Optimizing Process Based Reward Models through Reinforcement Learning for Verifiable Multi Step Reasoning in Large Language Model Architectures

Frederick Prescott
Department of Computer Science and Engineering
University of Nebraska–Lincoln
f.prescott@unl.edu

Samuel Thornton
School of Computing and Information Systems
Grand Valley State University
s.thornton@gvsu.edu

Abstract

The evolution of large language models has transitioned from simple predictive text completion toward complex, multi-step cognitive reasoning. However, traditional outcome-based reward models, which evaluate only the final correctness of a solution, often fail to identify logical fallacies or "hallucinations" occurring within intermediate steps. This paper explores the optimization of Process-Based Reward Models (PRMs) through reinforcement learning to enhance the verifiability and robustness of multi-step reasoning in large-scale model architectures. Unlike traditional approaches, PRMs assign value to each distinct stage of a reasoning chain, providing a more granular signal for training. This study analyzes the structural trade-offs involved in deploying these models at scale, focusing on the infrastructure requirements, the computational overhead of step-wise verification, and the socio-technical implications of automated reasoning governance. We argue that while process-based supervision significantly improves the reliability of models in high-stakes domains such as law, medicine, and engineering, it introduces unique challenges regarding system latency and the sustainability of human-in-the-loop feedback loops. By integrating reinforcement learning with process-oriented feedback, developers can foster a more transparent AI ecosystem where the path to a conclusion is as scrutinized as the conclusion itself. The discussion encompasses the broader implications for algorithmic fairness, the reduction of black-box opacity, and the policy frameworks necessary to govern verifiable machine intelligence in modern socio-technical infrastructures.

Keywords:
Large Language Models, Process Based Reward Models, Reinforcement Learning, Multi-Step Reasoning, System Architecture, Algorithmic Governance, Verifiable AI.

1. Introduction

The contemporary landscape of artificial intelligence is increasingly defined by the pursuit of reliable reasoning. As large language models are integrated into critical infrastructures, the necessity for these systems to provide not just answers, but verifiable and logical trajectories of thought, has become paramount [1]. Traditional training paradigms have largely relied on reinforcement learning from human feedback centered on outcome-based evaluations. While effective for stylistic alignment and factual retrieval, these methods frequently overlook the internal logic of the model [2]. When a model arrives at a correct conclusion through flawed reasoning, it reinforces a "silent failure" mode that can be catastrophic in specialized applications. This has led to the emergence of process-based reward models, which seek to decompose complex tasks into discrete, evaluative units [3]. The shift toward process-oriented optimization represents a fundamental change in how machine intelligence is audited and governed.

The optimization of these models through reinforcement learning requires a sophisticated understanding of system-level trade-offs [4]. To implement verifiable multi-step reasoning, researchers must balance the increased granularity of the reward signal against the exponential rise in computational demand. Every intermediate step that requires a reward assessment introduces a new layer of latency and a higher requirement for high-quality, annotated data [5]. From a socio-technical perspective, this movement toward process verification is not merely a technical refinement but a response to the growing demand for algorithmic transparency [6]. Stakeholders in government and industry are increasingly wary of black-box models that cannot explain their internal decision-making processes. Therefore, developing architectures that prioritize the verification of the reasoning path is essential for the long-term sustainability and public trust of autonomous systems [7].

2. Architectural Frameworks for Process-Based Supervision

Building an architecture capable of supporting process-based reward models necessitates a departure from monolithic inference pipelines. In a standard outcome-based setup, the model generates a full response before any evaluation occurs. In a process-based architecture, the system must be modularized to allow for the insertion of checkpoints where the reward model can intervene and assess the validity of the preceding segment [8]. This modularity demands a highly resilient infrastructure capable of managing asynchronous calls between the primary generative agent and the supervising reward model. The integration of reinforcement learning into this loop allows the system to learn which reasoning paths are most likely to lead to successful outcomes, effectively pruning the search space and focusing computational resources on viable trajectories [9].

The structural trade-offs in these architectures often manifest in the tension between depth and speed [10]. A high-density verification strategy, where every sentence or logical atomic unit is scrutinized, provides the most robust reasoning but at the cost of significant inference delays. Conversely, a sparse verification strategy might only check major milestones, risking the propagation of errors in the intervening steps [11]. Achieving an optimal balance requires a dynamic infrastructure that can adjust its verification intensity based on the perceived complexity of the query [12]. For instance, a simple factual inquiry might require minimal

process-based oversight, whereas a multi-layered engineering problem would trigger a comprehensive, step-by-step audit. This adaptive approach ensures that system resources are allocated efficiently while maintaining a high standard of logical integrity across various domains [13].

3. Reinforcement Learning Paradigms in Multi-Step Reasoning

Reinforcement learning serves as the primary engine for refining how process-based reward models interact with generative architectures. By treating the reasoning chain as a series of actions within a Markov decision process, researchers can apply policy gradient methods to optimize the likelihood of producing valid intermediate steps [14]. This goes beyond simple supervised fine-tuning by allowing the model to explore various reasoning strategies and receive feedback on their structural validity [15]. The challenge lies in the credit assignment problem at the step level. Determining exactly which step in a twenty-part reasoning chain contributed to an eventual failure—or which step was the pivotal moment of insight—requires a reward model that is itself highly sophisticated and tuned to the nuances of logical progression [16].

The training of these models often involves a hybrid approach where human-annotated process data is used to bootstrap the reward model, which then provides synthetic signals to the generative model during large-scale reinforcement learning runs [17]. This cycle creates a self-improving system, but it also introduces the risk of reward hacking, where the generative model learns to satisfy the criteria of the process-based reward model without actually improving its underlying logic [18]. To mitigate this, the architecture must incorporate multi-agent verification or diverse reward functions that evaluate the reasoning from multiple perspectives, such as factual accuracy, logical consistency, and linguistic coherence [19]. The goal is to create a robust training environment that resists the tendency of models to find shortcuts in complex reasoning tasks [20].

4. Infrastructure and Deployment Challenges

Deploying verifiable multi-step reasoning models into production environments presents substantial logistical hurdles. The increased complexity of the inference graph means that standard hardware configurations may struggle with the memory overhead and inter-process communication required for real-time verification [21]. Large-scale systems must be redesigned to support distributed inference where the reward model and the policy model can reside on separate nodes without introducing prohibitive latency. Furthermore, the sustainability of these systems is a growing concern [22]. The energy consumption associated with running multiple model passes for a single user query raises questions about the environmental impact of high-fidelity AI. Optimization at the hardware-software interface, such as specialized kernels for reward model integration, is necessary to make these architectures commercially and ecologically viable [23].

Beyond the physical hardware, the data infrastructure required to support process-based reward models is immense. Collecting high-quality reasoning chains that have been meticulously labeled for step-wise correctness is far more expensive than collecting simple

prompt-response pairs [24]. This has led to an increased reliance on synthetic data generation, where stronger models act as teachers for smaller models [25]. While this addresses the data scarcity problem, it creates a potential for systemic bias or the circular reinforcement of errors if the teacher model possesses its own logical blind spots. Establishing a reliable data governance framework that ensures the diversity and accuracy of the training corpus is a critical component of the deployment strategy [26]. Organizations must invest in long-term data curation efforts that involve domain experts to verify the ground truth of reasoning processes [27].

5. Verifiability and the Reduction of Hallucination

The primary technical motivation for process-based models is the reduction of hallucinations—instances where the model generates plausible-sounding but factually incorrect or logically inconsistent information. In multi-step reasoning, a single hallucination in the second step can derail a ten-step process, leading to a completely erroneous conclusion [28]. By implementing a process-based reward model, the system can catch these deviations as they occur. If a step fails to meet the verification threshold, the model can be programmed to backtrack and attempt an alternative path, much like a human researcher correcting a mistake in a draft. This iterative refinement process significantly increases the reliability of the final output [29].

Verifiability also has implications for the right to explanation in automated decision-making [30]. As AI systems are used to process insurance claims, evaluate creditworthiness, or assist in judicial research, the ability to audit the step-by-step logic becomes a legal and ethical necessity. A model that can prove its work is inherently more accountable than one that provides a black-box result. The challenge, however, is that verifiability is often subjective and dependent on the context of the task [31]. A logical step in a creative writing exercise is judged differently than a step in a mathematical proof. Architectures must therefore be flexible enough to handle different standards of verification across various cognitive domains, ensuring that the reward model is aligned with the specific requirements of the end-user [32].

6. Governance, Policy, and Socio-Technical Implications

The transition toward verifiable AI through process-based rewards is deeply intertwined with the broader landscape of algorithmic governance. As these systems become more capable, the pressure on regulators to define standards for logical sufficiency and procedural fairness increases [33]. Policy frameworks must evolve to address the complexities of multi-step reasoning, moving beyond simple accuracy metrics to evaluate the integrity of the process itself [34]. This involves establishing guidelines for how much human oversight is required during the training phase and what level of transparency must be provided to the end-user regarding the model's internal verification steps. There is a risk that the focus on process-based verification could lead to a transparency illusion, where the system provides a logical-looking path that is still fundamentally flawed, requiring sophisticated auditing tools to detect [35].

From a socio-technical perspective, the widespread adoption of these models could shift the

labor market for knowledge workers. If AI systems can reliably perform complex, multi-step reasoning, the role of the human expert may transition from creator to auditor. This requires a new set of skills focused on evaluating the reasoning chains produced by machines [2]. Furthermore, the question of fairness arises: if process-based reward models are trained primarily on Western logical traditions or specific academic datasets, they may inadvertently penalize alternative reasoning styles or cultural perspectives. Ensuring that the optimization process accounts for cognitive diversity is essential for creating equitable AI systems that serve a global population [10].

7. Robustness and Adversarial Resilience

A critical aspect of system-level discussion involves the robustness of process-based architectures against adversarial attacks. In traditional models, an adversary might attempt to find a specific jailbreak prompt that triggers an unethical response. In a multi-step reasoning model, the attack surface is much larger [14]. An adversary could attempt to inject subtle logical errors early in the reasoning chain that are designed to bypass the reward model's verification thresholds but ultimately lead to a compromised conclusion. Ensuring that the reward model is itself resilient to adversarial manipulation is a key area of ongoing research. This involves training the reward model on a wide range of near-miss reasoning chains and adversarial examples to improve its discriminative power [18].

Robustness also refers to the system's ability to maintain performance under distributional shift. If a model trained on medical reasoning is suddenly tasked with legal reasoning, its process-based reward model may no longer be accurate [12]. The architecture must include mechanisms for uncertainty quantification, where the system can signal when it is no longer confident in its ability to verify its own steps. This self-awareness is crucial for preventing over-reliance on AI in novel or high-stakes situations. By integrating reinforcement learning with Bayesian approaches to uncertainty, researchers can build models that know when to ask for human intervention, thereby enhancing the overall safety and reliability of the socio-technical system [4].

8. Sustainability and Resource Management in Large-Scale AI

The computational intensity of training and running process-based reward models cannot be ignored in the context of global sustainability goals. The carbon footprint of a single multi-step reasoning pass, when multiplied by millions of users, is substantial [22]. To address this, future architectures must prioritize efficiency as a core design principle. This might include the use of sparse transformers that only activate relevant portions of the network for specific reasoning steps, or the implementation of early exit strategies where the model stops reasoning once a sufficient level of certainty is reached. Reinforcement learning can be used to optimize not just for accuracy, but for a multi-objective function that includes energy efficiency as a key parameter [13].

Infrastructure providers also face the challenge of managing the lifecycle of these models. As new data becomes available, the reward models must be continuously updated to reflect the latest standards of knowledge and logic [26]. This requires a Continuous Integration and

Continuous Deployment pipeline for machine learning that can handle the complexities of multi-step verification. The policy implications of this are significant; who decides when a reward model needs to be updated, and what happens to the old reasoning standards? These are questions of governance that require collaboration between computer scientists, ethicists, and policymakers to ensure that AI development aligns with long-term societal interests and environmental constraints [19].

9. Case Studies and Empirical Observations

To ground the theoretical discussion, it is useful to examine specific applications of process-based reward models. In the field of automated theorem proving, for example, the use of step-wise verification has been instrumental in allowing models to solve complex mathematical problems that were previously out of reach [25]. By rewarding the model for every valid logical inference, researchers have been able to guide the generative process through vast search spaces. Similarly, in the domain of software engineering, process-based models are being used to audit the generation of code. Instead of just checking if the code runs, the reward model evaluates the architectural soundness and security of each function as it is written [30]. These cases demonstrate the tangible benefits of moving away from outcome-only evaluations.

Another relevant area is the use of multi-agent systems in biosecurity auditing. In these scenarios, one agent proposes a sequence of safety protocols, while a supervising agent—powered by a process-based reward model—audits each step for compliance with international regulations [8]. This creates a checks and balances system that is much more reliable than a single-agent approach. Empirical data from these deployments suggests that while the initial setup costs are higher, the long-term reduction in errors and the increased ease of auditing provide a strong return on investment. These observations highlight the importance of designing systems that are auditable by design rather than attempting to retroactively apply transparency measures to existing black-box models [32].

10. Cross-Domain Comparisons and Synthesis

When comparing the implementation of process-based rewards across different fields, such as financial forecasting and precision agriculture, distinct patterns emerge. In financial forecasting, the focus of the reward model is often on the mathematical validity of the trend analysis and the robustness of the data sources [8]. In contrast, in precision agriculture, the reasoning chain might involve interpreting sensor data and proposing irrigation strategies, where the verifiability is tied to physical constraints and biological models. Despite these differences, the underlying architectural requirement remains the same: a need for a modular, feedback-driven system that can handle sequential decision-making under uncertainty [16].

The synthesis of these diverse applications suggests that the future of large language model architectures lies in specialized verification. Rather than a single, general-purpose reward model, we are likely to see a library of domain-specific verifiers that can be plugged into a central generative core [1]. This plug-and-play architecture would allow for greater flexibility and easier updates as domain-specific knowledge evolves. It also facilitates a more nuanced

approach to fairness and bias, as verifiers can be tuned to the specific ethical considerations of a given field. This modularity is a hallmark of mature engineering systems and represents the next stage in the professionalization of artificial intelligence development [34].

11. Forward-Looking Perspectives on Cognitive Architectures

Looking ahead, the integration of process-based reward models is likely to lead to the development of system 2 thinking in artificial intelligence—a more deliberate, analytical, and self-correcting form of cognition [24]. Current models are often criticized for their system 1 nature, where they provide rapid, intuitive-sounding responses without deep reflection. By forcing the model to pass through multiple stages of verification, we are effectively slowing down the machine's thought process to ensure its quality. This has profound implications for the development of Artificial General Intelligence, as it provides a pathway toward models that can engage in long-term planning and complex problem-solving with a high degree of reliability [3].

The evolution of these architectures will also be shaped by the convergence of AI with other emerging technologies, such as quantum computing and edge processing [21]. Quantum-enhanced reward models could theoretically evaluate millions of reasoning paths simultaneously, while edge-based verification would allow for secure, private reasoning on local devices. As we move toward this future, the focus must remain on the human element. The goal of verifiable multi-step reasoning is not to replace human judgment but to provide tools that are more transparent, accountable, and aligned with human values [27]. The architectural choices we make today—prioritizing process over just outcome—will define the ethical and functional boundaries of the intelligent systems of tomorrow [35].

12. Conclusion

The optimization of process-based reward models through reinforcement learning represents a critical milestone in the development of verifiable and robust large language model architectures. By shifting the focus from final outcomes to the integrity of the reasoning path, researchers and engineers can address some of the most persistent challenges in AI, including hallucinations, lack of transparency, and adversarial vulnerability. However, this transition is not without its costs. The increased computational demand, the necessity for high-quality annotated data, and the complexities of governing multi-step reasoning processes require a holistic, system-level approach to design and deployment.

Ultimately, the success of these architectures will depend on our ability to balance technical performance with socio-technical responsibility. We must ensure that our systems are not only accurate but also fair, sustainable, and auditable. This requires ongoing collaboration between academia, industry, and the public sector to establish standards and policies that promote the ethical development of machine intelligence. As we continue to integrate AI into the fabric of our society, the commitment to verifiable reasoning will serve as a foundation for trust, allowing us to harness the power of large language models while mitigating the risks of their internal opacity. The journey toward more deliberate and transparent cognitive architectures is well underway, and the principles of process-based supervision will undoubtedly play a

central role in shaping the future of the field.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

5. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human feedback. Advances in Neural Information Processing Systems, 30.

6. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

8. Dou, Z., Zhao, Q., Wan, Z., Zhang, D., Wang, W., Raiyan, T., ... & Biswas, S. (2025). Plan Then Action: High-Level Planning Guidance Reinforcement Learning for LLM Reasoning. arXiv preprint arXiv:2510.01833.

9. Floridi, L., & Cowls, J. (2019). A unified framework of five-plus-one ethical principles for AI in society. Harvard Data Science Review, 1(1).

10. Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. International Conference on Machine Learning, 10949–10966.

11. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., ... & Steinhardt, J. (2021). Measuring mathematical problem solving with the MATH dataset. NeurIPS Datasets and Benchmarks Track.

12. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines.

Nature Machine Intelligence, 1(9), 389–399.

13. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

14. Lightman, H., Hunter, V., Kosaraju, V., Bavarian, M., Markosyan, N., Gominet, S., ... & Cobbe, K. (2023). Let's verify step by step. arXiv preprint arXiv:2305.20050.

15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Ziegler, D. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

16. Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

17. Shavit, Y., Agarwal, S., Brundage, M., Adler, S., & Campbell, R. (2023). Practices in governing agentic AI systems. OpenAI Policy Report.

18. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489.

19. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008–3021.

20. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.

21. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

22. Uesato, J., Kushman, N., Ramapuram, R., Figurnov, M., Huang, A., Lockwood, N., ... & Kohli, P. (2022). Solving math word problems with process-and outcome-based feedback. arXiv preprint arXiv:2211.14246.

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

24. Wang, X., Wei, J., Schuurmans, D., Quoc, L., Pang, B., Chi, E., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint

arXiv:2203.11171.

25. Wei, J., Wang, X., Schuurmans, D., Maeda, M., Zhao, T., Xia, V., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

26. Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Kak, A., ... & West, S. M. (2018). AI Now Report 2018. AI Now Institute at New York University.

27. Wu, Y., Remsing, R. C., Jozwik, M. G., Kocijan, S., Misra, S., Lin, J., ... & Goodman, N. D. (2023). Reasoning with language model is planning with a world model. arXiv preprint arXiv:2305.14992.

28. Yang, K., Klein, D., Russell, S., & Chen, A. (2024). Rewards-in-the-loop: Procedural optimization of reasoning chains. Journal of Artificial Intelligence Research, 79, 112–145.

29. Yao, S., Yu, D., Zhao, J., Shafran, I., McManus, T., Narasimhan, K., & Cao, Y. (2024). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36.

30. Zelikman, E., Wu, Y., Laskin, M., Snaider, J., Goodman, N. D., & Wu, C. (2022). Star: Bootstrapping reasoning with reasoning. Advances in Neural Information Processing Systems, 35, 15476–15488.

31. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. Advances in Neural Information Processing Systems, 32.

32. Zhang, M., & Li, J. (2023). Ethics of large language models in multi-step planning. Journal of Socio-Technical Studies, 15(2), 45-67.

33. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

34. Zhou, H., Li, C., & Wang, Y. (2024). Infrastructure for verifiable AI. ACM Computing Surveys, 56(4), 1-38.

35. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Christiano, P. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.