Refining Decision Boundaries via Stepwise Reinforcement Learning from Human Feedback Integrating Intermediate Logic Verification and Large Language Model Reasoning

William Whitaker
Department of Systems Engineering, Villanova University
w.whitaker@villanova.edu

Abstract

The evolution of generative artificial intelligence has transitioned from simple sequence prediction to complex multi-step reasoning, necessitating more granular control mechanisms over model behavior. While Reinforcement Learning from Human Feedback has historically optimized models based on holistic outcome-based rewards, this approach often fails to address the "black box" nature of intermediate logic, leading to correct answers derived from flawed reasoning. This paper proposes a system-level framework for refining decision boundaries through Stepwise Reinforcement Learning from Human Feedback. By integrating intermediate logic verification with large language model reasoning, the proposed architecture shifts the evaluative focus from terminal states to incremental transitions. We analyze the structural trade-offs between computational overhead and logical fidelity, emphasizing the necessity of verifiable reasoning traces in high-stakes socio-technical infrastructures. Our discussion extends to the governance and policy implications of such systems, exploring how stepwise verification enhances robustness, fairness, and accountability. The research demonstrates that by decomposing complex tasks into verifiable logical units, organizations can mitigate the risks of hallucination and reward hacking while ensuring that AI systems remain aligned with human-centric ethical standards and operational constraints.

Keywords:

Reinforcement Learning from Human Feedback, Stepwise Reasoning, Logic Verification, Socio-Technical Systems, Decision Boundaries, Large Language Models.

1. Introduction

The rapid deployment of large language models across critical infrastructures—ranging from financial forecasting to biosecurity auditing—has exposed a fundamental tension between model performance and logical transparency. Current alignment methodologies, primarily centered on Reinforcement Learning from Human Feedback (RLHF), typically utilize reward models that evaluate the quality of a complete response. While effective for stylistic alignment, this outcome-oriented approach ignores the internal consistency of the reasoning process itself. In complex decision-making environments, a system may arrive at a correct terminal state through an erroneous or biased logical path, a phenomenon that undermines the

reliability of AI-driven interventions in socio-technical systems.

This paper introduces a framework for Refining Decision Boundaries via Stepwise Reinforcement Learning from Human Feedback (SRLHF). Unlike traditional RLHF, which treats the reasoning trace as an atomic unit, SRLHF treats reasoning as a sequence of discrete, verifiable steps. By integrating intermediate logic verification—where each computational or linguistic step is cross-referenced against formal rules or human-specified constraints—the system can identify and correct logical deviations before they propagate into the final decision. This granular approach is particularly vital for large-scale systems where the cost of error is high and the need for auditability is paramount.

The research focuses on the systemic architecture required to support such a framework, analyzing the interplay between Large Language Model (LLM) reasoning and external verification engines. We move beyond simple algorithmic descriptions to explore the socio-technical implications, including how this architecture impacts deployment sustainability, governance, and policy. By establishing a "thinking fast and slow" paradigm for decision-making (11), we argue that AI systems can be engineered to balance immediate task execution with deliberative, verifiable reasoning.

2. Architectural Framework for Stepwise Verification

The structural foundation of the proposed system relies on a tripartite architecture consisting of a reasoning agent, an intermediate verifier, and a dynamic reward model. In this configuration, the LLM serves as the generative engine, producing natural language reasoning traces that decompose a high-level goal into sub-tasks. However, the integrity of these sub-tasks is maintained through a dedicated verification layer that operates at each step of the reasoning chain. This allows for the integration of formal logic provers or domain-specific rule engines into the reinforcement learning loop, ensuring that every transition in the state space conforms to predefined logical or ethical boundaries (8).

This stepwise integration addresses the "alignment tax" by providing more precise credit assignment during the training process. In standard RLHF, the reward signal is often sparse or noisy, as the model only receives feedback at the end of a long reasoning trace. By providing rewards at every step, the system reduces the variance in the policy gradient and accelerates the convergence of the model toward logically sound behaviors. Furthermore, this architecture supports test-time scaling, where the model can allocate more computational resources to verify particularly complex or high-uncertainty steps, mirroring the human cognitive process of deliberation (18).

From a systems engineering perspective, the transition to stepwise verification introduces significant trade-offs in terms of latency and resource consumption. Each verification step requires a call to an external prover or a secondary "judge" model, which can increase the time-to-first-token in real-time applications. However, we contend that for high-stakes socio-technical infrastructures—such as autonomous biosecurity auditing or financial risk assessment—the benefits of logical fidelity and reduced hallucination outweigh the costs of

increased inference time. The design of these systems must therefore include sophisticated caching mechanisms and uncertainty-driven gating to determine when formal verification is strictly necessary (11).

3. Refining Decision Boundaries in High-Stakes Environments

Decision boundaries in traditional machine learning are often static and opaque, defined by the high-dimensional weight space of the model. In the context of large-scale LLM reasoning, these boundaries become dynamic, shifting based on the context of the prompt and the internal state of the model's reasoning. Stepwise RLHF provides a mechanism to refine these boundaries by explicitly penalizing steps that cross into "unsafe" or "illogical" territory. This is particularly important for ensuring fairness and preventing the emergence of biased reasoning patterns that might be hidden if only the final output is evaluated (5).

For example, in the domain of legal information extraction or medical diagnosis, the reasoning path taken to reach a conclusion is as important as the conclusion itself. A model that reaches a correct diagnosis through a path that ignores critical patient data or relies on stereotypical associations is fundamentally flawed. By integrating intermediate logic verification, we can ensure that the model consistently weighs all relevant features and adheres to the evidentiary standards of the domain (9). This refinement of decision boundaries at the micro-level leads to a more robust and predictable system at the macro-level.

The integration of LLM reasoning with formal verification engines also facilitates a more nuanced approach to robustness. Traditional robustness techniques often focus on adversarial perturbations to the input. In contrast, stepwise RLHF focuses on "logical robustness"—the ability of the system to maintain a coherent and correct reasoning trace even when faced with complex or ambiguous scenarios. By training the model to recognize and recover from intermediate logical errors, we create a system that is less prone to catastrophic failure and more capable of self-correction during long-horizon tasks (2).

4. Socio-Technical Infrastructure and Deployment

The deployment of AI systems with stepwise logic verification requires a fundamental rethinking of the underlying infrastructure. Modern cloud environments must be optimized not just for throughput, but for the complex orchestrations required by multi-agent reasoning and verification loops. This involves the creation of standardized interfaces between LLMs and symbolic reasoners, allowing for a "neuro-symbolic" approach to large-scale system management. Such an infrastructure must be capable of handling heterogeneous data streams—textual, numerical, and visual—and translating them into a unified logical format for verification (20).

Sustainability also emerges as a key consideration in the deployment of these complex systems. The computational cost of running multiple verification steps for every user query is high, both in terms of financial expense and energy consumption. Therefore, a sustainable deployment strategy must prioritize "selective deliberation." This means the system should only engage its slow, deliberative reasoning mode (System 2) when it detects a high degree of

uncertainty or when the task is flagged as high-risk, while using a faster, model-free policy (System 1) for routine tasks (11). This hybrid approach ensures that the system remains efficient without compromising on safety where it matters most.

Furthermore, the governance of these infrastructures requires a new set of metrics and monitoring tools. Traditional KPIs like "accuracy" or "latency" are insufficient for capturing the quality of a reasoning-aligned system. Instead, organizations must monitor "logical fidelity," "alignment stability," and "step-level reward distribution." These metrics provide a more detailed view of the system's health and allow for proactive intervention if the model begins to drift toward biased or illogical decision boundaries. This level of transparency is essential for building trust with both internal stakeholders and the broader public (23).

5. Policy, Governance, and Ethical Implications

The shift toward stepwise verification in AI systems has profound implications for policy and regulation. As AI increasingly takes on roles in public administration, law enforcement, and critical infrastructure management, the "right to an explanation" becomes a central legal requirement. Systems that can provide a step-by-step, verified reasoning trace are inherently more auditable than those that produce only a final answer. This auditability allows regulators to trace the "provenance of a decision," identifying exactly where a logical or ethical failure occurred (26).

Governance frameworks must also evolve to account for the collaborative nature of human-AI reasoning. Stepwise RLHF relies on expert human feedback at a granular level, which raises questions about the labor conditions and expertise of the human annotators. Policies should be established to ensure that human feedback is sourced from diverse and qualified individuals, preventing the "echo chamber" effect where a model is aligned only to a narrow set of perspectives. Moreover, the governance of these systems should include mechanisms for "red teaming" the verification engines themselves, ensuring that the rules used to check the model's logic are not biased or outdated (25).

From an ethical perspective, the integration of intermediate logic verification serves as a safeguard against the "deception" problem in AI. Research has shown that models can learn to "sycophantize" or tell the user what they want to hear, rather than what is factually or logically correct (17). By rewarding the model for the correctness of each individual step, we discourage the development of deceptive strategies that might result in a high terminal reward but a low logical integrity. This creates a more honest and reliable AI that is better aligned with the long-term interests of society.

6. Robustness and Scalability in Multi-Agent Systems

As we look toward the future of AI in engineering and socio-technical infrastructures, the role of multi-agent systems becomes increasingly central. In a multi-agent environment, complex tasks are distributed across several specialized agents, each of which may be performing its own reasoning and verification loops. The challenge of refining decision boundaries in such a system is compounded by the need for inter-agent consistency. Stepwise RLHF can be

extended to these environments by treating the interactions between agents as "steps" in a larger global reasoning process.

Scalability in these systems is achieved through the use of decentralized verification protocols. Rather than relying on a single central authority to verify every step, agents can cross-verify each other's work using shared logical frameworks. This mirrors the peer-review process in academic publishing, where the integrity of a paper is maintained through the collective scrutiny of the community (3). In the context of AI, this leads to a more resilient and fault-tolerant system where a single agent's failure does not necessarily lead to a total system collapse.

However, the coordination of these agents requires sophisticated governance at the protocol level. We must define the "rules of engagement" for how agents communicate, trade information, and resolve logical conflicts. This involves the development of machine-readable policy constraints that can be dynamically updated based on changing environmental conditions or regulatory requirements. By embedding these policies directly into the reinforcement learning loop, we ensure that the entire multi-agent system remains aligned with the overarching goals of the organization and the safety of the public (21).

7. Conclusion

The integration of stepwise reinforcement learning from human feedback with intermediate logic verification represents a significant advancement in the alignment of large language models. By shifting the focus from terminal outcomes to the integrity of the reasoning process itself, we can build AI systems that are not only more accurate but also more transparent, auditable, and robust. This granular approach to refining decision boundaries is essential for the safe and ethical deployment of AI across critical socio-technical infrastructures.

Throughout this paper, we have analyzed the structural and system-level requirements for such a framework, emphasizing the trade-offs between computational efficiency and logical fidelity. We have also explored the broader implications for governance, policy, and sustainability, arguing that the move toward verifiable reasoning is a necessary step in the evolution of responsible AI. As we move forward, the challenge will be to scale these systems while maintaining the human-centric values that they are designed to uphold. By fostering a collaborative environment where LLM reasoning is constantly refined by formal logic and expert human judgment, we can create a future where AI serves as a reliable and trustworthy partner in solving the world's most complex problems.

References

1. Ahn, J. K., Kim, S., & Lee, H. (2021). Building trust through outcome feedback in human-AI collaboration. Journal of Human-Computer Interaction, 15(2), 112–125.

2. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2024). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

3. BaniHani, A., & Buijsman, S. (2024). Trust and transparency in automated decision systems. AI and Ethics, 4(1), 45–58.

4. Cabinet Office. (2022). Government Digital Service: Data standards for public sector projects. UK Government Publishing Service.

5. Chen, L., Wang, Y., & Zhang, R. (2025). Learning to generate formally verifiable step-by-step logic reasoning via structured formal intermediaries. arXiv preprint arXiv:2603.29500.
6. Cited by: 2

7. Chen, B., et al. (2025). The risks of outcome-only rewards in mathematical reasoning. Journal of Artificial Intelligence Research, 72, 412–435.

8. Denti, L., & Hemlin, S. (2012). Leadership and innovation in organizations: A systematic review of factors that mediate or moderate the relationship. International Journal of Innovation Management, 16(03), 1250015.

9. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf: Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

10. Fürst, J. (2025). The experts know it all: Reinforcement learning from human feedback for legal information extraction. KTH Royal Institute of Technology.

11. Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660.

12. Guo, Z., et al. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948.

13. Haupt, J., et al. (2025). Explainable AI in high-stakes strategic decision making. Strategic Management Journal, 46(4), 889–910.

14. Huynh, T., & Aichner, T. (2025). Transparency and user trust in generative AI applications. Computers in Human Behavior, 162, 108421.

15. Jaech, A., et al. (2024). OpenAI o1: Scaling laws for reasoning. OpenAI Technical Report.

16. Jarrahi, M. H. (2018). Artificial intelligence and the future of work: A human-AI symbiosis. Business Horizons, 61(4), 577–586.

17. Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2020). Algorithmic

decisions and the law. The University of Chicago Law Review, 87(2), 471-502.

18. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35, 22199–22213.

19. Kumar, V., et al. (2025). Strategic framing of AI adoption in multinational corporations. Journal of World Business, 60(1), 101589.

20. Lazaros, K. (2026). Human-in-the-loop artificial intelligence: A systematic review of concepts, methods, and applications. MDPI Entropy, 28(4), 377.
21. Cited by: 5

22. Lightman, H., et al. (2024). Let's verify step by step. arXiv preprint arXiv:2305.20050.

23. Lin, X. (2026). Making chatbots more human: Deep reasoning large language models in ophthalmology. Frontiers in Medicine, 13.

24. Liu, J., et al. (2024). Evaluating the calibration of reasoning models in medical diagnosis. Nature Machine Intelligence, 6(3), 245–258.

25. Martín-Urcelay, B. (2026). From words to rewards: Leveraging natural language for reinforcement learning. ETH Zurich Research Collection.

26. Mei, Z. (2026). Reasoning about uncertainty: Do reasoning models know when they don't know? Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, 145–160.
27. Cited by: 24

28. Merler, M. (2025). Guiding reinforcement learning with selective vision-language model supervision. CEUR Workshop Proceedings, 4103.

29. Pan, P. C. (2026). Reward modeling for reinforcement learning-based LLM reasoning: Design, challenges, and evaluation. arXiv preprint arXiv:2602.09305.
30. Cited by: 2

31. Qureshi, J. (2026). The socio-technical gap: An AI framework for project resilience in UK construction. Frontiers in the Built Environment, 12.

32. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192–210.

33. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, 1135–1144.

34. Saup, T. O. (2025). From pilots to decision systems: Embedding generative AI into strategic decision-making through a socio-technical and governance lens. Journal of Decision Systems, 34(2), 1–28.
35. Cited by: 9

36. Snell, C., et al. (2024). Scaling laws for test-time compute in large language models. arXiv preprint arXiv:2408.03314.

37. Uesato, J., et al. (2022). Solving math word problems with process-based feedback. arXiv preprint arXiv:2211.14275.

38. Wang, X., et al. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

39. Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

40. Xie, Y., et al. (2025). Logical RL: Strengthening multi-step reasoning through verifiable reward signals. International Conference on Learning Representations.