

Towards Hierarchical Multi-Step Reward Models for Enhanced Reasoning in Large Language Models

Teng Wang¹, Zhangyi Jiang², Zhenqi He¹,
Wenhan Yang¹, Yanan Zheng³, Zeyu Li⁴,
Zifan He², Shenyang Tong², Hailei Gong^{5*}

¹ the University of Hong Kong, ² Peking University,
³ National University of Singapore, ⁴ Georgia Institute of Technology,
⁵ Tsinghua University

wt0318@connect.hku.hk

Abstract

Recent studies show that Large Language Models (LLMs) achieve strong reasoning capabilities through supervised fine-tuning or reinforcement learning. However, a key approach, the Process Reward Model (PRM), suffers from reward hacking, making it unreliable in identifying the best intermediate steps. In this paper, we propose a novel reward model approach, **Hierarchical Reward Model (HRM)**, which evaluates both individual and consecutive reasoning steps from fine-grained and coarse-grained level. HRM performs better in assessing reasoning coherence and self-reflection, particularly when the previous reasoning step is incorrect. Furthermore, to address the inefficiency of autonomous generating PRM training data via Monte Carlo Tree Search (MCTS), we introduce a lightweight and effective data augmentation strategy called **Hierarchical Node Compression (HNC)** based on node merging (combining two consecutive reasoning steps into one step) in the tree structure. This approach diversifies MCTS results for HRM with negligible computational overhead, enhancing label robustness by introducing noise. Empirical results on the PRM800K dataset demonstrate that HRM, in conjunction with HNC, achieves superior stability and reliability in evaluation compared to PRM. Furthermore, cross-domain evaluations on MATH500 and GSM8K confirm HRM’s superior generalization and robustness across diverse reasoning tasks. The code for all experiments will be released at https://github.com/tengwang0318/hierarchical_reward_model.

1 Introduction

As the scale of parameters in Large Language Models (LLMs) continues to grow [1, 2, 3, 4], their general capabilities have significantly improved, surpassing human performance in various generative tasks such as text comprehension and data generation [5]. However, the upper bound and inherent limitations of LLMs in reasoning-intensive tasks—such as mathematical and logical reasoning—remain an open question [6, 7, 8, 9, 10, 11, 12]. Recent approaches, such as Chain-of-Thought (CoT)[13] and Tree-of-Thought (ToT)[14], have significantly enhanced reasoning performance. Despite these advancements, CoT lacks a mechanism to halt reasoning when an intermediate step is incorrect, leading to error propagation. Meanwhile, ToT does not inherently verify every intermediate step

*Corresponding author

| Feature | ORM | PRM | HRM |
|--------------------|------------------|-------------|-----------------------|
| Scoring Method | Rule-Based or RM | RM Only | RM Only |
| Granularity | Whole Process | Single Step | Few Consecutive Steps |
| Step-wise Feedback | No | Yes | Yes |
| Error Correction | Yes | No | Yes |

Table 1: Comparison of scoring methods, granularity, and feedback mechanisms across ORM, PRM, and HRM.

or guarantee retrieval of the optimal reasoning trajectory, which can limit its reliability in complex problem-solving scenarios.

To mitigate these limitations, recent efforts have focused on reward mechanisms that guide LLMs more effectively. There are two primary approaches to enhancing the reasoning capabilities of LLMs from the perspective of "how to reward LLMs": the Outcome Reward Model (ORM)[7, 8, 15, 16] and the Process Reward Model (PRM)[7, 8]. Each comes with its own limitations. ORM suffers from delayed feedback and credit assignment issues, while PRM is prone to reward hacking and incurs high costs for reasoning process annotation.

In this paper, we focus on addressing the limitations of PRM. To mitigate the impact of reward hacking in PRM, we propose the **Hierarchical Reward Model (HRM)**. Unlike PRM, which evaluates reasoning at a fine-grained step level, HRM considers both fine-grained and coarse-grained reasoning, enabling the reward model to reflect on its own judgments and assess multi-step reasoning coherence. Traditional PRM penalizes an incorrect step without accounting for potential corrections in subsequent reasoning. In contrast, HRM enables the reward model to identify and incorporate later steps that rectify earlier errors, resulting in a more robust and reliable evaluation. Table 1 compares the difference between ORM, PRM and HRM. The PRM800K [7] dataset comprises manually annotated reasoning trajectories, which serve as the foundation for training ORM, PRM, and HRM. We subsequently assess the performance of Qwen2.5-72B-Math-Instruct [4] as the policy model by employing the Best-of-N Search strategy across ORM, PRM, and HRM. Experimental results show that HRM is the most robust to reward hacking, with accuracy stabilizing at 80% as N increases. In contrast, PRM and ORM exhibit significant performance fluctuations, with accuracy degrading as N grows.

To fully exploit the capabilities of MCTS, we introduce a data augmentation framework termed **Hierarchical Node Compression (HNC)**, which consolidates two consecutive nodes from different depths into a single node. This approach effectively expands the training dataset while maintaining minimal computational overhead and enhancing label robustness through controlled noise injection. After evaluating HNC in the auto-annotation process by MCTS on the PRM-800K dataset, we find that fine-tuned HRM achieves more robust scoring within PRM-800K and exhibits strong generalization across other domains, including GSM8K [6] and MATH500 [7] dataset, outperforming PRM in robustness and consistency.

Our main contribution are as follows:

- We propose the **Hierarchical Reward Model (HRM)**, which enhances the reward model’s multi-step reasoning coherence and self-reflection by evaluating not only fine-grained reasoning steps but also coarse-grained reasoning trajectories. We validate HRM’s robustness on the PRM800K dataset using manually annotated data.
- We introduce **Hierarchical Node Compression (HNC)** in MCTS for autonomous annotation, enhancing reward model training with more diverse reasoning data while incorporating controlled noise to improve score robustness. Additionally, by filtering high-quality reasoning trajectories from MCTS, we refine the policy model through fine-tuning, further enhancing its reasoning performance.
- We investigate the generalization of HRM trained on PRM800K using auto-labeled reasoning processes from MCTS and HNC. Experimental results demonstrate that HRM achieves superior reasoning consistency and generalizes effectively across GSM8K and MATH500, outperforming PRM.

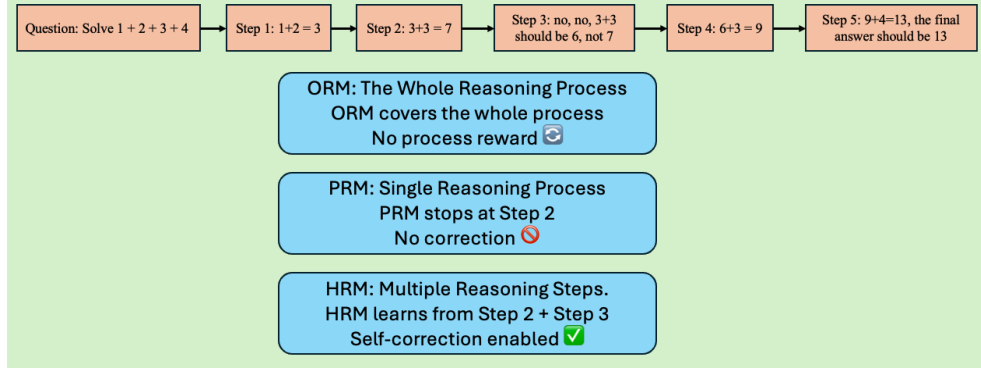


Figure 1: Illustration of how ORM, PRM, and HRM handle reasoning processes. ORM evaluates the entire reasoning chain, PRM assesses individual steps but stops at errors, and HRM considers multiple consecutive steps, enabling error correction.

2 Related Work

2.1 RLHF

Reinforcement Learning with Human Feedback (RLHF) [17] is a widely used framework for optimizing LLMs by incorporating human feedback signals. The core idea of RLHF is to use a Reward Model (RM) to distinguish between high-quality and low-quality responses and then optimize the LLM using Proximal Policy Optimization (PPO) [18] within a reinforcement learning framework. Ouyang et al. [17] first apply RLHF to train InstructGPT [17], aligning the model’s outputs with human preferences.

From the perspective of reward design, there are two main approaches: Outcome Reward Models (ORM) [6, 7, 9] and Process Reward Models (PRM) [7, 8, 10, 19]. ORM assigns rewards based on the whole output, while PRM evaluates intermediate reasoning steps to provide more fine-grained supervision. These reward mechanisms directly impact how LLMs learn to reason and optimize their outputs.

2.2 ORM

ORM suffers from delayed feedback and the credit assignment problem. Since rewards are only provided at the final outcome, ORM struggles to discern which intermediate steps contributed to success or failure [6, 7, 9]. This delayed feedback limits learning efficiency, making it harder to optimize critical decision points. Additionally, ORM is prone to spurious reasoning [6, 7, 9], where the model arrives at the correct answer despite flawed intermediate steps, reinforcing suboptimal reasoning patterns. However, DeepSeek-R1[15] integrates a rule-based ORM within the Group Relative Policy Optimization (GRPO) algorithm[16], demonstrating that rule-based reward mechanisms, rather than score-based reward models, can effectively guide LLMs toward generating long-chain-of-thought (long-CoT) reasoning and self-reflection, ultimately enhancing their reasoning abilities.

2.3 PRM

One of the most critical challenges in PRM is reward hacking, a phenomenon in which an RL agent exploits flaws or ambiguities in the reward function to achieve artificially high rewards without genuinely learning the intended task or completing it as expected [20, 21, 22, 23].

In the LLM domain, Wang et al. [24] find that when an LLM is used as a verifier to assess the quality of multiple outputs, its ranking can be easily manipulated simply by changing the order of candidates in the context. Wen et al. [25] demonstrate that RLHF can make AI models more persuasive to human evaluators without necessarily improving their correctness, leading to higher human approval of incorrect answers and increased evaluation error rates. Pan et al. [26] analyze In-Context Reward Hacking (ICRH), where an AI system engaged in a feedback loop optimizes a proxy objective—such

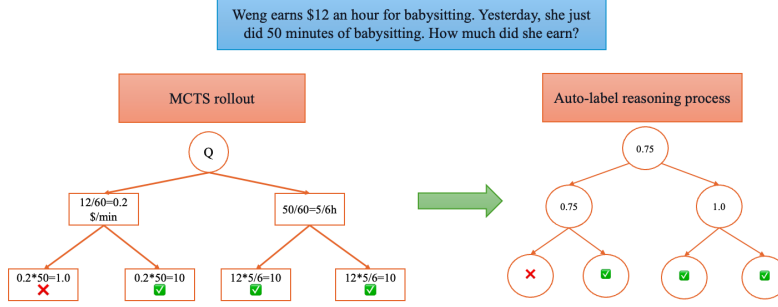


Figure 2: Illustration of Hierarchical Node Compression. The left part represents the original MCTS data annotation structure, while the right part shows the transformed MCTS structure after applying HNC.

as a writing assistant maximizing "user satisfaction scores"—and inadvertently learns to generate exaggerated claims or misinformation if these result in higher perceived satisfaction.

Furthermore, the annotation process required for training PRM models is prohibitively expensive [7, 8], making large-scale implementation impractical. To address this issue, Monte Carlo Tree Search (MCTS) has been proposed as an autonomous method for annotating reasoning trajectories [10, 19]. While MCTS reduces the need for human annotation, it incurs substantial computational costs due to the extensive simulations required for the MC-Score to achieve relative convergence. Moreover, in MCTS, the computational cost increases significantly as the depth and breadth of the search tree expand. To mitigate this, constraints are imposed on both tree height and width, limiting the number of simulation steps and thereby reducing the diversity of generated reasoning data.

3 Methodology

3.1 Hierarchical Reward Model

PRM [7, 8] focuses on fine-grained, step-wise reasoning, while ORM [7, 8] evaluates reasoning as a whole. To integrate the strengths of both, we propose the **Hierarchical Reward Model (HRM)**, which evaluates individual reasoning steps while also ensuring multi-step coherence by analyzing consecutive steps. The training dataset for HRM consists of consecutive reasoning sequences spanning from step 1 to N , as illustrated in Fig. 1 and Section 4.1.

HRM is designed with two primary objectives: (1) capturing both fine-grained and coarse-grained reasoning consistency, and (2) enabling self-reflection and error correction. Unlike PRM, which terminates evaluation upon encountering an error, HRM assesses whether subsequent steps rectify earlier mistakes, treating them as a cohesive unit rather than isolated errors.

3.2 Hierarchical Node Compression in MCTS

Although process supervision enhances the reasoning capabilities of policy models, the cost of human-annotated supervision is prohibitively high [7, 8]. To address this, autonomous annotation methods based on MCTS have been proposed [10, 19]. Fig. 2 illustrates the process of automatic reasoning annotation using MCTS. Given a ground truth and a corresponding question, MCTS generates multiple possible reasoning paths by simulating different step-by-step solutions. Each node in the search tree represents a reasoning step, and its score is calculated based on the proportion of correct steps in its subtree, reflecting the likelihood that the reasoning path is valid. However, these methods demand substantial computational resources, as ensuring the convergence of intermediate reasoning step scores requires a sufficiently deep and wide search tree; otherwise, the estimates remain biased. This exponential growth in complexity makes large-scale implementation challenging.

To address the limitations of autonomous process annotation, we propose a data augmentation method called **Hierarchical Node Compression (HNC)**, which maximally leverages MCTS-generated data. The key idea is to merge two consecutive nodes, each corresponding to a reasoning step, into a single node, thereby creating a new branch with minimal computational overhead.

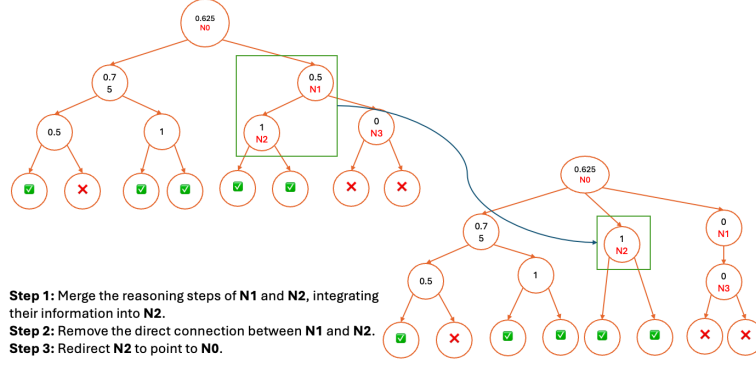


Figure 3: Illustration of Hierarchical Node Compression. The left part represents the original MCTS data annotation structure, while the right part shows the transformed MCTS structure after applying HNC.

As illustrated in Fig. 3, HNC assumes that each node has a sufficiently large number of child nodes. By randomly removing or merging consecutive nodes, it introduces controlled noise, enhancing the robustness of MCTS-based scoring. For instance, if each node initially has five child nodes contributing equally (20%) to the parent node’s score, merging one child node into a new branch redistributes the weights among the remaining four, increasing their contributions to 25%. This controlled perturbation improves the resilience of the scoring mechanism while expanding the dataset with minimal computational cost.

3.3 Self-Training

To filter high-quality reasoning data from MCTS, we adapt two approaches: using the MC-Score or leveraging PRM/HRM to assign scores. To mitigate reward hacking caused by the reward model, we apply a high-quality data filter based on MC-Score.

Due to computational constraints, we do not employ RL methods such as PPO [18] or GRPO [16]. Instead, we continue using supervised fine-tuning. To preserve the general capabilities of the policy model, we incorporate causal language modeling loss combined with KL divergence regularization using a reference model. The objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \log D_{\text{KL}}(P||Q), \quad (1)$$

where \mathcal{L}_{LM} represents the causal language modeling loss computed on high-quality reasoning sequences, and $D_{\text{KL}}(P||Q)$ denotes the KL divergence between the policy model’s output distribution P and the reference model’s output distribution Q . The term λ serves as a weighting factor to balance task-specific adaptation and retention of general capabilities.

Without proper KL regularization or with an insufficiently weighted KL loss (i.e., a very small λ), the KL divergence grows unbounded during training. Specifically, KL loss typically ranges from 0 to 20000, whereas the causal LM loss remains within 0 to 12, leading to a severe loss imbalance. This causes the optimization process to excessively minimize KL divergence at the expense of task-specific reasoning performance.

To address this, we apply a logarithmic scaling to $D_{\text{KL}}(P||Q)$, stabilizing the loss landscape and ensuring a balanced trade-off between preserving general language capabilities and enhancing reasoning ability. Further details are provided in Section 4.3.

4 Experiment

4.1 HRM

Given that the PRM800K dataset [7] consists of Phase1 and Phase2, where Phase1 includes manually annotated reasoning processes, we utilize these manual annotations to construct the training datasets

| Previous Step Label | Current Step Label | Label for Merged Step |
|---------------------|--------------------|-----------------------|
| Positive | Positive | Positive |
| Positive | Neutral/Negative | Negative |
| Neutral/Negative | Positive | Positive |
| Neutral/Negative | Neutral/Negative | Negative |

Table 2: Labeling strategy for constructing the HRM training dataset from manual annotations in PRM800K. PRM800K contains three label types: Positive, Negative, and Neutral. HRM extends PRM by incorporating multi-step reasoning.

| N | 2 | 4 | 8 | 16 | 24 |
|------------|--------------|--------------|--------------|------------|------------|
| ORM | 0.622 | 0.677 | 0.655 | 0.655 | 0.633 |
| PRM | 0.7 | 0.644 | 0.611 | 0.588 | 0.577 |
| HRM | 0.722 | 0.711 | 0.744 | 0.8 | 0.8 |

Table 3: Accuracy of Qwen2.5-72B-Math-Instruct [4] evaluated under the best-of- N strategy using ORM, PRM, and HRM, which are fine-tuned on manually labeled PRM800K data.

for ORM, PRM, and HRM. ORM training data comprises complete reasoning trajectories, while PRM training data consists of individual reasoning steps conditioned on preceding context. HRM training data extends PRM by incorporating multiple consecutive reasoning steps, allowing HRM to capture self-reflection and ensure reasoning coherence across sequential steps. Table 2 summarizes the labeling rules for merged reasoning steps in HRM.

We fine-tune Qwen2.5-1.5B-Math [4] as the reward model (RM) for classifying current step as correct or incorrect. Given an input reasoning step, RM predicts logits for the *positive* and *negative* classes, denoted as l_{positive} and l_{negative} , respectively. The confidence score is obtained by applying the softmax function:

$$P(y = \text{positive} \mid x) = \frac{\exp(l_{\text{positive}})}{\exp(l_{\text{positive}}) + \exp(l_{\text{negative}})}, \quad (2)$$

where $P(y = \text{positive} \mid x)$ denotes the probability of current step being correct. This probability serves as the reward score assigned by RM. Detailed training information is provided in Appendix A.1.

To evaluate the performance of ORM, PRM, and HRM, we employ Qwen2.5-72B-Math-Instruct [4] as the policy model and implement the best-of- N strategy. Specifically, ORM selects the best result from N complete reasoning trajectories, while PRM and HRM score N intermediate reasoning steps and select the most promising one at each step. For PRM and HRM, we consider the completion of a formula as an intermediate reasoning step, enabling a finer-grained evaluation mechanism. Table 3 presents the results, showing that ORM and PRM exhibit significant fluctuations, with accuracy decreasing as N increases. In contrast, HRM maintains stable performance, converging to an accuracy of 80% as N grows. Detailed information is provided in Appendix A.1

4.2 HNC

In this section, we utilize only the questions and ground truth from the PRM800K dataset [7], without relying on manually annotated data. We employ MCTS with Qwen2.5-7B-Math-Instruct [4] as an automatic annotation method to generate reasoning trajectories. As mentioned in Section 3.2, these auto-annotated reasoning trajectories from MCTS are used to train PRM, after which we apply the HNC data augmentation method to generate additional training data for HRM.

To balance computational efficiency and robustness, we configure MCTS with 5–6 child nodes per parent and a maximum tree depth of 7, ensuring reasoning completion within 7 steps. Since the computational cost of MCTS rollouts grows exponentially with tree depth and branching factor, we limit these parameters to maintain feasibility. The full MCTS simulation requires approximately 2,457 A100-80GB GPU-hours, while the HNC augmentation process takes around 30 minutes.

We perform supervised fine-tuning of Qwen2.5-1.5B-Math [4] for both PRM and HRM. To evaluate performance, we employ different policy models, including Qwen2.5-7B-Math-Instruct [4],

| Policy Model | Method | N | | | | | | | | | |
|------------------|--------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | 2 | 4 | 8 | 16 | 24 | 32 | 64 | 128 | 256 | 512 |
| DeepSeek-Math | PRM | 0.311 | 0.433 | 0.377 | 0.455 | 0.411 | 0.455 | 0.466 | 0.444 | 0.377 | 0.377 |
| | HRM | 0.311 | 0.388 | 0.444 | 0.455 | 0.455 | 0.422 | 0.533 | 0.522 | 0.455 | 0.500 |
| Qwen2.5-72B-Math | PRM | 0.233 | 0.344 | 0.411 | 0.422 | 0.488 | 0.522 | 0.600 | 0.566 | 0.666 | 0.700 |
| | HRM | 0.288 | 0.366 | 0.366 | 0.488 | 0.511 | 0.611 | 0.622 | 0.611 | 0.711 | 0.722 |
| Qwen2.5-7B-Math | PRM | 0.477 | 0.466 | 0.600 | 0.544 | 0.633 | 0.677 | 0.733 | 0.677 | 0.700 | 0.722 |
| | HRM | 0.500 | 0.566 | 0.655 | 0.600 | 0.666 | 0.711 | 0.711 | 0.766 | 0.777 | 0.766 |

Table 4: Accuracy of PRM and HRM under different policy models using the best-of- N strategy on the PRM800K dataset. The training data for both PRM and HRM are derived from MCTS with Qwen2.5-7B-Math-Instruct.

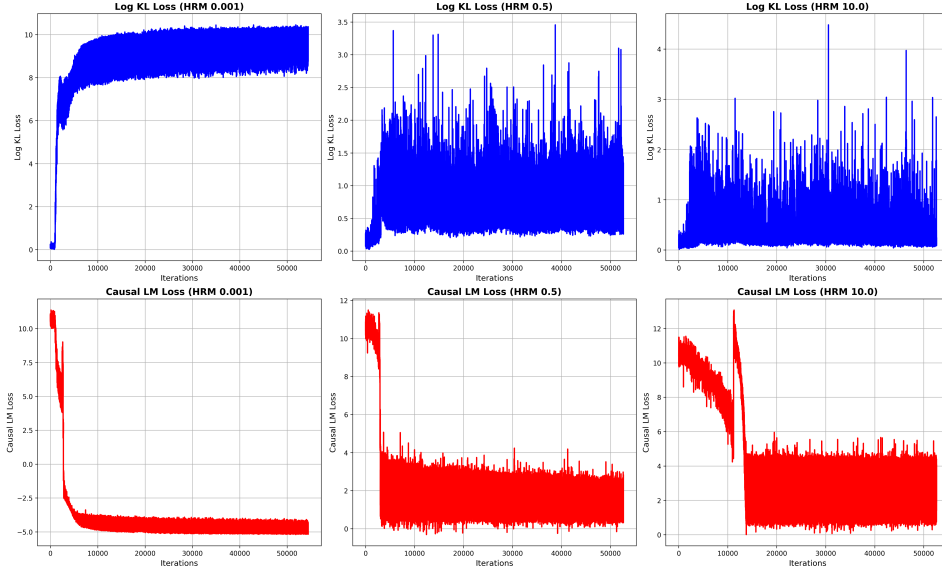


Figure 4: Loss dynamics during training across different KL loss weightings. Each column corresponds to a different λ value: 0.001 (left), 0.5 (middle), and 10.0 (right). The top row shows the log KL loss, while the bottom row depicts the causal language modeling loss.

DeepSeek-Math [16], and Qwen2.5-72B-Math-Instruct [4], applying the best-of- N strategy on the PRM800K dataset. Detailed training information is provided in Appendix A.2.

Table 4 presents the accuracy results of various policy models under PRM and HRM settings on the PRM800K dataset. Although both PRM and HRM training data are derived from MCTS with Qwen2.5-7B-Math, we evaluate the fine-tuned models using different policy models, where HRM consistently demonstrates greater stability and robustness compared to PRM.

The relatively lower performance of Qwen2.5-72B-Math-Instruct can be attributed to the tree height constraints imposed by MCTS, which necessitate answer generation within a predefined template and a fixed number of steps. Although Qwen2.5-72B-Math-Instruct demonstrates strong reasoning capabilities, its highly structured training process renders it susceptible to performance degradation when deviating from its learned format. To ensure computational feasibility, our MCTS framework enforces limitations on tree depth, which further amplifies this effect.

4.3 Self-Training

We adapt the method described in Section 3.3 to filter high-quality reasoning data and train the policy model. Fig. 4 illustrates that when λ is small (e.g., 0.001), the fine-tuned model rapidly loses its

| Policy Model | Method | N | | | | | | | | | |
|-----------------|--------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | 2 | 4 | 8 | 16 | 24 | 32 | 64 | 128 | 256 | 512 |
| Qwen2.5-7B-Math | PRM | 0.477 | 0.466 | 0.600 | 0.544 | 0.633 | 0.677 | 0.733 | 0.677 | 0.700 | 0.722 |
| | HRM | 0.500 | 0.566 | 0.655 | 0.600 | 0.666 | 0.711 | 0.711 | 0.766 | 0.777 | 0.766 |
| Qwen-7B-PRM | PRM | 0.477 | 0.555 | 0.533 | 0.655 | 0.655 | 0.677 | 0.711 | 0.700 | 0.744 | 0.744 |
| | HRM | 0.477 | 0.544 | 0.644 | 0.722 | 0.700 | 0.711 | 0.722 | 0.755 | 0.800 | 0.800 |
| Qwen-7B-HRM | PRM | 0.511 | 0.533 | 0.644 | 0.667 | 0.667 | 0.689 | 0.733 | 0.722 | 0.744 | 0.733 |
| | HRM | 0.489 | 0.589 | 0.722 | 0.722 | 0.722 | 0.744 | 0.744 | 0.789 | 0.789 | 0.789 |

Table 5: Comparison of fine-tuned policy model reasoning performance on the PRM800K dataset using the best-of- N strategy. Qwen-7B-HRM denotes the policy model fine-tuned on high-MC-score reasoning data from HRM’s training set, while Qwen-7B-PRM follows the same procedure for PRM’s training set.

| Dataset | Method | N | | | | | | | | | |
|---------|--------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | 2 | 4 | 8 | 16 | 24 | 32 | 64 | 128 | 256 | 512 |
| GSM8K | PRM | 0.784 | 0.828 | 0.858 | 0.882 | 0.884 | 0.893 | 0.905 | 0.917 | 0.927 | 0.918 |
| | HRM | 0.784 | 0.833 | 0.846 | 0.886 | 0.893 | 0.902 | 0.907 | 0.914 | 0.93 | 0.926 |
| Math500 | PRM | 0.468 | 0.572 | 0.598 | 0.624 | 0.658 | 0.658 | 0.656 | 0.662 | 0.686 | 0.688 |
| | HRM | 0.490 | 0.576 | 0.612 | 0.660 | 0.688 | 0.692 | 0.742 | 0.740 | 0.740 | 0.736 |

Table 6: Generalization performance of PRM and HRM fine-tuned on the PRM800K dataset and evaluated on GSM8K and Math500 using the best-of- N strategy. The policy model used for evaluation is Qwen2.5-7B-Math-Instruct.

generalization ability within just a few iterations, causing the KL loss to escalate to approximately 20,000. In contrast, the causal LM loss remains within the range of 0 to 12, leading to a significant imbalance. This discrepancy underscores the necessity of applying logarithmic scaling to the KL term in the objective function, as discussed in Section 3.3. Conversely, when λ is excessively large (e.g., 10.0), the model prioritizes adherence to the reference distribution, resulting in slower convergence and constrained improvements in reasoning capability.

We perform SFT using high-MC-score reasoning data from the PRM/HRM training datasets. Qwen-7B-HRM denotes the policy model fine-tuned on high-MC-score reasoning data from HRM’s training set, while Qwen-7B-PRM follows the same procedure for PRM’s training set. We set λ to 0.5. Table 5 further validates that SFT enhances the policy model’s reasoning capability by leveraging high-quality data, with HRM demonstrating greater robustness compared to PRM.

4.4 HRM Generalization Across Different Domains

To broaden the applicability of HRM and evaluate its generalization capability, we assess HRM and PRM, trained on the PRM800K dataset, on the Math500 [7] and GSM8K [6] datasets. Table 6 shows that HRM demonstrates superior generalization in Math500 compared to PRM, indicating its ability to handle complex mathematical reasoning tasks.

However, the performance difference between HRM and PRM in GSM8K is marginal, as the GSM8K dataset primarily consists of relatively simple arithmetic problems. A strong policy model can typically solve these problems within three steps on the GSM8K dataset, limiting the impact of HRM’s advantages, such as its ability to assess multi-step reasoning coherence and facilitate self-reflection through its training data. Nevertheless, as shown in Table 6, HRM still slightly outperforms PRM on simpler datasets like GSM8K. More importantly, in more complex datasets such as Math500, HRM demonstrates significantly greater robustness compared to PRM, highlighting its effectiveness in handling challenging reasoning tasks.

5 Conclusion

In this paper, we present **Hierarchical Reward Model**, which enhances multi-step reasoning evaluation by integrating fine-grained and coarse-grained assessments, improving reasoning coherence and self-reflection. We further introduce **Hierarchical Node Compression**, a data augmentation method that optimizes MCTS-based autonomous annotation, enhancing label diversity while expanding training data with minimal computational cost. Extensive experiments on PRM800K demonstrate HRM’s superior robustness over PRM, with strong generalization across GSM8K and MATH500. Additionally, MCTS-generated auto-labeled data enables policy model fine-tuning, further improving reasoning performance.

References

- [1] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. In *ArXiv e-prints*, 2023.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. In *ArXiv e-prints*, 2023.
- [3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. In *ArXiv e-prints*, 2024.
- [4] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. In *ArXiv e-prints*, 2024.
- [5] Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. Large language models are good multi-lingual learners: When llms meet cross-lingual prompts. In *COLING*, 2024.
- [6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. In *ArXiv e-prints*, 2021.
- [7] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *ICLR*, 2023.
- [8] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. In *ArXiv e-prints*, 2022.
- [9] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *ACL*, 2023.
- [10] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. Improve mathematical reasoning in language models by automated process supervision. In *ArXiv e-prints*, volume 2, 2024.
- [11] Adam Fisch Xinyang Geng Jacob Eisenstein Rishabh Agarwal Alekh Agarwal Jonathan Berant Aviral Kumar Amrith Setlur, Chirag Nagpal. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *ICLR*, 2025.
- [12] Teng Wang, Wing-Yin Yu, Ruifeng She, Wenhan Yang, Taijie Chen, and Jianping Zhang. Leveraging large language models for solving rare mip challenges. *arXiv preprint arXiv:2409.04464*, 2024.
- [13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

- [14] Jeffrey Zhao Izhak Shafran Thomas L. Griffiths Yuan Cao Karthik Narasimhan Shunyu Yao, Dian Yu. Deliberate problem solving with large language models. In *NeurIPS*, 2023.
- [15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. In *ArXiv e-prints*, 2025.
- [16] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. In *arXiv preprint arXiv:2402.03300*, 2024.
- [17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [18] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [19] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024.
- [20] Lilian Weng. Reward hacking in reinforcement learning, 2024. URL <https://lilianweng.github.io/posts/2024-11-28-reward-hacking/>.
- [21] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. In *arXiv preprint arXiv:1606.06565*, 2016.
- [22] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. ICML, 2022.
- [23] Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel. In *IJCAI*, 2017.
- [24] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *ACL*, 2024.
- [25] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf. In *ICLR*, 2025.
- [26] Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. Feedback loops with language models drive in-context reward hacking. In *ICML*, 2024.
- [27] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [28] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.
- [29] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.
- [30] Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. DeepSpeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15. IEEE, 2022.

- [31] Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019.
- [32] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

A Appendix

A.1 HRM Training Details

To accelerate the training process of reward model, we employ FlashAttention [27, 28], DeepSpeed [29, 30], and mixed-precision training [31, 32]. However, within the PRM800K domain, we frequently encounter the issue: *"Current loss scale already at minimum - cannot decrease scale anymore. Exiting run."* This indicates that the numerical precision is insufficient for stable training. To mitigate this issue and ensure reproducibility, we set `max_grad_norm` to 0.01, which effectively stabilizes the training process.

We define the completion of a reasoning step as the end of a formula, using `stop = ['\n\n', '\n\n', '# END!']` as boundary markers.

The following prompt is used in Section 4.1:

You are an expert of Math and need to solve the following question
and return the answer.

Question:
{question}

Let's analyze this step by step.

After you finish thinking, you need to output the answer again!
The answer should start with '# Answer', followed by two line breaks
and the final response.
Just provide the answer value without any descriptive text at the end.
And the answer ends with '# END!'
Below is a correct example of the expected output format:

Question: $1+2+3 = ?$

Firstly, solve $1 + 2 = 3$,
Then, $3 + 3 = 6$.

Answer

6
END!

"""

A.2 HNC Setting Details

To ensure the feasibility of autonomous annotation using MCTS, we impose constraints on both the width and height of the search tree. This limitation prevents us from treating the completion of a formula as a single reasoning step. Instead, we require the model to explicitly output # Step X at each step. Consequently, the training data for the reward model is segmented using # Step X as a delimiter. During inference, we also apply # Step X as a separator and employ the Best-of-N strategy for selecting the optimal reasoning path.

The prompt we use is as follows(`delimiter=['# END!', '# Step 2', '# Step 3', '# Step 4', '# Step 5']`):

""You are an expert of Math and need to solve the following question
and return the answer.

Question:
{question}

Let's analyze this step by step.

Begin each step with '# Step X' to clearly indicate the entire reasoning step. After you finish thinking, you need to output the answer again!

The answer should start with '# Answer', followed by two line breaks and the final response.

Just provide the answer value without any descriptive text at the end.

And the answer ends with '# END!'

Below is a correct example of the expected output format:

Question: $1+2+3 = ?$

Step 1

solve $1 + 2 = 3$,

Step 2

Then, $3 + 3 = 6$.

Answer

6

END!

"""

A.3 Self-Training

Initially, KL loss is not incorporated, causing the policy model to lose its generalization ability rapidly, despite a continuous decrease in evaluation loss. To address this issue, we introduce KL loss to regularize training from the reference model.

The logarithmic scaling and weighting factor λ are added to balance the impact of KL divergence. Without these adjustments, KL loss would range from 0 to 8000, while the language modeling loss remains between 0 and 12, leading to an imbalance. The logarithm ensures a more stable contribution of KL loss during training.

As illustrated in Fig. 4, setting $\lambda = 0.5$ achieves a balanced trade-off between KL loss and language modeling loss, preventing excessive divergence from the reference model while ensuring stable and effective training.