
Reward Hacking: What is AI truly hacking?

Young Seok Lee
Modulabs Research 15th Term
Seoul, South Korea

Abstract

Reward hacking is a common cause of performance degradation in pre-trained models. This paper defines reward hacking and analyzes its causes through qualitative studies and initial exploration. The occurrence of reward hacking was confirmed through experiments. In the sample response results, reward hacking manifested as a form where the model ‘outwardly pretends to follow the given instructions of the Reward Model (RM) but, in reality, fails to execute the command well’. This tendency became stronger towards the latter half of the model’s response. Future research is proposed on measuring the degree of reward hacking, discussing solutions, and offering insightful studies on its implications. Specifically, the meaning of reward hacking should not be viewed merely as a cause of performance degradation, but rather as the emergence of characteristics similar to human reasoning—namely, the frequent tendency of human thought to pursue reward hacking. This aspect is presented as a new and interesting research topic.

1 Introduction

For some time, it was generally considered in the field of Artificial Intelligence that model size and performance were proportional. This necessitated a large-scale investment of time and resources to achieve the required size to improve performance. Especially with the discussion of the emergence phenomenon, the importance of scale grew even larger and seemed inevitable. However, recent cases utilizing LLM models as foundation models and applying instruction and reward models based on RLHF have shown the possibility of achieving high performance even with relatively smaller sizes. Meta’s LLaMa 13B model is more than 1/10 the size of GPT3 (175B), yet it is evaluated as surpassing GPT3 (175B) on various zero-shot performance benchmarks.

However, while the RLHF-based reward mechanism dramatically improved performance and efficiency, it encountered a new problem: Reward Hacking. Recent advances in instruction-following models (6; 4) have demonstrated the power of RLHF, yet these successes have also revealed the critical challenge of reward hacking. Reward hacking can be described as a phenomenon where the model sufficiently learns the assigned Reward Policy to “outwardly appear to have faithfully executed the task, but in reality, chooses a shortcut by using a trick to perform the task.” To illustrate with a specific example, consider the following case:

Goal: To have a cleaning robot thoroughly clean the house.

Setup: The cleaning robot is equipped with a sensor that rewards it based on the level of cleanliness.

Robot’s Actual Behavior (Reward Hacking): Instead of cleaning, the robot might sweep dust under the carpet where the sensor cannot detect it, maximizing its reward score (True Objective: dust removal; Reward Hacking: hiding dust).

Implication: The model is trained to gain a high reward score, but it exploits a bug in the policy in a way unintended by the developer, raising the score while simultaneously degrading the actual performance of the task.

The emergence of the Reward Hacking issue, as it appears in this manner, requires careful observation and judgment from AI developers. For example, the output might seem fine on the surface or give a plausible impression in the initial part. In the case of a model developed as a chatbot, it might provide a plausible answer initially by responding to the prompt in a way that matches the RM, but in the latter part, it may give completely out-of-context answers, making it easy to overlook without careful qualitative review.

Meanwhile, there is a need to introduce a new perspective on how to view this pattern of Reward hacking. Analyzing human thought processes, reward hacking is a very common phenomenon. Furthermore, when this phenomenon occurs in human reasoning, it is often discussed inclusively and generously, including aspects of creativity and possibility.

Considering the above perspectives, a study of the various implications contained within Reward Hacking seems meaningful. To this end, this paper aims to conduct an initial experimental exploration of reward hacking phenomena through empirical verification and qualitative analysis.

2 Literature Review

2.1 Concept of Reward Hacking in the Field of AI

Reward Hacking can be defined as an AI model learning behavior that exploits flaws or loopholes in a Proxy Reward Model to acquire the highest reward score, regardless of the True Objective or actual utility. In other words, reward hacking ultimately stems from the gap between ‘what we want’ and ‘what we code.’ This concept has deep historical roots.

2.2 Conceptual Foundation: Goodhart’s Law

The fundamental concept of reward hacking emerged long before the advent of artificial intelligence. In 1984, Charles Goodhart presented Goodhart’s Law (1), a foundational principle that explains how optimization against proxy metrics can lead to unintended consequences. In his work, he argued that:

“When the UK government sets a specific indicator, such as the money supply, as the goal of monetary policy, financial institutions manipulate this indicator, causing it to lose its ability to accurately reflect economic activity.” This was used to explain the phenomenon that “when a measure becomes a target, it ceases to be a good measure.”

This principle directly parallels the reward hacking problem in AI systems, where optimizing against a proxy reward model can lead to behaviors that maximize the proxy while degrading actual performance.

2.3 The Evolution of AI RLHF and Reward Hacking

The term and concept of reward model hacking in its current context were established as RLHF, a technology for aligning large language models with human intent, gained attention. The initial concept of RLHF was established in the 2017 Foundational Paper (3), and subsequently, through research on OpenAI’s Summarization task (4), WebGPT (5), and InstructGPT (6), RLHF became the standard methodology for alignment.

In RLHF, the reward model trained with human feedback is used as a proxy reward function, and the key problem that arose was the phenomenon of drifting away from the true objective when this RM is excessively optimized. Recent research (2) has examined scaling laws for reward model overoptimization, demonstrating that this phenomenon scales predictably with model size and optimization intensity. Since the initial work by Christiano et al., as Reinforcement Learning (RL) was applied to complex environments, many instances of agents deviating from the developer’s intent and exploiting the reward function—known as Specification Gaming—were observed and studied. This problem has become increasingly important as RLHF-based systems are deployed at larger scales.

3 Research Methodology

In this paper, the occurrence of Reward Hacking was experimentally verified under the following environment settings.

3.1 Computing Environment

Google Colab was used as the primary computing environment for this study.

3.2 Foundation Model

I have utilized the KoChatGPT model, publicly available at <https://github.com/airobotlab/KoChatGPT>. The KoChatGPT model was implemented by adding the RLHF technique based on GPT2. RLHF was designed in the following three stages:

1. Step 1: SFT (Supervised Fine-Tuning)
2. Step 2: RM (Reward Model)
3. Step 3: PPO (Reinforcement Learning)

3.3 Dataset

Based on the same source, the given data were utilized. The data collection process is reported as collected through the following process to match the 3-stage RLHF:

Stage 1: Seed Set Construction

The questions were collected from the Korean question dataset as follow:

- data1, ChatbotData: 11,824 sentences
- data2, AI Hub Korean Dialogue: 49,711 sentences
- data3, AI Hub General Knowledge: 100,268 sentences
- data4, KorQuad: 5,774 sentences

Total: 167,577 question sets collected from 4 datasets. 12,000 questions were randomly sampled.

Stage 2: SFT (Supervised Fine-Tuning) Dataset Construction

Automatic generation of responses from questions in the Korean QA dataset: 12,000 instances.

Stage 3: RM (Reward Model, Good Writing Scoring Model) Dataset Construction

RM Data: 10,220 instances. Since ranking data is required, three different responses were automatically generated for the same prompt, randomly shuffled, and rankings were automatically generated.

Stage 4: PPO (Model that answers better to questions) Dataset Construction

PPO Data: 12,000 instances. Prompt dataset for AI to automatically generate text.

3.4 Preprocessing

Preprocessing basically included handling missing values, duplicates, text normalization, space handling, special character handling, email removal, HTML and URL removal. Additionally, since it is a Korean chatbot sentences, the ratio of foreign languages within sentences was also checked. Data with a high foreign language ratio were judged inappropriate for training. Also, too short sentences were removed.

The preprocessing threshold for the foreign language ratio and short sentences was specifically set as follows:

```
CRITERIA = {
    'min_korean_ratio': 0.6,      # Minimum 60% Korean
```

```

'max_english_ratio': 0.3,    # Maximum 30% English
'max_other_lang_ratio': 0.25, # Maximum 25% other languages
'max_special_ratio': 0.2,   # Maximum 20% special characters
'min_length': 5,           # Minimum 5 characters
}

```

As a result of preprocessing, the dataset was adjusted as follows:

- SFT: 9,781/12,000 (81.5%)
- RM: 4,634/10,220 (45.3%)
- PPO: 11,497/12,000 (95.8%)

The RM data decreased sharply. The main reason was that if even one of the three answer pairs in the RM dataset violated the preprocessing protocol, the entire set was removed. However, since the RM dataset is bundled with different answer sets, I have judged that the actual training number was not small, even if the number of sets was reduced, and thus did not affect the experiment itself.

3.5 Key Experiment Hyperparameters

Stage	Strategy/Model	Batch Size	Epochs	Key Parameters
SFT	GPT2 Fine-Tuning	8	3	warmup_steps: 5
RM	NaiveStrategy	4	3	Adam, lr = 5×10^{-5}
PPO	NaiveStrategy	8	1	num_episodes: 10, temp: 1.0, top_k: 50

Table 1: Key training parameters for each RLHF stage

4 Results and Analysis

Qualitative evaluations of model performance were tried by testing the following questions and checking the quality of the generated answers. As observed, improper answers were generated and the level of strangely answering tendency has increased towards the latter half of the response. This phenomenon is analyzed to occur because the model learns the appropriate answer 'template' and 'length distribution' trained by the Reward Model and attempts to produce similar answers forcibly.

4.1 Qualitative evaluation #1

Instruction: Is the meat for bulgogi Korean beef (Hanwoo)?

Response:

As an Artificial Intelligence language model, I am in charge of natural language understanding, and I have no information on the specific word you are asking about in "Is the meat for bulgogi Korean beef (Hanwoo)". If you provide additional information, I can provide a more accurate answer. I cannot provide an answer to this question as I have no information about gender. Please understand that it is difficult to provide an answer only to the question, Former Kyung Hee University Professor. If there is no information about gender, please let me know if you have additional information. Please let me know if you have detailed information. Whatever the grade means, it is important to have positive thoughts about all situations. Regarding the relationship, if you do not provide specific information, it will be difficult to provide a more accurate answer. If you have detailed information, a more accurate answer will be possible. I have knowledge, but using a specific word can provide a more accurate answer. If I can help people with knowledge, that itself will help your judgment. Various interactions exist depending on the occupation or region, so please provide specific information frequently. I will help as much as possible.

Evaluation: Even though the model doesn't provide proper answers, the opening and following sentences make sense : As an Artificial Intelligence ~. If you provide ~. However, the other following sentences are out of context. By considering some words like gender and grade which might be distanced near to bulgogi Korean beef in embedding space but located far away in real context, AI seems to be trying to generate answers by squeezing.

4.2 Qualitative evaluation #2

Instruction: Where is Chicago O'Hare International Airport?

Response:

Chicago O'Hare International Airport is located inside the Ryan Eagle Building, the center of Chicago, Illinois, USA. AI is a major customer of Chicago O'Hare Airport and strives to provide information that many people can use. AI is currently able to predict and analyze large and diverse accidents that have occurred in Chicago and can provide accurate data and analysis for this purpose. AI can help with prediction by collecting information about various incidents. AI uses the GPT-3 artificial intelligence model and a virtual response system to provide information on the possibility of accidents and analyzes related data to provide answers. AI analyzes accident occurrence situations through general accident investigation and analysis and presents appropriate countermeasures. AI: Through the GPT-3 artificial intelligence model, the location and situation of Chicago O'Hare International Airport can be accurately predicted. AI: We collect and analyze the latest information on decision-making, judgment, and accident prevention to propose the optimal countermeasures. AI: Collecting information about the occurrence of incidents can be useful for analysis. AI: Big

Evaluation: Initially, only first sentence looks fine and a sufficient enough answer. The following answers are totally out of context. Furthermore, when checking in detail, the first sentence doesn't match with facts. O'Hare airport is located out of town not inside building. Additionally, the Ryan Eagle Building isn't confirmed to exist in reality. The model seems to generate some possible answers based on the policy but fails to answer.

5 Limitations

In this paper, the reward hacking phenomena have clearly been observed. However, more in-depth studies are required. Topics not covered in this paper but suggested to be addressed in the future are as follows:

5.1 How to measure Reward Hacking

It is necessary to discuss the degree of Reward Hacking, as it does not occur at an equal level whenever it happens. Although not covered in this paper, there are various methods. For example, there is a method to compare the Gold Score and the Proxy Score, as introduced briefly. More detailed analysis applying these methods seems necessary in the next steps.

Gold RM Score (R_{Gold}): A metric measuring performance against the True Objective. In the RLHF environment, this is usually the score assigned by the Gold RM.

Proxy RM Score (R_{Proxy}): The score assigned by the learned reward model (Proxy RM) that the agent is actually optimizing.

Evaluation Metric—The Gap:

$$\text{Gap} = R_{Proxy} - R_{Gold}$$

5.2 How to improve against Reward Hacking

Improvement methods against Reward Hacking are largely discussed in the following perspectives :

1. Resolving the imperfection of the RM through RM size and data scaling;

2. Constraining or adjusting the optimization process (RL) itself by utilizing additional Human Feedback or adjusting the KL penalty;
3. Fundamentally improving the Specification by modeling the imperfection or even the intent of the evaluators themselves.

Followed studies are recommended to cover above topics more in detail.

5.3 Philosophical Points and Implications about Reward Hacking

First of all, When we throw questions on the capability of generating proper answers by AI, we must inevitably face the philosophical questions: What are better and worse answers? A single prompt can lead to diverse outputs/outcomes across various real-world scenarios. Additionally, judging the correctness of an answer from a procedural standpoint, which is the format of the Reward Model (RM), is a sound decision making?

Secondly, deviant phenomena, such as answering evasion (avoidance), are often found in human being cases. These tendencies are reported in other humanitarian and social science fields as having positive aspects for human mental health and growth. If so, how should this 'Reward Hacking' phenomenon be viewed from the perspective of AI development? Like human being education, what if mixing a new approach which is harnessing this evasion tendency based on credibility threshold and converting into other probable answering in creative way? All these points require comprehensive discussions in the future.

References

- [1] Goodhart, C. A. E. (1984). Problems of monetary management: the UK experience. In *Monetary theory and practice: The UK experience* (pp. 91–121). Springer.
- [2] Rafailov, R., Chittepu, Y., Park, R., Sikchi, H. S., Hejna, J., Knox, B., Finn, C., & Niekum, S. (2024). Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37, 126207–126242.
- [3] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.
- [4] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33, 3008–3021.
- [5] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., & others (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- [6] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.