A Brief Survey of Large Language Model based Agents in Game Systems

Peixian Ma (stuid: 50011853) ✉*[1]

[1]The University of Science and Technology, Information Hub, DSA Thrust

ABSTRACT

Introduction

Large language model based agents(LLMs-based agents) have shown wonderful performance in many tasks, which leverage the powerful processing capacity of large language models for training and iteration[] By reasonably designing the framework of large language model based agents, researchers find that the ability of LLMs-based agents in using environmental information and human-like thinking could be effectively enhanced.[1; 2; 3]. Because of these abilities, LLMs-based agents can be applied in less constrained and more complex environments, such as game systems. Researchers are able to construct autonomous game systems based on large language model based agents, and incorporate more exploration and attractive elements[4].

In recent years, large language models have become a crucial component in the field of artificial intelligence, and continue to develop at an impressive rate. With the development of prompt learning and alignment techniques, by scaling the both size of transformer based models[5] and dataset in a reasonable way, researchers explored that large language models can show great performance and capability in many tasks. In the existing research, many base models have shown powerful potential capabilities, including GPT-4[6], LLaMA[7], GLM[8; 9]. These models have been successfully applied in various applications such as chatbots, automatic writing, sentiment analysis, text classification, and question answering systems[10; 11]. Through these techniques, LLMs play a core role in building specific agents for automatic system, by supporting the data analysis and decision making. One of the important directions is to construct agents for the game systems[12; 13; 14; 15]. Researcher try to equip these agents with multiple human-like abilities, such as memory, planning, action, and try to prompt them to complete common tasks in game systems, like dialogue interaction, story telling or battle[14].

In this paper, we firstly provide a brief review of large language model based agents in game systems in recent years, and demonstrate their specific points. Then, we describe some evaluation metrics for the performance of LLMs-based agents in game systems. Finally, we give a conclusion of these agents and discuss the future directions.

Recent Works

With the development of deep learning methods and hardware techniques in recent years, large language models have attracted the attention from all over the world. A series of large language models, represented by GPT[6], achieved unprecedented results on hundreds of fundamental task tests and demonstrated greater generalization capabilities than all previous deep learning models[16; 17]. Meanwhile, In-context Learning(ICL)[18], Chain of Thought(CoT)[19], Lora[20] and other techniques or tricks enable researchers to build LLMs-based agents quickly and conveniently with fewer resources. Therefore, LLM has been applied in many scenarios and get favourable feedback[4]. Likewise, LLMs-based Agents play an important role in game system field, which are used to generate more immersive and interactive experiences for human players, providing a richer and deeper game experience. Researchers have already developed a series of prototypes and demos for some specific game scenarios, which have epoch-making significance.

Currently, the application of Multiple LLMs-based agents(Multi-Agents) in game systems is a novel research direction, which provides the possibility of Agent-based automatic games. Generative Agents is an interactive human behavior simulation framework based on LLM proposed by Stanford University[14]. It contains some human-like modules, such as observation, memory retrieval, planning, and response, which are used to achieve realistic human behaviors. Generative Agents runs in a small town environment with 2D pixel image rendering. 25 Agents are initialized in this environment and to achieve interactive dialogue, action, and other functions. Xu et al. explored the issue of using multiple LLMs based proxies in Werewolf games[21]. Researchers guide individuals to use past communication and experience retrieval and reflection for self-improvement and generate a new round of interactive behavior without adjusting LLM parameters. In

*Co-corresponding Author, pma929@connect.hkust-gz.edu.cn

the study, agents based on LLMs began to exhibit clear strategic thinking and behavior, making the game exhibit effects similar to human player participation.Inspired by human group dynamics, AgentVerse has been proposed to collaborate and dynamically adjust the composition of multiple agents[22]. It can effectively deploy groups of multi-agents to exhibit performance that exceeds that of a single agent. In addition, AgentVerse can effectively alleviate the possible negative behaviors of multi-agent groups and improve their collaboration potential.

In addition, researchers will also develop LLMs-based agent for specific gaming environments to explore the understanding ability of large language models in gaming systems. DEPS[23] and GITM[24] explores and clarifies the challenges faced by LLMs-based agents in Minecraft, an open gaming world environment, and summarize some experience for design and use of LLMs-based agents as follows: Firstly, due to the long-term nature of the task, executing planned actions in an open world environment requires accurate multi-step reasoning; Secondly, due to the fact that planners may not have taken into account the difficulty of the current agent executing tasks when formulating plans, the resulting plans may be inefficient or even infeasible. Researchers developed a DEPS interactive planning method based on the aforementioned difficulties and assisted LLMs-based agents in effectively completing dozens of tasks in the Minecraft environment. Also, Voyager is the first LLMs-based embedded agent developed for Minecraft games, constantly exploring the Minecraft world without human intervention, acquiring various skills, and making new discoveries[25]. Voyager consists of three key components: an automated course that maximizes exploration, an expandable skill library for storing and retrieving complex behaviors, and a self validation prompt iteration mechanism that combines environmental feedback, execution errors, and program improvement. Voyager interacts with GPT-4 through black box queries, bypassing the need for fine-tuning model parameters. The skills developed by Voyager are scalable, interpretable, and combinatorial in terms of time, which quickly enhances the agent's capabilities and alleviates catastrophic forgetting. Voyager exhibits significantly better learning abilities than ordinary Minecraft players and demonstrates extraordinary proficiency.

In the development process of LLMs-based agents game systems, the shortcomings of large language models would also affect the performance of LLMs-based agents. For example, hallucination can easily cause large language models to produce incomprehensible answers or fail to understand the input, resulting in LLMs-based agents losing their original functions. GameGPT is also a multi-agent collaboration framework for LLMs-based agents in game systems[26]. However, it focuses on addressing the illusion and redundancy that can occur with large language models. It proposes collaboration and hierarchical methods to identify and deal with the illusion and redundancy generated by A during interaction, and improves its ability to automate interaction.

In short, although the development of LLMs-based agents in game system is at an early stage, the basic framework of them has been formed[1]. Existing studies have focused on using LLMs-based agent to mimic player behavior, improve the efficiency of game automation, or improve the immersion of players in the game.

Evaluation Metrics

Evaluation metric is a crucial aspect of the development and optimization for large language model and their applications, which enable researchers to assess the performance of these models, identify their strengths and weaknesses, and improve their efficiency[17]. At present, research on evaluating agent capabilities mainly focuses on the traditional direction, which is the ability to perform advanced tasks in a limited environment, including multiple agent dialogues, plan formulation, etc. However, there is still a lack of environmental awareness in identifying potential risks.

Park proposed a social simulation evaluation technique to evaluate the performance of LLMs-based agents in dialogue systems[27] . This technology takes information from proxies such as goals, rules, and member roles as input and generates design examples with simulated behaviors, including social behaviors such as posts and replies. This study demonstrates that LLMs-based agents can adjust the corresponding text output appropriately according to the scene during dialogue, optimizing the dialogue effect.

It is of significance to explored the similarity between LLMs-based agents and human performance. Argyle's work[28] found that algorithmic biases in medium to large language models are fine-grained, and under appropriate conditions, LLMs-based agents can accurately simulate reactions and behaviors from humans. This study also compares the performance of LLMs-based agents with that of actual humans, demonstrating that the large language model represented by GPT-3 contains a large amount of detailed and multifaceted potential information, reflecting the complex interactions between ideas, attitudes, and socio-cultural backgrounds. This makes LLMs-based agents have enormous potential value in simulating social operations, especially in games.

The combination of professional tools and LLMs-based agents can improve efficiency and accuracy in specific game systems. Qin's work[13] developed a generic tool learning framework for evaluating the potential of LLMs-based agents for proficiency in tool use, including instruction understanding, task decomposition and tool learning. The research also demonstrate the methods of improving agents' learning and generalization capacity in using tools.

In general, accurate evaluation methods are essential for assessing the performance of large language models. The combination of human evaluation and automated evaluation metrics provides a comprehensive assessment of model performance, while tool metrics and similarity assessment can provide additional insights into the agent's behavior.

Discussion

Since the proposal of large language models, there has been a explosion of LLMs-based agents research and applications, with the a trend of growing. In game system field, LLMs-based agents are regarded as a vital role aiming to provide players with brand new and immersive experience felling. Currently, results of related experiments have demonstrated that LLMs-based agents can effectively improve the playing efficiency, interact with players or other agents and show a certain social and strategic. LLMs-based agents will be a momentous research direction in artificial intelligence.

However, there are still many challenges in developing agents in game systems. The defects of the large language model itself can lead to agent anomalies or even serious errors. For example, catastrophic forgetting(CF) can cause the agent to lose a lot of learned knowledge and information, resulting in a precipitous decline in its performance[29]. While there are a number of benchmarks and evaluation criteria available to evaluate agents' abilities, there has not yet been a comprehensive and objective metric to measure how agents differ from human players in a given scenario. The safety of LLMs-base agents is also one of the future research priorities. Large language models can be vulnerable to various types of attacks, such as adversarial examples, poisoning attacks, and data poisoning attacks. These attacks can cause the model to misclassify inputs or produce unintended outcomes[30; 31]. Methods for enhancing the model's robustness against such attacks include adversarial training[32], defensive distillation[33], and defensive sampling. Also, in the process of using LLMs-based agents, effective measures need to be taken to ensure that the their performance is aligned with human values and eliminate the outputs and reactions that may be harmful to humans.

In conclusion, the application of large language models based agents in games helps to enhance the game experience, providing players with smarter NPCS, richer dialogue, and more challenging play experiences. At the same time, these models show potential in game design and development, providing new tools and ideas for game developers.

References

1. Wang, L. *et al.* A survey on large language model based autonomous agents (2023). 2308.11432.

2. Brown, T. B. *et al.* Language models are few-shot learners. (2020). 2005.14165.

3. Liang, E. *et al.* RLlib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning (ICML)* (2018).

4. Xi, Z. *et al.* The rise and potential of large language model based agents: A survey (2023). 2309.07864.

5. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, Online, 2020).

6. Orrù G, C. C. e. a., Piarulli A. Human-like problem-solving abilities in large language models using chatgpt. *Front. Artif. Intell.* (2023).

7. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

8. Zeng, A. *et al.* Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).

9. Du, Z. *et al.* Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335 (2022).

10. Liu, P. *et al.* Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. (2021).

11. Wang, Y. *et al.* Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966* (2023).

12. Qin, Y. *et al.* Toolllm: Facilitating large language models to master 16000+ real-world apis (2023). 2307.16789.

13. Qin, Y. *et al.* Tool learning with foundation models (2023). 2304.08354.

14. Park, J. S. *et al.* Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23 (Association for Computing Machinery, New York, NY, USA, 2023).

15. Shen, Y. *et al.* Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580* (2023).

16. Zhao, W. X. *et al.* A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

17. Chang, Y. *et al.* A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).

18. Dong, Q. *et al.* A survey for in-context learning (2022). 2301.00234.

19. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).

20. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314* (2023).

21. Xu, Y. *et al.* Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658* (2023).

22. Chen, W. *et al.* Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848* (2023).

23. Wang, Z., Cai, S., Liu, A., Ma, X. & Liang, Y. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560* (2023).

24. Zhu, X. *et al.* Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144* (2023).

25. Wang, G. *et al.* Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv: Arxiv-2305.16291* (2023).

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

27. Park, J. S. *et al.* Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18 (2022).

28. Argyle, L. P. *et al.* Out of one, many: Using language models to simulate human samples. *Polit. Analysis* **31**, 337–351 (2023).

29. Luo, Y. *et al.* An empirical study of catastrophic forgetting in large language models during continual fine-tuning. (2023). 2308.08747.

30. Xu, G. *et al.* Cvalues: Measuring the values of chinese large language models from safety to responsibility (2023). 2307.09705.

31. Tian, J. *et al.* Chatplug: Open-domain generative dialogue system with internet-augmented instruction tuning for digital human (2023). 2304.07849.

32. Song, C., He, K., Wang, L. & Hopcroft, J. E. Improving the generalization of adversarial training with domain adaptation. *arXiv preprint arXiv:1810.00740* (2018).

33. Liu, Y., Li, Z., Backes, M., Shen, Y. & Zhang, Y. Backdoor Attacks Against Dataset Distillation. In *NDSS* (2023).