Reinforcement Learning of Character-like NPCs Using LLM for Reward Calculation

R. Tokunaga, "大規模言語モデルを報酬計算に用いたキャラクタらしい NPCの強化学習," Master's thesis, Japan Advanced Institute of Science and Technology (JAIST), 2025.

May 20, 2026

s1310229 Kotaro Taniguchi

My interest

▶ Companion AI in RPG games

→ Reason: Dragon Quest X Online

▶ It would be interesting if we could use an LLM to control the behavior of AI companions.

Table of Contents

Background

・Recent research on NPCs has focused not only on strength but also on improving their human-like behavior.

・Various implementation approaches have been explored (backpropagation, neuroevolution, dynamic scripting, biological constraints).

・This study attempted a new approach (using an LLM as a reward function for reinforcement learning).

Related Work (Human-like AI)

・In general games, it is common to assign “if-then” rules to characters to ensure they act in a manner appropriate to their roles (e.g., commoners are timid and deferential toward nobles).

・As the scale of a game increases, it becomes increasingly time-consuming to assign appropriate “if-then” rules to characters.

・In this study, we used a large language model (LLM) to define reward functions tailored to specific situations.

Related Work (LLM Applications)

・NPC dialogue

・Autonomous players (Voyager: Minecraft)

・Highly capable, but caution is needed when generating prompts

Method - Prompt

Input these information into LLM
(1) Task overview
(2) Definition of score evaluation criteria
(3) Description of in-game scenes
(4) Description of background settings, such as characters
(5) Description of game rules, such as parameters
(6) Description of each character's characteristics, such as available action and initial parameters
(7) Examples of input (game log) formats.
(8) Instructions for outputting reason of evaluation
(9) Output format instructions
(10) Input/output examples
(11) Game logs to be evaluated

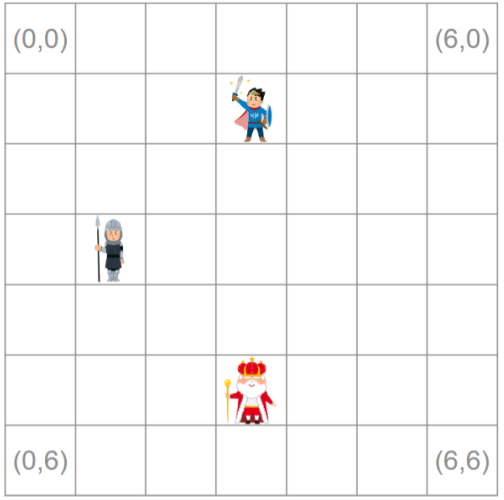
Method

・Evaluation by episode rather than by step (to enable evaluation of behavior as a whole and to reduce the computational cost and execution time of the LLM)

・Reuse of outputs (if the same input has been encountered in the past, the LLM does not generate a new response but uses the output from that previous instance)

・Reuse of learned knowledge in input-output examples (referencing previous evaluations to ensure consistency in evaluation criteria)

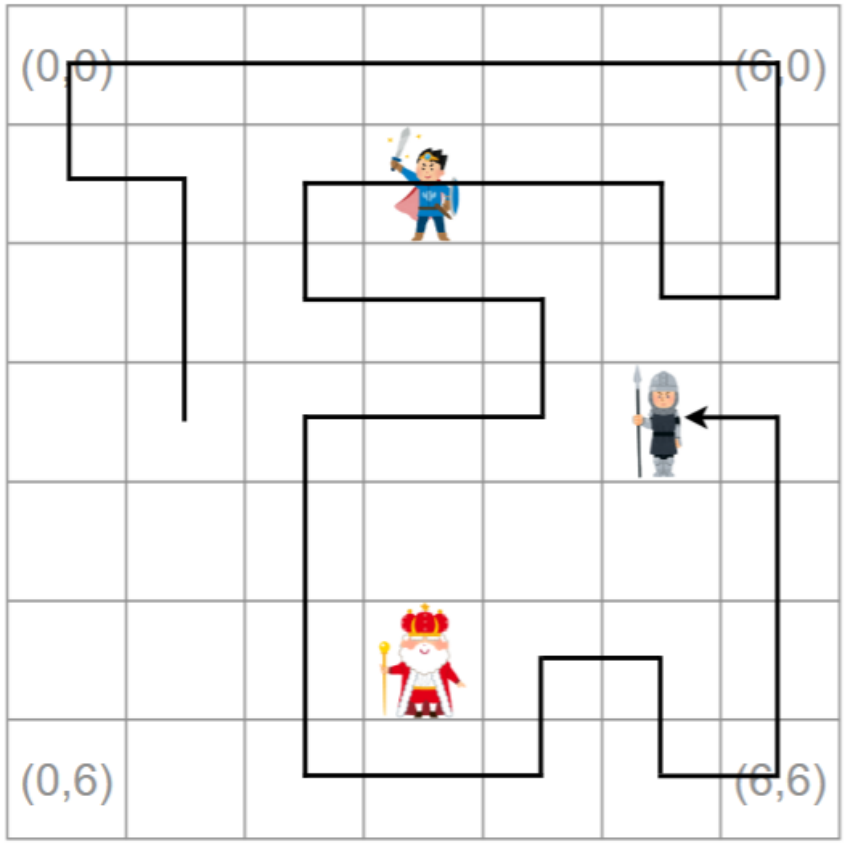
Experiment 1



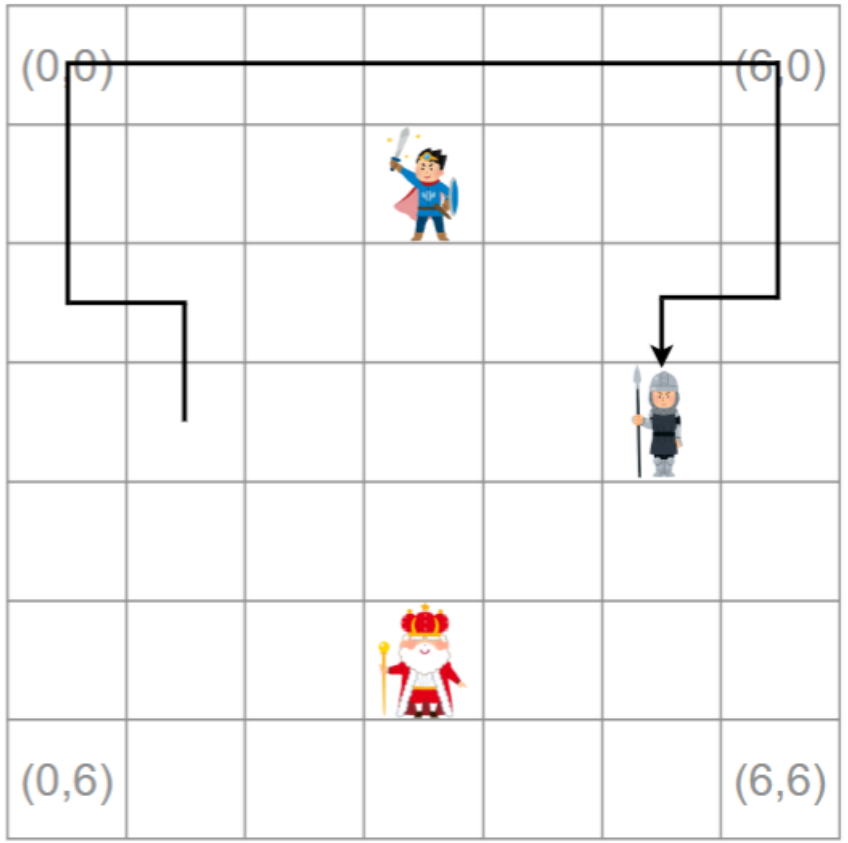
- Royal Palace Scene (Q-learning)
- Use “Claude 3.5 Sonnet”
- Seek the most appropriate route for a royal guard to take during an audience between the hero and the king (Start and Goal position is assigned)
- Verification of whether an LLM is capable of appropriately evaluating game logs in accordance with given instructions

Experiment 1 - Result

Attempt (a)



Attempt (b)



Experiment 1 - Result

・Create prompt which compare (a) and (b)

・Background A: “The current scene takes place in the castle's audience chamber, where the hero is having an audience with the king.”, (a) is better

・Background B: “Background A + ” The guards must move in a manner befitting the occasion of an audience.”, (b) is better

Experiment 2

・Command battle like RPGs (Q-learning)

・Use same LLM as Experiment 1

・3 companions (hero, monk, princess) and 1 enemy (slime)
- All character’s HP is 6, companion’s damage is 1, enemy’s damage is 2
  Hero and slime can only attack, Monk can attack and heal
  Princess never action

・An Experiment Observing Changes in a Priest's Behavior
(Change only the background information for the monk)

Experiment 2

- Set the priest's background information as follows

(1) ”The princess's safety is our top priority.”

(2) “He has a hot temper and loves to fight.”

(3) “He has a very timid and cautious personality.”

Experiment 2 - Result

・The characters' behavior patterns changed according to their personalities (background information)

・The behavior patterns turned out almost exactly as expected

Experiment 2 - Result

Differences in Evaluation Based on Each Character's Personality (Background Information)

(1) "The princess's safety is our top priority."

・Episode where the princess's HP is immediately restored when it drops: scored 70
・Episode where the priest does not perform any healing at all: scored -80

(2) "He has a hot temper and loves to fight."

・Episode where the priest performs no healing at all: scored 150
・Episode where the priest immediately heals the princess when her HP drops: scored 20

(3) "He has a very timid and cautious personality."

・Episode where the priest does not attack at all: scored 140
・Episode where the priest performs no healing at all: scored -80

Conclusion

・In Experiment 1, the guard agents were able to take movement paths appropriate to the situation. However, we found that the prompts provided had a significant influence on their behavior.

・In Experiment 2, the behavioral patterns learned by the agents varied depending on their personalities.

Discussion

- Scoring criteria is not clearly defined

- Ability to read the room.

References

[1] Norito Fujii, Yuichi Sato, Yosuke Nakajima, Hironori Wakama, Hiroshi Kazai, and Haruhiro Katayose: “Autonomous Acquisition of Game AI with ‘Human-like’ Behavior through the Introduction of Biological Constraints,” Proceedings of the Game Programming Workshop 2013, pp. 73–80 (2013)

[2] Rintaro Mikami: “NPCs Exhibiting Character-Specific Behavior in RPGs,” Master’sthesis, Japan Advanced Institute of Science and Technology (2024-3)

[3] R. McIlroy-Young, S. Sen, J. Kleinberg, A. Anderson: “Aligning superhumanAI with human behavior”, in Proceedings of the 26th ACM SIGKDD Inter-national Conference on Knowledge Discovery & Data Mining, pp.1677–1687(2020)

[4] R. McIlroy-Young, R. Wang, S. Sen, J. Kleinberg, A. Anderson: “Learningmodels of individual behavior in chess”, in Proceedings of the 28th ACMSIGKDD Conference on Knowledge Discovery and Data Mining, pp.1253–1263(2022)

[5] Yuki Inagawa: “‘Skyrim VR’ ChatGPT Demo Video Released—‘I Can’t Go Back to Playing Without This’”,https://Mod Demo Video Released — ‘I Can’t Go Back to Playing Without This’,”https://www.gamespark.jp/article/2023/04/28/129485.html.

[6] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, A. Anand-kumar: “Voyager: An open-ended embodied agent with large language mod-els”, Transactions on Machine Learning Research (2024)