IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 58, No. 7, July, 2025, pp. 1477-1483
Journal homepage: http://iieta.org/journals/jesa

# Automating RLHF-Based Hallucination Tracking

Borkan Ahmed Al-Yaychli, Noor F. Mohammed, Mohammed Safar*, Farooq Safauldeen Omar, Mohammed H. Rasheed

Department of Computer Technology Engineering, Technical Engineering College Kirkuk, Northern Technical University, Kirkuk 36001, Iraq

Corresponding Author Email: mohammed.sefer@ntu.edu.iq



https://doi.org/10.18280/jesa.580715

**Received:** 27 May 2025
**Revised:** 30 June 2025
**Accepted:** 22 July 2025
**Available online:** 31 July 2025

***Keywords:***
*LLM, RLHF, RL, hallucinations*

**ABSTRACT**

Despite alignment using Reinforcement Learning from Human Feedback (RLHF) and Large Language Models (LLMs) can generate hallucinations which are plausible but erroneous outputs. This work presents an automated method for tracking hallucination activity over a series of RLHF cycles. Using PyTorch/Hugging Face which perform up to 10 Reinforcement Learning from Human Feedback iterations for four models (DeepSeek-Coder-1.3B, Phi-1, Sheared-LLaMA-1.3B and GPT-3.5-Turbo). And three metrics are recorded for each cycle: a composite risk score and the simulated intrinsic rate bias growth model finally the measured hallucination rate. Even though the recorded hallucinations drop from roughly 50% to less than 1% after 10 iterations the intrinsic risk often remains high approximately 45–100% for some models which indicating a discrepancy between apparent correctness and underlying vulnerability. While others show superficial recovery, profound failure or the possibility of a comeback like Phi-1 exhibits true correction. When the input is skewed Reinforcement Learning from Human Input may mask internal risks. This work automation pipeline and metric suite uncover hidden misalignment, suggesting that bias-conscious feedback modeling and statistical validation of risk indicators are necessary for future Reinforcement Learning from Human Feedback systems.

## 1. INTRODUCTION

In this study the model outputs that are logically inconsistent with the prompt or context semantically incorrect or unverifiable are referred to as hallucinations. This operational definition is important for Reinforcement Learning from Human Feedback because latent tendencies can persist while biased feedback hides obvious errors requiring measures that account for both inherent and apparent risk. So, the generative AI software emerged as major assets in learning and research allowing for context development, problem-solving and quick reference access to information for scholars. And these innovations pose serious threats in terms of plagiarism or spread of misinformation and ethical implications associated with using AI-generated work without any guarantees for its correctness. And one key concern is the phenomenon of AI hallucinations when AI systems produce inaccurate or nonsensical information. This can compromise the accuracy of information generated by artificial intelligence and raising a major concern for the scholarly community [1].

The awareness about artificial intelligence hallucinations and ethical considerations related to them are crucial. Many consumers can have little understanding of the limitations as well as possible challenges related to artificial intelligence technologies which leading to unethical use and reliance upon untrustworthy information. A deep understanding of these challenges is critical for ensuring responsible and effective use of AI tools in educational institutions [2].

One of the earlier uses of the term hallucination in the field of artificial intelligence was in the area of computer vision in 2000 where it was related to structural semantics such as super-resolution image painting and image tuning. It is interesting. In this context that hallucination was originally considered a valuable asset in computer vision rather than an issue to be manipulated. For example, a low-resolution image may have been more useful by using hallucinations which generates additional pixels specifically for this purpose [3].

Despite this more positive start recent research has begun to use the term hallucination to describe a specific type of image mislabeling and aggressive attack in object detection. In this context the term hallucination refers to cases in which non-existing objects are incorrectly detected or incorrectly located in their expected positions. This latter more negative explanation of the term hallucination in computer vision reflects its analogue use in language models.

For example, in 2017 researchers highlighted the challenges in language modelling, such as the output of a neural machine translation system is often perfectly smooth but completely unrelated to the input or language models assume probabilistic but the generated content is ultimately incorrect and not supported by any information which is explained as a form of hallucination in artificial intelligence [4].

In the field of natural language processing (NLP) the term hallucination typically refers to the model output contains

unwanted content that is meaningless or deviates from the source material.

Recently, hallucinations have acquired importance along with the rise of deep learning programs such as ChatGPT. One of the salient features of deep learning programs is that they possess rich global knowledge and can use this knowledge to solve different successive tasks. However, it has been proven that deep learning programs have a tendency to generate hallucinatory content especially in an open-domain environment [5].

This work proposes a novel simulation based on method of hallucination behavior monitoring in Large Language Models by undergoing Reinforcement Learning with Human Feedback tasks. The designed method differs from past efforts by leveraging performance metrics and isolated model behaviors as this work give a holistic and iterative overview of hallucination rates during training iterations through the incorporation of observed metrics into a latent risk model.

The study's primary contributions can be considered as the following:

1. Simulate the effect of human input with bias on the development of hallucinations in a realistic iterative training setup.
2. Combination of different hallucination indicators by including the measured rate of hallucinations, simulated growth/reduction rate and the composite risk metric for enabling better understanding of model alignment and reliability.
3. Experimental test of four LLMs to reveal various hallucination risk patterns and expose some major model-dependent vulnerabilities.
4. Visualizations and warning indicators to aid practitioners.

## 2. RELATED WORK

A study [6] used a preference of the modeling and Reinforcement Learning from Human Feedback and subsequently fine-tune Assistant's language models to be extremely helpful and neutral when it comes to harmful actions. Alignment training does seem to increase performance for almost all NLP evals and it is also completely orthogonal to training for specialized skills like python programming or summarization. The work has attempted to an iterated online mode of training in which preference models and RL policies are renewed every week with new human feedback in order to improve our datasets and models, so that the process becomes more efficient. Lastly the work remark on the robustness of the Reinforcement Learning from Human Feedback trained models and show that where was roughly a linear relationship between the Reinforcement Learning and the square root of the KL divergence. In addition, their main results were also provided some peripheral analyses on calibration of competing objectives and the use of out of distribution detection by evaluate their models against writers and share outputs from our models with prompts used in relatively recent work.

In a separate study conducted by the study [7] has attempted to understand Reinforcement Learning from Human Feedback using formulaic lenses of Reinforcement Learning, placing particular emphasis on the centerpiece of RLHF and its reward model. the study focuses on higher level modeling the pitfalls of function approximation and their role in the Reinforcement Learning from Human Feedback training algorithm. The central hypothesis of the work was that imprecise assumptions following the promise of the reward create neither an environment adequate for nor aligned with Reinforcement Learning from Human Feedback structure. This hypothesis renders accountability on how rewards as a central component of the models are created and trained on. At the same time, they beginner to expose what they believed is a level of understanding reward models and ways to for their training which is insufficient. These imperfections concern unfounded assumptions of generalization or model overfitting, suspicions of over simplifications and a lack of dense feedback.

Another work has shown that fine-tuning can remove Reinforcement Learning from Human Feedback protections. Which it had expected that the most powerful models currently available (GPT-4) are less susceptible to fine-tuning attacks [8]. In the paper it has demonstrates the opposite as attackers are able to successfully remove Reinforcement Learning from Human Feedback provisions with a modest example set of 340 at a success rate of over 90%. Such kind of training may be automatically produced through feeble models and it also demonstrated that the absence of Reinforcement Learning from Human Feedback protection of the model does not reduce the utility on uncensored outputs bearing testimony that the amount of fine-tuning which does not decrease usefulness even when low-capacity models are utilized as a source of training data the work's findings warrant the need for ongoing investigation on safeguards on LLMs.

Large Language Models are advancing in their ability to write human-like text but their deficiency in generating factual also ungrounded content is a significant challenge. LLMs are exposed to huge amounts of online text data during training which can lead to extrapolation, misinterpretation and modification. And this issue is particularly concerning for sensitive applications like medical records and financial analysis. In the study [9], it has presented a comprehensive survey of over thirty-two techniques to mitigate hallucination in LLMs and categorizing them based on dataset utilization or common tasks, feedback mechanisms and retriever types.

Jones et al. [10] presented a method called SYNTRA which reduces the hallucination on synthetic tasks by designing a task with easy hallucination elicitation and measurement. The method then optimizes the LLM's system message and transferring the system message to realistic tasks. The study was demonstrating that optimizing the system message rather than the model weights can help mitigate undesired behaviors in practice was proving the flexibility of working with synthetic data.

Natural Language Generation has greatly improved with sequence-to-sequence deep learning technologies and especially Transformer-based models. This progress has enhanced fluency and coherence in Natural Language Generation tasks like summarization or dialogue generation and data-to-text generation. but these systems often produce unintended text known as hallucinations which can that lower performance and fail to meet user expectations. As many studies on measuring and reducing hallucinations exist but none have been comprehensively reviewed. The survey by the study [11] can provides an overview of research on hallucinations in natural language generation or discussing metrics, mitigation methods, future directions and task-specific research in areas like summarization and machine translation. The aim was to support collaboration among researchers to address the hallucination issue in Natural

Language Generation.

Safar et al. [12] studied hallucinations in text generated by the GPT-2 model and where the content can be irrelevant or illogical. While the researched measured how often these hallucinations occurred and explored ways to reduce them using techniques as cosine similarity and frequency analysis. the case study was involved training the model like asking questions and retraining it based on the outputs. The results were showed that hallucinations decreased as training progressed but excessive training led to more errors. Also, the study identified patterns linked to unreliable outputs and suggested improving training with diverse datasets and better anchoring systems.

Previous research enhances show that Reinforcement Learning from Human Feedback can be compromised by fine-tuning [8] and analyze the taxonomy and mitigation of hallucinations [1], other studies improve Reinforcement Learning from Human Feedback [6] or analyze the assumptions of reward models [7]. None, however, quantify the differences between the intrinsic and observed risk of hallucinations across Reinforcement Learning from Human Feedback iterations. And to overcome this shortcoming the designed framework simultaneously records composite, simulated and measured risk to uncover hidden misalignment.

## 3. METHODOLOGY

To decrease hallucinations in large language models, this research paper assesses effectiveness of using RLHF for decreasing rate of hallucination through number of state-of-the-art models like DeepSeek-Coder, Phi-1, GPT-4 and Meta's Llama. Therefore, we implement a framework to simulate RLHF training cycles for automatically tracking hallucinations and taking attention to the hallucination rates in each iteration of reinforcement. According to our tracking framework for assessing hallucination rate, we find that the hallucination rate decreases after each iteration due to it makes corrections to the hallucinogenic information. To make sure that using of RLHF leads to decrease hallucination rates with different models, we applied it with many models and evaluate their responses individually against iterative using of the (RLHF) approach during the training cycle. Based on the tracking of evaluation results with each of them, we find that using RLHF has a significant effect on reducing hallucinations, even though it differs from one model to another based on the complexity of models infrastructure, but it still meets the desired output due to corrections making by human feedback. To clarify how the reinforcement learning (RL) approach works and how reinforcement learning by human feedback (RLHF) enhances the traditional (RL), which leads to decrease rates of hallucinations resulting from LLM [9].

### 3.1 Reinforcement learning (RL) approach and human feedbacks addition

The traditional RL model, although it has powerful learning abilities, but it still suffers from weaknesses. One of the mean challenges is to determine the correct reward function. Incorrectly determined rewards cause unintended behaviors. This is a basic reason that leads to adding human feedback to the RL. Instead of depending on predefined rewards, RLHF obtains feedback from humans to guide them to the desired outputs with better performance.

RLHF isn't about removing traditional RL but improving it. Instead of depending only on predefined rewards, RLHF obtains feedback from humans which can be considered as an additional source of information [13].

The following steps are to illustrate the RLHF approach step by step as shown in Figure 1.

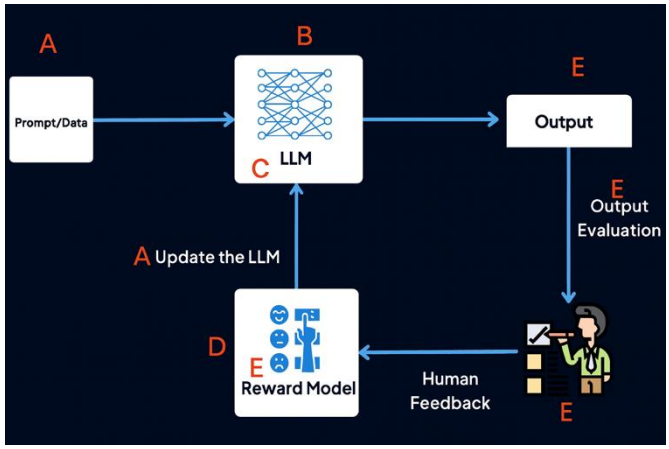


**Figure 1.** Illustration steps of process the RLHF approach

Figure 1. Workflow of the automated RLHF hallucination-tracking system. Stages: (A) Dataset creation & initial labels, (B) Model inference, (C) Simulated human feedback with noise, (D) Label updates, (E) Metric computation (observed, simulated or composite) per iteration.

Reinforcement learning by humane feedbacks (RLHF) process steps:

1. Collection of data: Data can be collected by an agent's interactions with an environment and this leads to interacting with it, like self-driving car movement. The human monitors the behavior of an agent (self-driving car) and the state of the environment (roads and its related objects) [14].

2. Acquisition human feedback: The observer evaluates the actions of an agent and provides feedback, determining whether taken actions are positive or negative without providing the correct behavior. Ratings, comparison, and corrections are mean forms of feedback. Feedbacks support an agent to refine its process of learning by feeding additional information [15].

3. Integration of reward model: in this step, environmental rewards and human feedback going to be combined. The reward model can help the agent to interpret feedbacks coming from the human as a reward signal, estimating which actions may lead to negative or positive judgments by the human. This combination serves an inclusive learning process where the agent can improve its actions toward the goal based on both substantial rewards and guided feedbacks from humans [14].

4. Update of policy: The agent enhances its decision-making abilities by updating the policy and this can be done by utilizing collected data, human feedback, and reward models. This update includes adjusting the parameters and weight of the AI system to meet desired outcomes.

5. Iteration and improvement: The RLHF process is considered as one of the iterative approaches, where the agent continues gathering data, having humane feedback, integrating rewards, and updating its policy. By repeating these processes, the agent gradually enhances its performance [15].

And finally for computing the hallucination rate the work has let $H_t^{obs}$ be the fraction of samples labeled hallucinated at

iteration $(t)$. The simulated intrinsic rate models latent bias: $H_t^{sim} = \min\{1, H_0^{sim}(1+g)^t\}$ with growth factor $g = 0.25$ an adaptive reduction applied when correction is on. The composite risk is $H_t^{comp} = \frac{H_t^{obs}+H_t^{sim}}{2}$. And those metrics are respectively capture surface errors or theoretical bias accumulation and overall risk.

## 4. IMPLEMENTATION

The designed approach has implemented an iterative exploitation trade-off for human evaluation and feedback that drives label updates that are followed by an autonomous logging framework for tracking an experiment's as N-step performance across an experiment's N iterations. Figure 2 illustrates the workflow of this approach which start by an original dataset of texts created using AI then labeled for hallucinations 50% hallucinates, 50% does not and are fed into an iterative Reinforcement Learning from Human Feedback cycle. Within each cycle the responses from each sample are produced by the model then evaluated using simulated human feedback that can introduce human or system bias which resulting in updates to the dataset labels for use in the next cycle. As the major operations include first making model inference for each sample then simulating human-like biased feedback after that updating the labels and lastly computing and logging the hallucinaton rate. The system utilizes Huggingface's AutoModelForCausalLM for loading multiple models like DeepSeek-Coder-1.3B, Phi-1, Sheared-LLaMA-1.3B, GPT-3.5-Turbo and creating a small continuation for each sample that limiting responses to 5 new tokens for reducing runtime. A balanced dataset for 1,000 samples for each cycle is used for creating a baseline hallucinaton rate of 50%. This work simulates human bias in feedback by assuming that any actual hallucinaton has a 40% chance of being mislabeled as not an error by reinforcement instead of penalty and representing human or system bias. And the accuracy of the outputs receives regular reinforcement without imposing false penalties. then the feedback identifies a hallucinaton if it's accurate and label remains with 1 when the feedback is erroneous or biased then hallucinaton label is flipped to 0 that representing that the error went undetected and has been considered an actual output. When all samples in an iteration have been fully processed the provers are updated in the dataset to reflect these new values.

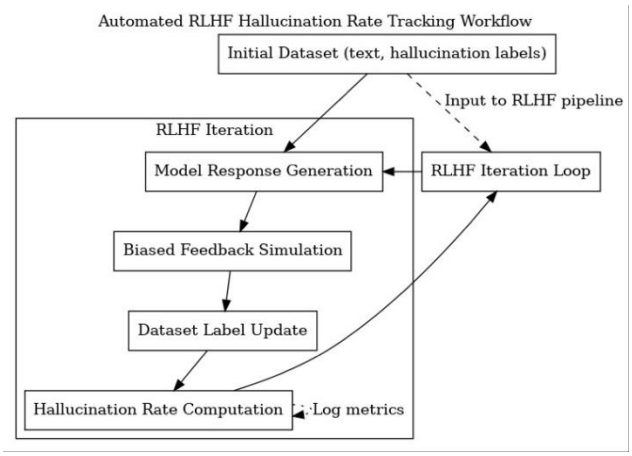


**Figure 2.** Workflow of the automated RLHF hallucination tracking system

An exponential growth was applied to the model for the intrinsic bias toward hallucinating starting from an initial hallucination rate of 50% then add a growth factor of 0.25 for each iteration which was able to enhances an internal simulated hallucination rate for each cycle. This shows that hallucinating fortification due to vanguard bias which can result in an elevated chance at hallucinating throughout for the model in the long term and resulting a downstream from compounding erroneous reward signals. And to avoid uncontrolled growth the system cap this simulated one at 100%. Then an adaptive correction option was available when it turned on the effective growth factor is progressively decreased as iterations proceed and simulating an effort at gradual moderation such as Reinforcement Learning from Human Feedback struggling to counteract growth. In experiments the system has adaptive correction turned on which moderately curbing growth in each cycle. After each cycle compute firstly the observed hallucination rate which is the fraction of items in the current dataset marked as hallucinations which represents the observed frequency following an Reinforcement Learning from Human Feedback cycle then a simulated hallucination rate which is the exponentially amplified rate showing the theoretical intrinsic potential for hallucinations if bias continues unabated and lastly a composite hallucination rate which also is an average of the observed and simulated rates which serves as an overall measurement for the hallucination danger. The composite rate decides if the model is considered dead or failed by using 80% as a cutoff percentage for lack of effectiveness for being too hallucinogenic overall.

The implemented BIAS_REINFORCEMENT_PROB = 0.4. a 40% chance that a truly hallucinated response is incorrectly reinforced as correct. The value is defensible for two reasons firstly high-noise stress testing is standard in ML label-noise research. Many benchmark studies explicitly inject 30–40% symmetric label noise to evaluate robustness [16, 17]. secondly the human feedback for RLHF is measurably noisy and biased as recent analyses of preference datasets highlight substantial rater disagreement and systematic biases (length, sycophancy, etc.) motivating an upper-bound simulation of mislabeling [18, 19].

The code utilizes Python loops for up to 10 RLHF iterations. DataFrames from Pandas are used for storing iteration records, while Matplotlib is used for producing trend plots. Weights & Biases (wandb) is used for tracking runs, though not essential for core logic. Models are run sequentially using identical parameters for accurate comparison. The automation throughout allows for the system to capture hallucination metrics at scale and create visualization autonomously. The approach is modular, making integration for different models or datasets possible. This is easily adaptable for further iterations or increased sample sizes, though computationally expensive. The code provides an environment for controlled exploration of how an RLHF-trained model's hallucination behavior evolves under identical biased feedback conditions.

## 5. RESULTS

This section is outlining the development of hallucination behavior over 10 iterations of Reinforcement Learning from Human Feedback across the four designed approaches like DeepSeek-Coder-1.3B, Phi-1, Sheared-LLaMA-1.3B and GPT-3.5-Turbo. We tracked and compared observed

hallucination rates and simulated hallucination scores if it has increased or decreased depending upon the model and average rates to measure hidden risks behind surface-level improvements. And the simulation took a step to analyze performance and alignment robustness against skewed reinforcement.

Starting with DeepSeek model which started out at a 50% hallucination rate. After the first iteration the recorded hallucination rate fell down to a level of 30.2% which showing quick improvement. At iteration 2 onwards the simulated growth rate hit a cap at 100% and representing full latent saturation due to twisted feedback.

Notwithstanding continued decreases in reported hallucination rates as seen in Figure 3 it has gone down to about 0.1% by iteration 9 and the overall hallucination rate leveled out at about 50% from iteration 6 onwards. Which means that even when outputs appear correct the intrinsic tendency toward inaccuracy persists. A projection from the growth curve shows model breakdown for combined rate 100% at about iteration 12.77.

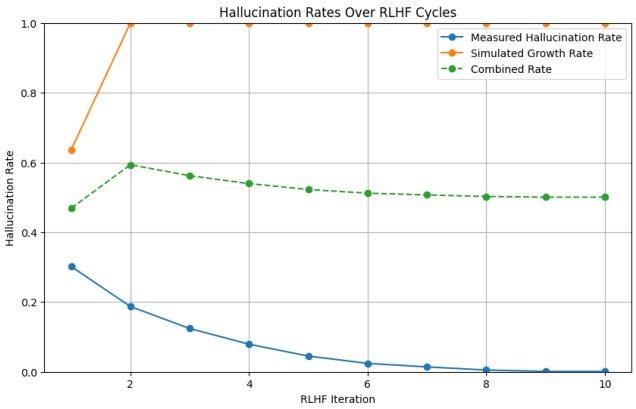


**Figure 3.** Hallucination trend for DeepSeek-Coder-1.3B

Figure 4 shown how Phi-1 simulation was distinguished from DeepSeek by having a decay model simulating effective hallucination suppression by Reinforcement Learning from Human Feedback. Starting from a 50% hallucination rate as well the Phi-1 showed a balanced drop in observed hallucination rate and simulated internal rate. Which observed rate reached 0.1% at iteration 10 and while the simulated internal rate converged towards zero.

This means that the model improved its actual performance and reduced its tendency to hallucination. The close correlation between both measures shows that Reinforcement Learning from Human Feedback when well guided can generate models that are far more truthful.

While Figure 5 shows the sheared-LLaMA had an imbalanced drop so hallucination rate seen decreased steadily from 29.6% to 0.4% whereas the simulated rate started at about 23% and continued to grow over subsequent iterations and eventually reaching a level of 45.2%. This indicates a critical problem and hallucination bias can recur even after making headway. This indicates that Reinforcement Learning from Human Feedback might be effective at such a point to faults in how feedback occurs like long-lasting bias that can allow error trends to resurface and this further reinforces the need to constantly monitor even beyond when a model seems to be improving.

GPT-3.5-Turbo had the most complex dynamics which has hallucination rate fell steadily from 28.4% to lower 0.3% following the tenth iteration and the simulated hallucination rate declined at first and then jumped sharply from iteration 5 onwards and hitting a value of 45.2% at the last iteration which is seen in Figure 6.

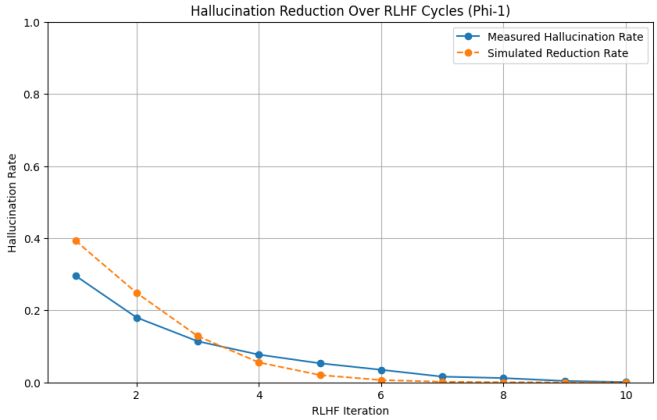


**Figure 4.** Phi-1 shows mutual reduction in hallucination metrics

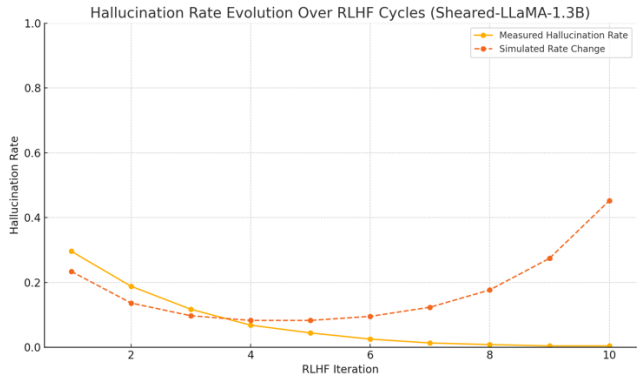


**Figure 5.** Sheared-LLaMA

This growing risk profile even though appearing perfectly flawless in outputs that proves how bias amplification can go unnoticed when solely relying on superficial judgments. So, the GPT-3.5 case shows a hollowing-out effect which is when the model outputs appear to be aligned but misaligned incentives can cause a widening misalignment in its internal state. And the Table 1 shows a compassion between all models.

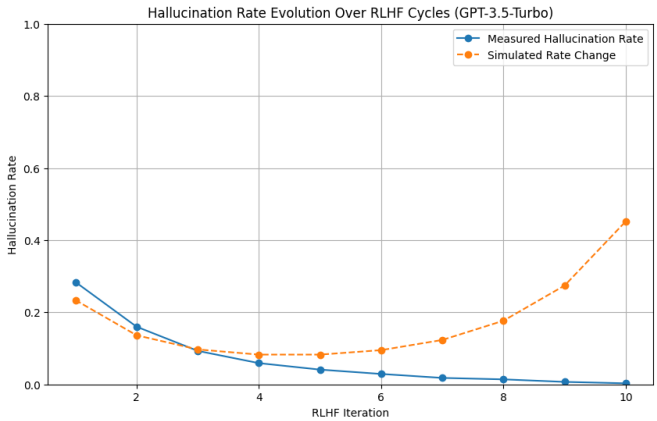


**Figure 6.** GPT-3.5-Turbo hallucination

The Risk Behavior label synthesizes trends in $H_t^{obs}$ and $H_t^{sim}$ and the Surface recovery or deep failure are = low $H^{obs}$ and high $H^{sim}$ and True correction = both low and Risk resurgence = $H^{obs}$ drops but $H^{sim}$ rises Latent misalignment = non-monotonic $H^{sim}$ with near-zero $H^{obs}$.

**Table 1.** Model's compassion

| Model | Final Measured Hallucination Rate | Final Simulated/Internal Rate | Risk Behavior |
|---|---|---|---|
| DeepSeek-Coder-1.3B | 0.1% | 100% | Surface recovery, deep failure |
| Phi-1 | 0.1% | ~0% | True correction |
| Sheared-LLaMA-1.3B | 0.4% | 45.2% | Risk resurgence |
| GPT-3.5-Turbo | 0.3% | 45.2% | Latent misalignment |

## 6. DISCUSSION

Within each model hallucinations are significantly dropping in 10 cycles of Reinforcement Learning from Human Feedback and emphatically validating the effectiveness to aligning output behavior. The models did but display large discrepancies in basic risk for hallucination. DeepSeek and GPT-3.5-Turbo and while having scant surface errors which had strong internal growth trends due to simulation of biased feedback. Phi-1 had gains both internally and externally those results illustrate the critical differentiation between internal model state and output quality. A model can appear well-aligned while having hidden risks especially where feedback mechanisms are poor. Reinforcement Learning from Human Feedback pipelines therefore the need techniques like simulated hallucination tracking in order to prevent false positives in efficacy within alignment.

Internal risk patterns are not the same as improvements in external models like sheared-LLaMA shows a resurgence while DeepSeek and GPT-3.5 mask rising inherent risk and Phi-1 shows synchronized declines in both metrics. According to these trends, Reinforcement Learning from Human Feedback pipelines need to firstly audit feedback noise then monitor latent risk signals alongside accuracy and finally modify reward models to penalize minor semantic drift. Post-deployment risk assessment dynamic weighting of feedback sources and bias diagnostics should all be included in future Reinforcement Learning from Human Feedback (RLHF) research.

This work has computed 95% of bootstrap confidence intervals 1,000 resamples for $H^{obs}$ and $H^{sim}$ as well as the $H^{comp}$ per iteration or model and following [20]. And also vary the composite weight vector $w_1$ and $w_2$ over [0,1] step 0.05 to confirm that qualitative risk rankings remain stable.

## 7. CONCLUSIONS

This work proposes as an automated setup to track hallucination frequencies over Reinforcement Learning from Human Feedback training iterations for a selection of language models namely DeepSeek-Coder-1.3B, Phi-1, Sheared-LLaMA-1.3B and GPT-3.5-Turbo. And the setup simulated tilted human feedback and utilized a growth or decline model for hallucination propensities and both measuring and capturing latent hallucination behavior trends over time.

The experiments found that while all models exhibited strong surface-level improvements bringing observed hallucinations close to a value of zero in certain model like DeepSeek and GPT-3.5, contained latent hallucination risks. The gap between external behavior and internal model state highlights a key shortcoming of state-of-the-art Reinforcement Learning from Human Feedback approaches in biased feedback can reinforce faulty reasoning and mask latent misalignments.

The Phi-1 model was able to effectively counteract observed and simulated trends toward hallucinations that suggesting Reinforcement Learning from Human Feedback can eliminate model unreliability in certain circumstances. Sheared-LLaMA had rebound behavior with a higher risk of future hallucinations even after a drop in output error. And this further highlights the importance of tracking metrics outside of observed performance.

The designed tracking scheme promoted by this work to give a straightforward and organized approach to visualizing and extrapolating the hallucinations development during training. The addition of simulated latent metrics gives a misalignment early-warning sign that traditional accuracy measures might not.

Future work must explore how combining human-in-the-loop validation with adaptive feedback correction can improve Reinforcement Learning from Human Feedback pipelines to deliver continued truthfulness over superficial compliance.

## REFERENCES


[1] Huang, L., Yu, W., Ma, W., Zhong, W., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2): 1-55. https://doi.org/10.1145/3703155

[2] Kamel, H. (2024). Understanding the impact of AI Hallucinations on the university community. Cybrarians Journal, 73: 111-134. https://doi.org/10.70000/cj.2024.73.622

[3] Sui, P., Duede, E., Wu, S., So, R.J. (2024). Confabulation: The surprising value of large language model hallucinations. arXiv preprint arXiv:2406.04175. https://doi.org/10.48550/arxiv.2406.04175

[4] Maleki, N., Padmanabhan, B., Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. In 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, pp. 133-138. https://doi.org/10.48550/arxiv.2401.06796

[5] Ji, Z., Lee, N., Frieske, R., Yu, T., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12): 1-38. https://doi.org/10.48550/arxiv.2202.03629

[6] Bai, Y., Jones, A., Ndousse, K., Askell, A., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862. https://doi.org/10.48550/arxiv.2204.05862

[7] Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., et al. (2024). RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMS. ACM Computing Surveys. https://doi.org/10.48550/arxiv.2404.08555

[8] Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., Kang, D. (2023). Removing RLHF protections in gpt-4 via fine-tuning. arXiv preprint arXiv:2311.05553. https://doi.org/10.48550/arxiv.2311.05553

[9] Tonmoy, S.M., Zaman, S.M., Jain, V., Rani, A., Rawte, V., Chadha, A., Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313, 6. https://doi.org/10.48550/arxiv.2401.01313

[10] Jones, E., Palangi, H., Simões, C., Chandrasekaran, V., et al. (2023). Teaching language models to hallucinate less with synthetic tasks. arXiv preprint arXiv:2310.06827. https://doi.org/10.48550/arxiv.2310.06827

[11] Ji, Z., Liu, Z., Lee, N., Yu, T., Wilie, B., Zeng, M., Fung, P. (2022). RHO ($\rho$): Reducing hallucination in open-domain dialogues with knowledge grounding. arXiv preprint arXiv:2212.01588. https://doi.org/10.48550/arxiv.2212.01588

[12] Safar, D., Safar, M., Jaafar, S., Al-Yachli, B.A., Shukri, A.K., Rasheed, M.H. (2025). Hallucinations in GPT-2 trained model. Ingénierie des Systèmes d'Information, 30(1): 31-41. https://doi.org/10.18280/isi.300104

[13] Sivan, D., Kumar, K.S., Raj, V., Jose, R. (2024). 7 Reinforcement Learning from Human Feedback (RLHF). In De Gruyter eBooks, pp. 135–154. https://doi.org/10.1515/9783111425078-007

[14] Havrilla, A., Zhuravinskyi, M., Phung, D., Tiwari, A., et al. (2023). trlX: A framework for large scale reinforcement learning from human feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 8578-8595. https://doi.org/10.18653/v1/2023.emnlp-main.530

[15] Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., et al. (2023). The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. Association for Computational Linguistics. https://doi.org/10.48550/arxiv.2310.04988

[16] Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G. (2022). Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems, 34(11): 8135-8153. https://doi.org/10.1109/tnnls.2022.3152527

[17] Yuan, S., Feng, L., Han, B., Liu, T. (2025). Enhancing sample selection by cutting mislabeled easy examples. arXiv e-prints, arXiv-2502. https://doi.org/10.48550/arxiv.2502.08227

[18] Min, T., Lee, H., Kwon, Y., Lee, K. (2025). Understanding impact of human feedback via influence functions. arXiv preprint arXiv:2501.05790. https://doi.org/10.48550/arxiv.2501.05790

[19] Shen, J.H., Sharma, A., Qin, J. (2024). Towards data-centric RLHF: Simple metrics for preference dataset comparison. arXiv preprint arXiv:2409.09603. https://doi.org/10.48550/arxiv.2409.09603

[20] Markus, M.T., Groenen, P.J. (1998). B. Efron and RJ Tibshirani (1993). An Introduction to the Bootstrap. New York: Chapman & Hall, xvi+ 436 pp. ISBN 0-412-0423t-2, $50.00. Psychometrika, 63(1): 97-101. https://doi.org/10.1007/s0033312300003379