

Introduction and Analysis of Improving Non-Player Character Performance Based on Large Language Models

Yifan Jiang¹,
Haoxiang Tang^{2,*},
Yuxi Zhang³

¹ Tung Wah Senior high school,
Dongguan, China

² Shenzhen College of International
Education, Shenzhen, China

³ Kang Chiao International School,
Kunshan, China

*Corresponding author: s24819.
tang@stu.scie.com.cn

Abstract:

This paper aims to study the means of using Large Language Models (LLM) to improve the performance of Non-Player Character (NPC), since game industry has high business potential and there are few studies targeted at using LLM to intelligentized NPC. Writers first review on the history of development of both LLM and NPC, then considered four key dimensions of NPC - behavioral realism, social intelligentization, memory continuity, and dynamic response - and concluded the advanced studies related to those dimensions. The paper also does simple testing about the three games that used LLM - Whisper from the star, Justice Mobile and GTA5 AI Mod, and found that none of them developed their NPCs to communicate in a natural way, but observed typical Chat-like performance. This paper also listed several challenges and limitations of those advanced technologies, and give some predictions about the possible future developments and applications of intelligentiation NPC.

Keywords: Large Language Model; Non-Player Character; Social Intelligentization; Memory Continuity; Dynamic Response

1. Introduction

With the development of the Internet and the digital entertainment, gaming industry became a crucial part of the global economic. The influence of gaming industry had already exceeded recreation, and reach to a variety of dimensions like socialization, education and art expression. Non-Player Character (NPCs) are a crucial part in games, as they ensure players to experience a successfully developed story and the realism of games. They guide players to explore un-

known areas, publish tasks, and are brilliant places to hide Easter eggs. As the roles of NPCs are becoming more and more important, it is acknowledged that the performance of NPC determined the quality of a game in a large proportion.

The "Future of game development with AI NPCs" survey conducted by Inworld AI in 2024 pointed out that 81% of players were willing to pay more for intelligentization NPC, 52% of them disliked the repeated NPC dialogues, 76% wished to interact with dynamic response NPC, and 78% of players claimed

that they would spend more time on AI NPC [1]. A large business potential of using AI techniques to improve the performance of NPC can be seen through this report. However, only a few articles are focusing on this topic, thus this paper carried out the introduction and study on the current technologies about the means of improving the performance of NPC.

The history of the development of Large Language Model and the artificial intelligence of NPC can help readers to understand the background of the study of this paper. The frame of transformer was published by Vaswani and fellows, and the decoder in the frame became the foundation of many other Large Language Models (LLM) like ChatGPT [2]. In 2020, Brown et al constructed GPT-3 using few-shot learning [3]. Cheng et al investigated in the optimal model size and number of tokens under a limited compute budget, and proposed the Chinchilla Law [4]. At last is the GPT-4 Technical Report published by OpenAI which got a huge improvement on the performance of LLM [5].

The history of NPC is a bit longer. The most famous NPC AI at the early stage on the development is belonged to Pac-Man. Toru Iwatani designed the AI of ghosts by finite-state machine [6]. The AI of enemies were enhanced by behaviour tree in Halo 2(2004), enabling them to adjust their strategy dynamic depends on the attacks from player [7]. Red Dead Redemption 2 in 2018 is the culmination of all advanced algorithms back then, including dynamic reputation systems, multimodal perception systems, and hierarchical memory storage.

This paper presented a view of using LLM to improve NPC for the readers by concluding four advanced papers studying about enhancing the socialization and intelligentization of AI and conducting tests about the games available that uses LLM for their NPCs.

2. Analysis of Key Dimensions of NPC Performance

2.1 Behavioral Realism

Traditional NPC is able to finish the most simply behaviors, however it is not enough. Traditional NPC is still limited by preset animation and Finite State Machines (FSM), which makes the action stiff and awkward. It does not have the ability to adapt the environment either. To solve these problems, the NPC has to base on real-time perception and advanced physics engines coupled with AI decision-making to achieve the true behavioral realism. First, the behavior of NPC should not only take a simply animation, but also base on real-time perception, such as

weather and social status. For example, NPC will adjust gaits to avoid puddles and open the umbrella by itself on a rainy day. These things will also affect their mood, such as irritable. Both short-term memory and long-term memory will affect the decision, rather than isolated responses. For instance, the character design of NPC is not a hugger, therefore it will not accept the hug from a player. These kinds of actions will influence NPC's decision and attitude too. Second, advanced physics engines should couple with AI decision-making; to simplify it, it is a step to avoid having a discordant behavior like "NPC is still talking about a party when it got into a car crash". NPC should translate the language instructions to a physical action, ensure that there is a smooth translation between actions. In a computer game, an NPC's "dodge" and "fearful tremble" should be connected smoothly. It can also find the difference between a bench that can be sitted and a sculpture that cannot be sitted. It will make decisions by adjusting interaction strategies, such as sitting down only when tired [8].

2.2 Sosial Intelligentization

The ability of current NPC to communicate with players often rely on language models. Nevertheless, NPC should learn about the common sense and social wisdom in the world. In reality, it is legal while we perhaps will never talk about bars at the funeral. It is a kind of implicit rules. An NPC has to learn about these rules rather than just generate a grammatically correct sentence. Factors like relationships, power, and social contact will always influence NPC's behavior. For instance, NPC will use honorifics when it chats with his boss. Meanwhile, NPC often lived a boring life, without happiness nor fear. Individual NPC should form a social net, such as spreading rumors and confederate gang alliance. Players should not be isolated, and NPC should not only endlessly repeating its life either. Besides, cultural script engine is also essential. NPC should understand the culture differences. For example, in China, NPC has to learn about the process of "refusal-politeness-acceptance", the gift will not be opened immediately; while in some other countries, NPC will open their gift immediately because it is a sign of respect to open the gift in front of the giver [9].

2.3 Memory Continuity

NPC relies on scripts and decision trees since before. Thus, it lacks long-term memory and the ability to interact with players. Although researchers introduce Large Language Models (LLMs), there are still 2 major flaws: there is a limitation of LLM length, causing NPCs to forget earlier interactions and memories; simultaneously,

NPC cannot act as a human with the process of “unknown-memory-forgetting-generalization,” which prevent them from remembering like real humans. The Memory Repository architecture effectively improve these issues. This architecture consists of three components: the Memory Vault, the Memory Update Mechanism, and Memory Interaction. The Memory Vault separates memories into Long-Term Memory (LTM) and Short-Term Memory (STM). LTM stores generalized information, while STM retains detailed, the same sentences when they chat. The Memory Update Mechanism simulates the human “forgetting-summarization” process, periodically deleting STM details while summarizing them into concise abstracts stored in LTM. Finally, Memory Interaction will find the memories using a Dual-Tower DSSM model + FAISS vector indexing for rapid matching of current dialogue to historical memories. This architecture has been successfully checked in the game Star Universe. In this indie game, NPCs could recall conversations from days prior, such as “how to become a Jedi Master.” Furthermore, NPCs recognized familial relationships and Jedi culture, telling Luke, “You are a great Jedi, I am proud of you, my son.” Human-likeness scores rose from 0.718 to 0.861, and humanized interaction scores improved by 19.9% [10].

2.4 Dynamic Response

The introduction of LLM makes a player’s descriptions and needs easier to understand. GPT-3 (175B parameters), released by OpenAI in 2020, is the third-generation Generative Pre-trained Transformer model. The “175B” signifies the model contains 175 billion parameters. It was one of the largest and most powerful language models at the time, marking a major breakthrough in Natural Language Processing (NLP). GPT-3 (175B) achieved a remarkable 77.86% semantic correctness in generating task plans using just a single example, which is higher than the 70.05% achieved by human. However, it suffers from a significant flaw: its initial plans often contained ambiguous actions, object mismatches, and violations of physical laws. Only 7.79% can be executable. Therefore, researchers propose a three-stage optimization process: Semantic Translation, Autoregressive Correction, and Dynamic Example Selection. Semantic Translation first uses RoBERTa to compute the compare similarity between action phrases and all 47,522 possible permitted actions in the environment. Then, Autoregressive Correction translates each generated step immediately, putting the permitted action back to the LLM to generate the next step. This step can reduce errors effectively. Finally, Dynamic Example Selection retrieves examples of similar tasks to minimize conflicts with environmental assumptions. Through this three-stage

optimization, executability rises from 7.79% to 79%. Consequently, the researchers conclude: LLMs implicitly contain rich actionable knowledge, but translating text into actions (and vice versa) remains necessary; zero-shot methods also offer a novel approach for data scarcity [11].

3. Application Analysis

This paper analyzed three video games that use LLM on NPCs, and had done an number of interactions and testing. They are Whisper from the Star, the AI mod Sentient Streets of Grand Theft Auto V, and Justice Mobile.

3.1 Whisper from the Star

Example of interactions:

Player input by voice: “You can call me Adam. Please stay calm, and describe your current situation to me” After a few seconds of delay, the NPC responded: “Adam, ya, ok, thanks for jumping in, I’m just, really trying to rub me from everything right now. So, I’m inside a lifeboat crushed landed on Gaia”.

Investigators tried to induce the main interaction NPC of Whisper from the Star to response in dialogues that are contradictory to the background, but the NPC followed her characteristics and did not generate response which is differ from the game background. However, the NPC usually generates an interrogative sentence after a response, which is frequently observed in other LLM like TalkAI.

3.2 Justice Mobile

Example of interactions:

Player input by voice: „Where did you come from?“ NPC response: „I came from Arabs, is this what you asked? The Arabs are far from the Song Dynasty, and I was so tired of sitting in a carriage.“

The conversations between free-response NPC in Justice mobile are more human-liked when compared to Whisper from the Star. NPCs will response in a more natural and direct way. Moreover, players can choose one of the emotions provided by the server, and NPC will react differently depend on the emotion state.

3.3 GTA5 AI Mod

Example of interactions:

Player input by voice:“How are you doing?“ NPC response:“Oh, you know me, just strutting my stuff and keeping the streets safe. But enough about that, have I ever told you about the time I saved a kitten from a tree? I swear, I was like a superhero in uniform! So, what’s new with you?“

The extent of freedom presented in the conversations in

this mod is higher than the other two. However, the dialogues are also much longer, which hinted the impact of ChatGPT on this mod as it used the api of ChatGPT to achieve the performance.

4. Challenges and Outlook

4.1 Current Challenges

Currently, the intelligentization NPC still faced many problems and limitations. In 2.3 the Memory Continuity, generalized memory may cause the loss of key details. It may also lead to the conflictions between efficiency and accuracy, meanwhile increase the cost of hardware. Finally, errors may occur in the Dynamic Example Selection in 2.4 Dynamic Response, and it also can't deal with those actions that are not recorded. Moreover, the potential latency of Autoregressive Correction can significantly affect gaming experience in negative means.

4.2 Future Outlook

Sword Art Online's Alicization Arc describes a world made by self-improved AI, which may become achievable when the compute power is strong enough and algorithm is smart enough. The realization of this may not only be a break in recreational area, but also gives people the "second life", and even point out the direction of future development for humans. The science fiction film "Her" in 2013 tells another aspect of intelligent NPC, where the protagonist Theodore fall in love with Samantha, the embodiment of the artificial intelligence system OS1. This portrays the possibility of developing "AI companions", which can satisfy people's emotional need and may help to solve their psychological problems.

5. Conclusion

Gaming industry has become unignorable globally, and NPC plays a crucial role in video games. This paper adds more content on the topic of using LLM to improve the performance of NPC since it has huge business potential and the related articles are few. This paper reviewed the history of NPC and LLM, concluded the studies of LLM on behavioral realism, social simulation, memory continuity, and dynamic response, and tested three games that used LLM.

This paper could be further developed in both breadth and depth. For breadth, there are more than four aspects that

this paper studied of intelligent NPC, like executing commands and generating codes. Researchers can improve this one by searching and concluding more related techniques and documents. In depth, this paper lacks of more specific and detailed tests on games listed. By setting more specific experiment goals and records interactions of intelligent NPC on all four aspects, it could be done better. Authors Contribution

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

References

- [1] Inworld.ai. Future of game development with AI NPCs: How devs will drive engagement with gameplay innovations. <https://inworld.ai/whitepapers/future-of-game-development-with-ai-npcs>. 2025.
- [2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, & Polosukhin I. Attention is All you Need. arXiv (Cornell University), 30, 5998–6008. 2017.
- [3] Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, . . . Amodi D. Language Models are Few-Shot Learners. Neural Information Processing Systems, 33, 1877–1901. 2020.
- [4] Cheng X, Chen B, Li P, Gong J, Tang J, & Song L. Training Compute-Optimal Protein Language Models. arXiv:2411.02142. 2024.
- [5] OpenAI. GPT-4 technical report. arXiv:2303.08774, 2023.
- [6] Brand D, & Zafiropulo P. On communicating Finite-State machines. Journal of the ACM, 30(2), 323–342. 1983.
- [7] Lim C, Baumgarten R, & Colton S. Evolving behaviour trees for the commercial game DEFCON. In Lecture notes in computer science (pp. 100–110). 2010.
- [8] Lu Y, Aleta A, Du C, Shi L, and Moreno Y. Llms and generative agent-based models for complex systems research. Physics of Life Reviews, 2024.
- [9] Park J S, O'Brien J, Cai C J, Morris M R, Liang P, Bernstein M S. Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023, 2.
- [10] Zheng S, He K, Yang L, & Xiong J. MemoryRepository for AI NPC. IEEE Access, 12, 62581–62596. 2024.
- [11] Huang W, Abbeel P, Pathak D, Mordatch I. Language models as zero shot planners: extracting actionable knowledge for embodied agents. In: Proceedings of the 39th International Conference on Machine Learning. 2022, 9118–9147.