Title	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
Author(s)	徳永, 遼太
Citation	
Issue Date	2025-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19840
Rights	
Description	Supervisor: 池田 心, 先端科学技術研究科, 修士 (情報科学)

Using Large Language Models for Reward Calculation in Reinforcement Learning to Create Character-Like NPCs

2110421 Tokunaga Ryota

In recent years, artificial intelligence (AI) technology has rapidly advanced, leading to extensive research and applications across various fields, including games. In particular, large language models (LLMs) such as ChatGPT have gained significant attention because of their high capabilities and broad applicability.

One example of AI technology applied to games is the behavior control of non-player characters (NPCs), which are not operated by human players. NPCs serve as enemies, teammates, or inhabitants of the game world, playing a crucial role in enhancing the overall gaming experience. Traditionally, they have been implemented primarily using rule-based methods designed by developers. However, in recent years, AI techniques such as supervised learning, reinforcement learning, and tree search have been actively adopted for NPC behavior control.

In the era when AI technology was still immature, most complaints about NPC behaviors stemmed from issues such as “not acting appropriately, being too weak.” However, as AI players have surpassed human skill levels in many games, including Go, Mahjong, and StarCraft, a new set of challenges has emerged. AI-controlled NPCs can sometimes be overly strong or exhibit behavior that appears unnatural to human players, which may negatively impact the gaming experience. As a result, research has increasingly focused on developing NPCs that prioritize human-like behavior over skill levels.

For example, Maia is a model trained through supervised learning on a large dataset of amateur chess games. It has been shown to predict human moves more accurately than players developed using tree search or reinforcement learning. In another approach, Fujii et al. created a human-like Super Mario player by incorporating biological constraints, such as “cognitive fluctuations,” “delays between cognition and action,” and “operation fatigue,” into reinforcement learning agents.

In story-driven games with unique world settings, such as role-playing games (RPGs), NPCs are expected to behave not only like generic humans but also in a manner that aligns with their specific character roles. Even among warriors with the same abilities, there can be various characterizations, such as a timid warrior, a brave warrior, an attention-seeking warrior, or a warrior who secretly wishes for their teammates to die. Each of these character roles requires different behaviors.

Modern games feature a vast number of NPCs, and their appropriate behaviors are required not only in predefined maps and events but also in randomly generated situations. In such cases, implementing character behaviors using rule-based systems is impractical, and the diversity of situations makes it difficult to collect a large amount of training data. To address this, reinforcement learning has been explored as an approach where NPCs receive higher rewards for behaving in a manner consistent with their character. However, designers still need to define "what kind of behavior is considered character-like," and as the complexity of the problem increases, this becomes increasingly challenging.

In this work, we hypothesized that LLMs have the ability to "understand a given world setting and situation and reasonably judge desirable states and actions." Based on this assumption, we explored the idea of leveraging LLMs for reward calculation in reinforcement learning. Roughly speaking, we propose a framework where the LLM is given an instruction such as: "In this world setting, the character has the following roles. The current situation is as follows. The agent has taken this sequence of actions, leading to this outcome. Evaluate whether this action sequence is appropriate for the character, providing a score with reasoning." The obtained evaluation score is then used as an episodic reward in reinforcement learning to guide the NPC's behavior.

The first experiment involved a scenario where a royal guard needed to reach a destination without crossing in front of the hero, who was having an audience with the king. In many cases, the agents successfully learned the necessary and sufficient detour route we had intended, and the reasoning provided by the LLM for its evaluations aligned with our expectations. However, in some trials, the LLM highly rated routes where the guard wandered unnecessarily around the room, leading to the unintended learning of such behavior. We consider this an unintended side effect of our approach where episodes with high evaluations were used as samples to further refine the LLM's evaluations.

The second experiment involved a scenario where a party consisting of a hero, a princess, and a cleric engaged in battle with a slime. The task was to train the cleric, who could perform both attacks and healing actions, to behave according to different character roles. We tested three distinct role setting: (1) prioritizing the princess's safety above all else, (2) being aggressive and enjoying battles, and (3) being extremely cautious and timid. As a result, the cleric exhibited behaviors appropriate to each role settings: sacrificing their own well-being to protect the princess, attacking the slime even when the hero or princess was injured, and exclusively focusing on healing.

Through the experiments in this work, we observed that LLMs have a certain ability to “reasonably judge desirable states and actions.” However, we also found that LLMs’ judgments can vary significantly depending on how the instructions are phrased. Future research will need to explore methods to improve consistency in these evaluations.