and even influence the character's relationships with other individuals.

2.3 The Degree of Openness in the Game World and the Diversity of Choice

First is the open world. Many great RPGs offer an open world where players can freely explore, not limited to completing tasks. The world is filled with rich details and secrets waiting to be discovered, enhancing the joy of exploration.

The second is the consequences of player choices. Players' decisions in the game should have substantial consequences, influencing both the storyline and the world's state. For example, a player's choices could lead to shifts in political situations in certain regions or alter relationships with NPCs.

2.1 Detailed World-building and Environment Design

First is the immersive world. A successful RPG needs to design a virtual world with a rich historical background and internal logic. The design of the environment, architecture, culture, language, and other aspects should make players feel that this world is alive.

The second is the interactive world. Not just the map and quests, every element in the world should have its purpose, with players able to interact with the environment, NPCs, and other characters in various ways. For example, through communication, trade, combat, or quests, players should feel their impact on the world.

3 LLMS OVERVIEW

LLMs are the third stage in the development of language models. Based on the Transformer architecture, these models learn the grammar, semantics and contextual information of language by

training on large amounts of text data. In June 2018, OpenAI released the GPT-1 model, marking a new era for language models, namely the era of LLMs. Released in June 2020, GPT-3 is regarded as a major breakthrough in large language models, demonstrating significant advancements across multiple natural language processing tasks, particularly in the quality and diversity of generated text. Subsequent iterations like GPT-4, launched in 2023, further enhanced these capabilities with improved contextual reasoning, reduced error rates, and expanded multimodal functionalities. Table 1 provides a comparison of parameter counts across different models.

Table1: Parameter Count Comparison.

Models	Release Time	Developers	Parameter size/10⁸	Sample size/10⁹
GPT-1	2018	OpenAI	1.17	10
BERT	2018	Google	3.40	34
GPT-2	2019	OpenAI	15.00	100
Fairseq	2020	Meta	130.00	—
GPT-3	2020	OpenAI	1750.00	4990
GLaM	2021	Google	1200.00	16000
LaMDA	2022	Google	1370.00	15600
GPT-4	2023	OpenAI	1750-2800	—

In addition, Rudolph et al. analyzed the ChatGPT, Google Bard (now Gemini), Microsoft Bing Chat (now Copilot), and Baidu Ernie—in educational applications, analyzing their capabilities in knowledge delivery, interactive learning support, and pedagogical effectiveness. In the field of education, materials such as textbooks, exercises, solutions, and various reports, such as summaries of each student's learning progress, often require extensive creative work. Since these are all text-related tasks, the education field is highly compatible with text generation (Rudolph, Tan, & Tan, 2023). In addition to requiring text generation capabilities, the education field demands that educational content be accurate and thematically relevant. When users ask questions or request exercise solutions, the model must effectively understand the query's content and provide appropriate answers. Furthermore, the generated text must not only be clearly expressed but also logically structured, which necessitates robust comprehension and data processing capabilities within the model itself. Therefore, the model's performance in this study's testing phase can serve as key evidence for evaluating the relative strengths of different models. Table 2 presents the performance comparison results.

Table 2: Performance Comparison.

index	GPT-4	Bing Chat	Bard	Ernie
Generate accuracy	High	Medium	Low	Medium（good in Chinese）
Timeliness	None	High	Low	None
Academic ability	Excellent	Average	Poor	Strong academic ability in Chinese
User experience	Excellent	Average	Average	Chinese-friendly
Problems	Hallucination, logical errors	Prejudice, low academic accuracy	Incoherent responses	Content restrictions due to censorship.

4 THE APPLICATION OF LLMS IN RPGS

4.1 Dynamic Dialogue Generation

In RPGs, LLMs are utilized to generate dynamic dialogues for NPCs, imbuing them with character depth and contextual responsiveness that enhances narrative immersion. In this way, NPCs are no longer limited to a single reaction to player behavior and dialogue, but can instead exhibit more diverse and personalized conversations based on different contexts, player choices, and game progression. This enables more vivid NPC characterization, enhancing both the immersion and interactivity of the game, which significantly boosts the player's overall gaming

experience. Nananukul et al. proposed a framework for enhancing NPC dialogue with a more narrative-driven approach using large language models, focusing on games like Final Fantasy VII and Pokémon. The goal is to enable NPCs to respond with personality-appropriate reactions and tone in specific scenarios (Nananukul & Wongkamjan, 2024). By gathering character details, situational descriptions, skills, and personality traits, a knowledge graph is created to structure the data of game characters and scenarios. Tailored prompt templates are then developed for different games and characters, providing character personality, describing specific contexts, and outlining the dialogue goals and style. The relevant information from the knowledge graph is incorporated into the prompts, improving the contextual relevance of the model's generated dialogues. In Final Fantasy VII, battle scenario dialogues were tested, requiring characters to generate dynamic responses based on the battle state, such as enemy health or skill usage. In Pokémon, NPC dialogues were tested, with Red being given different personalities (e.g., talkative, confident) to evaluate the diversity of the generated responses. The results indicate that the quality of dialogue generation is relatively high, with GPT-4 being able to understand the character's behavior and reactions in specific contexts and generate reasonable and natural dialogues. It performs well in expressing simple personalities, such as talkative or shy, with accuracy. However, its ability to express more complex personalities, such as mature or introverted, is limited. The generated content may seem overly positive or superficial, particularly in expressing nuanced personalities like maturity or introspection. Additionally, when generating repetitive dialogues or overly positive tones, it does not always align with the character's established traits, such as Cloud's cold personality.

Huang discussed and analyzed the performance of GPT-4 and GPT-3.5 Turbo. He used GPT-4 for NPC dialogue generation in RPG games, leveraging its large number of parameters and enhanced context window to generate more natural, coherent, and highly contextually relevant text (Huang, 2024). By providing the model with the game's background and NPC character settings, and capturing dynamic information, NPCs can generate dialogue that better aligns with the NPC's background and the game's current state.

Chubar designed an RPG game and used GPT to generate procedural content, creating a game with rich narratives and dynamic content. The study tested the performance of the two models by comparing the number of worlds and total units generated under two different diversity conditions (Chubar, 2024). With a temperature of 0.6, GPT-3.5 Turbo generated 15 worlds, totaling 489 units, with an average of approximately 32.6 units per world. In comparison, GPT-4 generated 5 worlds, totaling 210 units, with an average of 42 units per world, showing a higher unit density. At a temperature of 1.2, the generated worlds were the same as those at 0.6, but GPT-3.5 Turbo produced fewer total units (437), suggesting that higher temperature may lead to greater content diversity at the cost of unit density. Meanwhile, GPT-4 generated fewer total units (195), with a slight decrease in the average number of units per world (39). GPT-4's world unit density was higher at a lower temperature (0.6) (42 units per world), indicating that its generated content is superior to GPT-3.5 Turbo in terms of control and detail richness. As the temperature increased from 0.6 to 1.2, the number and density of units slightly decreased, but remained relatively high, indicating that GPT-4 can maintain a high level of quality while increasing generation diversity. In contrast, the content generated by GPT-3.5 Turbo showed a more significant decline.

4.2 Task Description

LLMs can also be used for task description generation. Due to their exceptional text generation capabilities, using LLMs for task generation in games can significantly reduce the workload of developers. By simply setting the game background and relevant content for the LLMs, high-quality task descriptions can be generated. Värtinen et al. explored the application of GPT-2 and GPT-3 in generating task descriptions for Role-Playing Games (RPGs). As players' demand for game content continues to rise, developers face the challenge of manually designing tasks. The paper proposes an improved version of GPT-2 called Quest-GPT-2, designed to automatically generate RPG task descriptions (Värtinen, Hämäläinen, & Guckelsberger, 2022). The study evaluated the model's performance through calculated metrics and a large-scale user study involving 349 players. Task data was extracted from six classic RPG games—Baldur's Gate 1 & 2, Oblivion & Skyrim, Torchlight II, and Minecraft. The GPT-2 fine-tuned version, Quest-GPT-2, and GPT-3 were used to generate 500 task descriptions. These descriptions were then rated by the 349 RPG players based on three criteria: "Does the task description match the game task?", "Is the text coherent and logical?", and "Does it match the RPG task style?". The results showed that the task descriptions

generated by the large models were mostly acceptable, with some even outperforming those written by humans. However, issues such as repetition, character confusion, and logical errors were present, while the task descriptions generated by GPT-3 were more natural and logically clear.

Koomen collected thousands of task data from four multiplayer role-playing games (World of Warcraft, The Lord of the Rings Online, Neverwinter Nights, The Elder Scrolls Online), which included task objectives, context information, and task descriptions. Using GPT-J-6B, he fine-tuned the model on the task data from each game, generating four different models to generate tasks for each game. The quality of the generated tasks was then evaluated through user surveys and focus group interviews, comparing them to the original task descriptions. The results showed that the generated tasks contained more elements of surprise and mystery, but also revealed issues such as logical inconsistencies and repetition (Koomen, 2023). Although the generated task descriptions showed some improvement over those written by humans, challenges remain in generating complex backgrounds and task motivations. For example, the background motivation of some tasks may not be fully captured, leading to missing or inconsistent motivations in certain tasks.

4.3 Dungeon Master in Some Contexts

LLMs can also serve as game masters (GM) for tabletop role-playing games (TRPGs), such as in Dungeons & Dragons, or as an auxiliary system. In TRPGs, the GM plays a crucial role, ensuring the game runs smoothly and being responsible for creating and describing the world, environment, backstory, and plot within the game.

Sakellaridis explored the feasibility of using LLMs as Dungeon Masters (DM) for Dungeons & Dragons (D&D) (Sakellaridis, 2024). The study was based on text data from the popular Critical Role D&D livestream series, fine-tuning ChatGPT to adapt it for the DM role. The reference adventure module "The Sunless Citadel" was chosen to ensure the LLM agent adhered to D&D rules. A text-based TRPG environment was created, where players input text commands, and the LLM agent responds. Finally, through player feedback on both human DMs and LLMs, game dialogue data, and statistical analysis, the study compared the strengths and weaknesses of using an LLM as a DM. The results showed that LLMs excel at creating immersive worlds but perform less well than human DMs in terms of storyline flow. In some scenarios, LLMs tended to force the plot forward, reducing player freedom and struggled to fully adapt to nonlinear story branches, leading to issues with narrative continuity.

Zhu et al. discussed the potential of using LLMs as an assistant system for DMs, helping them generate scenes, summarize monster information, extract key details, and provide creative suggestions. The study explored how LLMs can assist DMs in game narration, enhancing the immersion of the game (Zhu et al., 2023). This research invited 71 players and DMs to participate in an experiment, testing the game on a Discord-based D&D server. Quantitative data analysis and user interviews were used to evaluate the quality of the content generated by the LLM, the player experience, and the role of AI in the game. The study concluded that LLMs can assist the DM, but they are not yet able to independently guide the game.

4.4 Assist in Game Development

Another important use of LLMs is to assist game developers in game development. However, research in this area is still relatively limited. The main reason is that the accuracy of LLMs in generating code needs improvement. Their understanding of the global context of code is still limited, especially when dealing with cross-file references, complex architectures, or large projects, which can lead to missing key information and thus affect the correctness and accuracy of the code. Additionally, LLMs can only generate code statically and lack the code execution and error-checking capabilities of an IDE or interpreter, making it difficult to detect potential bugs. Furthermore, LLMs also face issues such as hallucinations and redundancy, which further limit their use in game development applications.

Chen et al. explored the potential of using LLMs for automated game development to improve game development efficiency. The study introduced a model called GameGPT, which can play a role in various stages of game development, including planning, task classification, code generation, and task execution (Chen et al., 2023). By collaborating with small expert models, LLMs can help reduce redundancy and hallucination issues. Additionally, breaking the code into smaller segments for GameGPT to generate can improve code accuracy. The study incorporated a review role at each stage to test the reliability of the model. The results showed that it not only optimized the generation process but also enhanced the system's reliability and scalability through a feedback loop.

In addition to assisting with programming, LLMs can also aid game developers in the creation of game worlds. Nasir et al. studied the use of LLMs to transform stories into playable game worlds. They proposed a system called Word2World, which extracts information about characters, settings, objectives, and interactive elements from a story and generates the game world's environment, layout, plot, and interactive content. The system can also be continuously optimized through algorithmic checks and feedback loops (Nasir, James, & Togelius, 2024). The experiment showed that the generated game worlds were playable in 90% of cases, but LLMs faced certain challenges in creating large-scale 3D games or animation effects.

5 CONCLUSIONS

By investigating the development direction and current situation of LLMs in RPG games, this paper summarizes four directions that LLMs can play: generating dynamic dialogue for NPCS, generating more narrative mission descriptions, acting as a host for games, and assisting game development. Although progress has been made in these directions, the LLM's data compression and illusion problems caused by inconsistencies, as well as the redundancy problems caused by the model itself or its search strategy, hinder the LLMs from performing better in RPG games.

With the development of the current LLMs model, there will be 70 models in the Chatbot Arena in 2024 that perform better than the GPT-4 in 2023, and the deepseek V3 model will be released in early 2025. With the update of algorithms and the upgrade of hardware, it can be predicted that LLMs will certainly have rapid development in the future, and various models will definitely be upgraded in the direction of more intelligence. Under such conditions, LLMs can play a very important role in RPG games, and even in the whole electronic games, and have an important role in the development of the game industry.

REFERENCES

Chen, D., Wang, H., Huo, Y., Li, Y., & Zhang, H. (2023). Gamegpt: Multi-agent collaborative framework for game development. arXiv preprint arXiv:2310.08067.

Chubar, A. (2024). Generating game content with generative language models.

Huang, J. (2024). Generating dynamic and lifelike NPC dialogs in role-playing games using large language model.

Koomen, S. B. (2023). Text generation for quests in multiplayer role-playing video games (Master's thesis, University of Twente).

Nananukul, N., & Wongkamjan, W. (2024). What if Red can talk? Dynamic dialogue generation using large language models. arXiv preprint arXiv:2407.20382.

Nasir, M. U., James, S., & Togelius, J. (2024). Word2World: Generating Stories and Worlds through Large Language Models. arXiv preprint arXiv:2405.06686.

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. Journal of Applied Learning and Teaching, 6(1), 364-389.

Sakellaridis, P. (2024). Exploring the Potential of LLM-based Agents as Dungeon Masters in Tabletop Role-playing Games (Master's thesis).

Värtinen, S., Hämäläinen, P., & Guckelsberger, C. (2022). Generating role-playing game quests with GPT language models. IEEE Transactions on Games, 16(1), 127-139.

Xu, X., Wang, Y., Xu, C., Ding, Z., Jiang, J., Ding, Z., & Karlsson, B. F. (2024). A survey on game playing agents and large models: Methods, applications, and challenges. arXiv preprint arXiv:2403.10249.

Zhu, A., Martin, L., Head, A., & Callison-Burch, C. (2023, October). CALYPSO: LLMs as Dungeon Master's Assistants. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 19, No. 1, pp. 380-390).