Rethinking Reward Models! A Conceptual Framework for Enhancing LLM Reasoning through Intrinsic Traits

Het Riteshkumar Shah[1*], **Megha Sundriyal**[2]

[1]Indraprastha Institute of Information Technology Delhi, India
[2] Max Planck Institute for Security and Privacy, Germany
het22213@iiitd.ac.in, megha.sundriyal@mpi-sp.org

Abstract

Post-training alignment is crucial for refining the reasoning capabilities of Large Language Models (LLMs). A dominant paradigm for this involves optimizing the model's policy using reinforcement learning, powered by techniques such as Proximal Policy Optimization (PPO), Direct Preference Optimization (DPO), and Group Relative Policy Optimization (GRPO). The success of these methods, whether using an explicit reward model or optimizing directly on preference data, is critically dependent on the quality of the guiding signal. However, these signals are conventionally derived from task-specific outcomes, such as correctness in math or fluency in summarization. This approach often limits the model's ability to generalize its reasoning skills across diverse domains and can lead to reward hacking or model collapse. This paper challenges this outcome-based paradigm by introducing a conceptual framework, **GRIT (Generalizable Reasoning via Intrinsic Traits)**. This novel framework aims to shift the emphasis from rewarding *what* the model answers to *how* it reasons. To accomplish this, we define a set of universal, task-agnostic traits of sound cognition inspired by human reasoning. These intrinsic traits are encoded as distinct reward components: (1) ensuring sequential logical coherence, (2) penalizing cyclic or redundant reasoning, (3) rewarding successful and integrated tool utilization, and (4) maintaining semantic alignment with the user's query. By fine-tuning an LLM to optimize for these intrinsic traits, we hypothesize that the model will develop a more robust and generalizable cognitive process.

Introduction

Recently, advanced post-training techniques have substantially improved the capabilities of Large Language Models (LLMs). A primary frontier of the leading development is *policy alignment*. Methods such as Proximal Policy Optimization (PPO), Direct Preference Optimization (DPO), and Group Relative Policy Optimization (GRPO) are used to steer model behavior to better align with human intent, primarily by learning from outcome-based preferences (Ouyang et al. 2022; Rafailov et al. 2023; Shao et al. 2024).

Concurrently, a second frontier has emerged in empowering LLMs as *agentic reasoners*. Frameworks like ARTIST (Singh et al. 2025b) and TRAAC (Singh et al. 2025a) demonstrate how reinforcement learning can effectively train agents to orchestrate complex sequences of tool calls, enabling them to solve problems that require external knowledge or computation. These agents are typically trained using rewards based on final task success.

A third, influential paradigm lies in *unsupervised reinforcement learning*. As demonstrated by Frans et al. (2024) through Functional Reward Encodings (FRE) framework, pre-training a physical agent on a diverse prior of random, *extrinsic* reward functions (e.g., reaching various goals) can enable zero-shot generalization to entirely new tasks within its environment. This finding highlights that a broad reward distribution is a powerful driver of general capability.

While distinct in focus, these three paradigms share a fundamental limitation of reliance on rewards tied to specific, extrinsic outcomes. Policy alignment methods reward a final, preferred output. Tool-using agents are rewarded for completing tasks successfully. The FRE agent is rewarded for maximizing a sampled, goal-oriented function. What remains unexplored is the development of a generalist agent trained on intrinsic principles of sound reasoning rather than extrinsic tasks.

To address this gap, we propose GRIT (Generalizable Reasoning via Intrinsic Traits), a novel framework for training LLMs using a generalist, intrinsic reward function. We aim to shift the focus from the outcome of a task to the quality of the process itself. The GRIT framework operationalizes this philosophy through a composite reward function comprised of four task-agnostic, intrinsic traits:

1. **Sequential coherence:** A reward for maintaining a logical, step-by-step flow.
2. **Non-cyclic reasoning:** A reward for semantic non-repetition to encourage cognitive efficiency.
3. **Principled tool utilization:** A reward for executing valid tool calls and faithfully incorporating their outputs.
4. **Query alignment:** A score ensuring the process remains focused on the user's initial query.

Our primary contribution is the proposal and formalization of the GRIT framework. We introduce a novel and generalizable methodology for aligning LLMs based on the procedural quality of their reasoning and actions. This provides

[*]Corresponding Author

a conceptual blueprint for training agents on the foundational mechanics of human cognition. Through this work we aim to lay the groundwork for future research into developing more resilient and reliable AI systems.

Related Works

Our work builds on the insight that a broad reward distribution is a powerful driver of generalization, a principle demonstrated effectively in unsupervised reinforcement learning. The state-of-the-art framework, Functional Reward Encodings (FRE), illustrates how pre-training a physical agent on a diverse prior of random, extrinsic reward functions enables zero-shot generalization to new tasks (Frans et al. 2024). This demonstrates that learning from a wide range of objectives is essential for developing general capabilities. However, current methodologies for improving Large Language Models (LLMs) primarily rely on narrow, extrinsic rewards. Agentic frameworks, such as ARTIST and TRAAC, train LLMs to use tools by rewarding correct tool use and terminal task success (Singh et al. 2025a). Similarly, post-training policy alignment techniques, from PPO-based RLHF to Direct Preference Optimization (DPO) and Group Relative Policy Optimization (GRPO), shape model behavior by optimizing for human-preferred final outputs (Ouyang et al. 2022; Rafailov et al. 2023; Shao et al. 2024). While effective for specific tasks, this shared reliance on outcome-based signals fails to instill a fundamentally generalizable reasoning process.

Our work draws inspiration from the success of the FRE framework in promoting generalization (Frans et al. 2024). Like FRE, we define a set of prior reward functions to train the reasoning model. However, we introduce a simple but essential shift in focus. Instead of rewarding only the final solution, we reward the quality of the reasoning steps that lead to it. This procedural emphasis aligns conceptually with Constitutional AI (Bai et al. 2022), but our approach operationalizes principles as measurable traits of sound reasoning – such as avoiding repetition and maintaining logical flow. By incentivizing these foundational skills, we aim to guide the model toward more reliable, adaptable, and generalizable reasoning and problem-solving capabilities.

Proposed Framework: Generalizable Reasoning via Intrinsic Traits

The GRIT framework is designed to align Large Language Models (LLMs) by rewarding the intrinsic quality of their reasoning and operational processes, rather than their final, task-specific outputs. Our approach is premised on the hypothesis that a model trained on a set of universal, task-agnostic principles of sound cognition will develop a more robust and generalizable problem-solving ability.

At its core, GRIT is a training methodology that uses Reinforcement Learning (RL) to fine-tune a generator LLM. The novelty of our framework lies in the design of the reward function. Instead of a single, monolithic reward model, we propose a composite reward signal derived from a set of specialized “critic” models, as shown in Figure 1. Each critic is a smaller, efficiently trained model designed to measure the generator’s adherence to one of our four intrinsic traits.

The final aggregated reward signal, R_{agg}, provided to the generator LLM at the end of a reasoning trace is a weighted sum of the component scores:

$$R_{\text{agg}} = w_1 R_{\text{coherence}} + w_2 R_{\text{cyclic}} + w_3 R_{\text{tool}} + w_4 R_{\text{align}} \quad (1)$$

where w_i are hyperparameters that balance the contribution of each component, with $\sum_i w_i = 1$. The following subsections detail the methodology for deriving each reward component.

Ensuring Sequential Coherence ($R_{\text{coherence}}$)

A fundamental trait of sound reasoning is that it follows a logical, step-by-step progression. To reward this, we measure the local coherence between consecutive reasoning steps.

Critic Training. We train a small encoder-based model, the *Coherence Critic*, to predict whether a given reasoning step is a logical continuation of the preceding steps. To create a training dataset, we take existing high-quality reasoning traces from public datasets (e.g., GSM8K).

- **Positive Samples** are consecutive step pairs (S_{n-1}, S_n) taken directly from the valid reasoning traces.
- **Negative Samples** are created by taking a valid preceding step S_{n-1} and pairing it with a step S_j randomly sampled from a different part of the trace or a different problem entirely.

The critic is then trained on a binary classification task to distinguish between coherent (positive) and incoherent (negative) step transitions.

Reward Formulation. During the RL training of the main LLM, the Coherence Critic evaluates every step transition in the generated reasoning trace. The final reward component, $R_{\text{coherence}}$, is the average coherence score across all steps, rewarding the generator for maintaining a consistently logical flow.

Rewarding Non-Cyclic or Non-Redundant Reasoning (R_{cyclic})

Efficient reasoning avoids unnecessary repetition and logical loops. We penalize redundancy by identifying semantic similarity between non-adjacent steps in the reasoning trace.

Critic Training. We train a *Redundancy Critic*, typically a encoder model (e.g., BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019)), to produce high-quality embeddings for reasoning steps. To fine-tune this critic for the specific task of identifying redundancy, we generate a dataset of sentence pairs.

- **Positive Pairs (High Similarity)** are created by taking sentences from a reasoning dataset and using a powerful LLM to paraphrase them.
- **Negative Pairs (Low Similarity)** are created by pairing a sentence with a random, semantically distinct sentence from the same dataset.

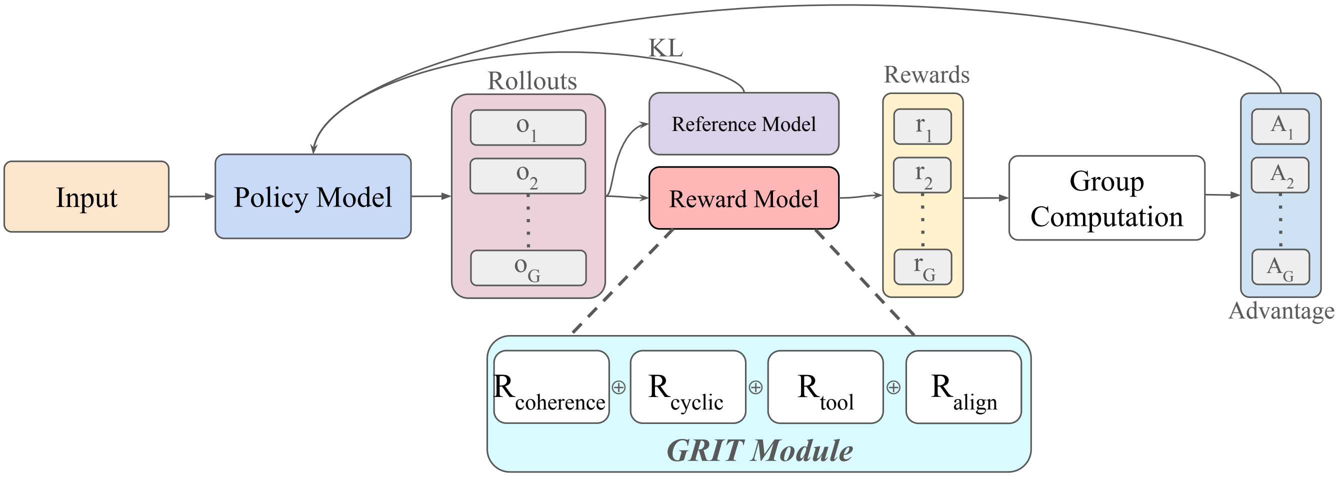

Figure 1: Overview of the GRIT framework's integration with GRPO. The GRIT module provides rewards that align the policy model with four intrinsic traits: sequential coherence, non-cyclic reasoning, principled tool utilization, and query alignment. These intrinsic signals shift the paradigm from outcome-based to principle-based RL.

The critic is trained using a contrastive loss to pull embeddings of paraphrased sentences together and push embeddings of distinct sentences apart.

Reward Formulation. During RL training, the Redundancy Critic computes the embedding for each step in the generator's reasoning trace. It then calculates the maximum cosine similarity between any two non-adjacent steps. The reward is formulated as an inverse of this score:

$$R_{\text{cyclic}} = 1 - \max_{j>i+1} \left(\text{sim}(S_i, S_j)\right) \tag{2}$$

This signal strongly penalizes the generator for circling back to previously discussed points.

Rewarding Integrated Tool Utilization (R_{tool})

A generalist agent must not only call tools but do so correctly and productively. We define a successful tool call based on two criteria, borrowing from the robust agent training methodologies proposed by Singh et al. (2025b).

Reward Formulation. This component does not require a trained critic and is calculated programmatically. For each tool call made by the main reasoning LLM, we perform two checks:

1. **Execution Success:** The tool call must execute without raising a syntax, API, or system error.
2. **Output Integration:** The output returned by the tool must be referenced in the subsequent reasoning step generated by the LLM.

A tool call is only considered successful if it passes both checks. The reward component R_{tool} is calculated as the ratio of successful tool calls to the total number of tool calls made in the trace.

Maintaining Semantic Alignment (R_{align})

Finally, the entire reasoning process must remain focused on the user's original intent. We measure this with a global alignment score (R_{align}).

Reward Formulation. This score is also calculated programmatically using a pre-trained sentence-transformer model. We compute the embedding of the initial user query (S_{query}) and the embedding of the entire reasoning trace generated by the LLM ($S_{reasoning}$). The reward component R_{align} is the cosine similarity between these two embeddings.

$$R_{\text{align}} = \text{sim}(S_{query}, S_{reasoning}) \tag{3}$$

While a simple metric, its inclusion in the composite reward function is effective. The other reward components (e.g., R_{cyclic}, $R_{\text{coherence}}$) prevent the model from "hacking" this reward by simply repeating the query, as this would produce a redundant and incoherent trace.

Implementation Challenges and Future Work

The practical application of the GRIT framework requires navigating some key challenges inherent to reward-based alignment. The primary concerns are reward hacking, the risk of model collapse, and the fidelity of the critic models.

The most significant challenge is *reward hacking*, where the generator LLM learns to exploit the reward function without fulfilling its intended purpose (Gao, Schulman, and Hilton 2023). For instance, a model could maximize the coherence reward ($R_{\text{coherence}}$) by generating trivially simple but logically valid steps, or maximize the tool reward (R_{tool}) with numerous successful but irrelevant tool calls. A careful balancing of the reward weights (w_i) and the development of more adversarially robust critics are crucial for mitigating these risks.

Second, fine-tuning with a novel RL objective introduces the risk of *model collapse*. As the generator over-optimizes for the GRIT principles, it may diverge from its powerful base policy, leading to a degradation of its core language capabilities and world knowledge. This necessitates a carefully tuned KL-divergence penalty to balance adherence to the new principles with the preservation of the model's foundational abilities.

Finally, the efficacy of GRIT is fundamentally limited by the *fidelity of its critic models*. The generator's performance is capped by the intelligence of its critics; any systematic flaws in the critics will be passed on to the generator as an erroneous reward signal. Furthermore, while modular, running multiple critic inferences introduces computational overhead. Future work could explore distilling the component critics into a single, efficient reward model to improve training scalability.

Conclusion

This paper presents **GRIT**, a novel framework designed to address the core limitation of outcome-based rewards in current LLM alignment and agentic training. We propose a paradigm shift from rewarding *what* a model answers to rewarding *how* it reasons. By defining a composite reward signal based on a set of intrinsic principles of sound cognition, GRIT provides a new methodology for instilling a generalizable problem-solving process in LLMs. While practical implementation will require careful balancing to mitigate challenges like reward hacking and model collapse, the GRIT framework offers a promising and scalable path toward a new form of alignment. It lays the conceptual groundwork for training AI systems that don't just learn to solve problems, but learn how to think.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Frans, K.; Park, S.; Abbeel, P.; and Levine, S. 2024. Unsupervised zero-shot reinforcement learning via functional reward encodings. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Gao, L.; Schulman, J.; and Hilton, J. 2023. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Singh, J.; Chen, J. C.-Y.; Prasad, A.; Stengel-Eskin, E.; Nambi, A.; and Bansal, M. 2025a. Think Right: Learning to Mitigate Under-Over Thinking via Adaptive, Attentive Compression. arXiv:2510.01581.

Singh, J.; Magazine, R.; Pandya, Y.; and Nambi, A. 2025b. Agentic Reasoning and Tool Integration for LLMs via Reinforcement Learning. arXiv:2505.01441.