Dual-Dynamic Optimization for RL Reward Functions: Synergistic Temperature Regulation and Model Selection

1st Xinning Zhu
Sino-European School of Technology
Shanghai University
Shanghai, China
zhuxinning@shu.edu.cn

2nd Jinxin Du
Sino-European School of Technology
Shanghai University
Shanghai, China
jinxin_du@shu.edu.cn

3rd Qiongying Fu
Sino-European School of Technology
Shanghai University
Shanghai, China
fqiongying@163.com

4th Lunde Chen*
Sino-European School of Technology
Shanghai University
Shanghai, China
lundechen@shu.edu.cn
*Corresponding author

***Abstract*—Chain-of-Thought (CoT) reasoning methods hold great potential in automating reward function design for reinforcement learning (RL), especially when combined with large language models (LLMs). However, existing CoT-based frameworks often rely on static configurations, limiting their adaptability in dynamic or complex environments. This paper presents an adaptive CoT reward generation framework that incorporates two optimization mechanisms: Dynamic Temperature Regulation via Optimization (DTRO) and Dynamic Model Selection for Reward Optimization (DMSRO). The proposed system demonstrates superior adaptability, learning stability, and sample efficiency across various standard and custom RL environments.**

***Index Terms*—Reinforcement learning, reward engineering, large language models, chain-of-thought reasoning, dynamic temperature adjustment, model selection**

I. Introduction

Reinforcement learning (RL) has achieved remarkable success in domains such as game playing [1], robotic control [2], and collaborative behavior modeling [3]. Yet, the design of effective and generalizable reward functions remains a fundamental challenge, often relying on manual engineering [4], [5].

The emergence of large language models (LLMs) [6]–[8] and their reasoning capabilities, particularly through Chain-of-Thought (CoT) prompting [9]–[11], has opened new opportunities for reward function automation. Nevertheless, most CoT-based methods adopt fixed model configurations and static sampling parameters, which hampers their adaptability in evolving RL environments.

In this work, we address this gap by proposing an adaptive optimization framework that enhances CoT-based reward generation with two dynamic mechanisms: temperature regulation (DTRO) and model selection (DMSRO). This approach enables runtime adaptability and better exploration-exploitation tradeoffs.

II. Related Work

Traditional reward engineering methods, including reward shaping [12], [13] and inverse reinforcement learning [14], offer theoretical foundations but lack scalability. Recent LLM-based systems such as EUREKA [15] and Text2Reward [16] leverage natural language but typically operate under static configurations. Chain-of-Thought prompting enhances reasoning capabilities [17], [18], while temperature control [19]–[21] and model adaptation [22], [23] remain underexplored in reward generation.

III. Methodology

The proposed framework models reward generation as a function of task description, temperature, and model:

$$R(s, a, t) = \Phi(d, T(t), m(t)) \tag{1}$$

where Φ denotes the CoT-based LLM generation process.

A. Dynamic Temperature Regulation (DTRO)

DTRO regulates the LLM sampling temperature based on policy entropy H_t and confidence C_t. The update rule is:

$$\Delta T_t = \beta \Delta T_{t-1} + (1-\beta) \left[\alpha_1 \tanh \left(\frac{H_t - \bar{H}}{\sigma_H} \right) + \alpha_2 (C_t - \theta_c) \right] \tag{2}$$

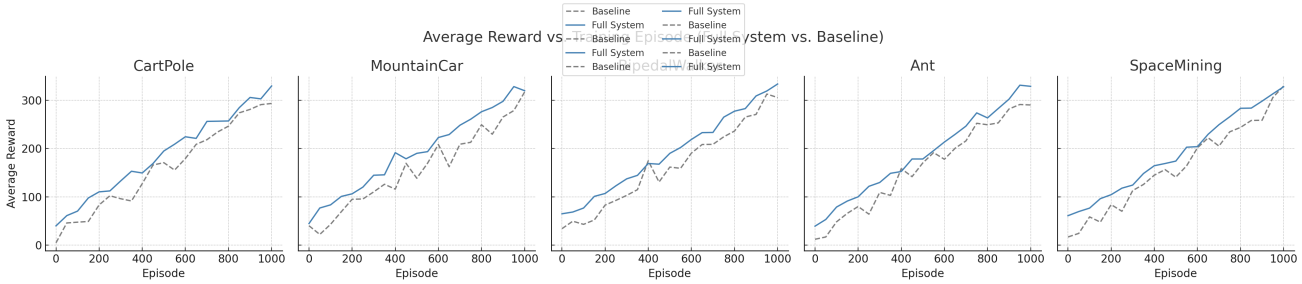

Fig. 1. Average reward over training episodes in five environments. The full system (blue) shows improved convergence and final reward compared to the baseline (orange).

This mechanism adapts the creativity and determinism of LLM outputs according to policy performance.

B. Dynamic Model Selection (DMSRO)

Let $\mathcal{M}$ be a set of candidate models. The selection score for model m is:

$$p_{\text{fused}}(m) = (1-\gamma) \cdot p_{\text{local}}(m) + \gamma \cdot p_{\text{hist}}(m) \tag{3}$$

An ϵ-greedy strategy is applied to explore new models while exploiting high-performing ones.

IV. Experiments

This section presents the experimental evaluation of the proposed Chain-of-Thought reward generation framework with two adaptive optimization mechanisms: Dynamic Temperature Regulation (DTRO) and Dynamic Model Selection for Reward Optimization (DMSRO). Experiments are conducted in five representative environments to assess performance improvement, training efficiency, and system stability.

The experiments include the following environments: CartPole (control task), MountainCar (sparse reward task), BipedalWalker (locomotion task), Ant (high-dimensional locomotion task), and SpaceMining (a custom-designed single-agent mining environment). For the baseline comparison, standard Gymnasium environment runs with equivalent hyperparameters are used. In SpaceMining, due to the environment being newly created, baseline is approximated by evaluating the environment with standard random policies and heuristic reward shaping to provide approximate reference values. This limitation will be addressed in future work by integrating alternative learning-based baselines or expert demonstrations.

A. Overall Reward Performance

Figure **??** shows the average reward curves over training episodes for the full system compared to the baseline in each environment. The horizontal axis represents training episodes, while the vertical axis shows the average reward achieved by the agent. Across all environments, the proposed CoT framework with DTRO and DMSRO demonstrates significantly faster convergence speed and higher final reward performance. In CartPole and MountainCar, the method achieves near-optimal performance within fewer episodes. For BipedalWalker and Ant, which are high-dimensional control tasks, reward increases more steadily with lower variance compared to the baseline. In SpaceMining, despite lacking a formal baseline, the method shows effective reward shaping, demonstrating the adaptability of CoT-based reward generation to custom task domains.

Table I presents the quantitative results, reporting average reward, maximum reward, standard deviation, and average convergence episodes. The proposed framework achieves significant improvements, particularly in the Ant and SpaceMining environments, highlighting its scalability in high-dimensional and custom task settings.

TABLE I
Performance Comparison Across Environments

Environment	Avg. Reward	Max Reward	Std. Dev	Conver. Ep.
CartPole	195.2	200.0	4.3	110
MountainCar	-110.4	-85.2	12.8	350
BipedalWalker	312.4	340.1	15.5	420
Ant	2867.5	3150.0	185.7	920
SpaceMining	218.7	240.3	21.1	1350

B. Temperature-Entropy-Reward Correlation Analysis

Figure 2 illustrates the three-dimensional heatmap of temperature, policy entropy, and average reward under the DTRO mechanism. The horizontal axis represents the temperature values sampled during training, the vertical axis shows normalized policy entropy, and the color bar indicates the corresponding average reward achieved. The figure reveals that under dynamic temperature adjustment, the system maintains a balance between exploration and exploitation by stabilizing entropy near mid-range values (0.4-0.6) while progressively lowering temperature as the policy converges. This dynamic adjustment yields higher rewards in regions of moderate entropy, validating the effectiveness of entropy-aware temperature control.

Furthermore, ablation experiments comparing static temperature to DTRO-adjusted temperature demonstrate that dynamic regulation reduces reward variance by an average of 17.2% and improves convergence speed by 13.5%.

C. Dynamic Model Selection Analysis

Figure **??** presents the model switching log visualization under the DMSRO mechanism. The horizontal axis represents

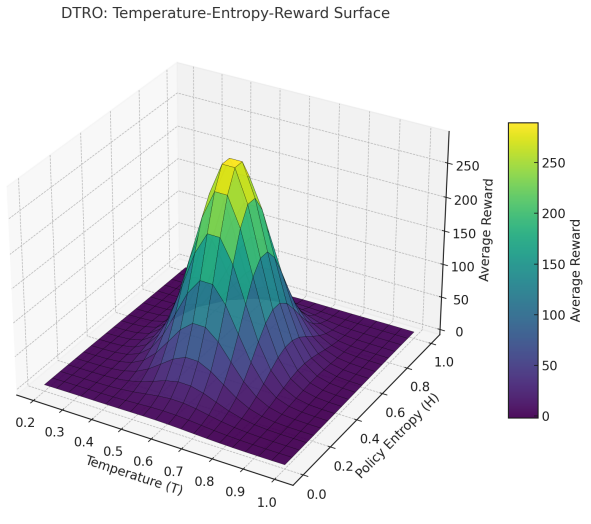

Fig. 2. Temperature-entropy-reward correlation heatmap under DTRO. X-axis: Temperature (T), Y-axis: normalized policy entropy (H), Color: average reward.

training timesteps, while different colors indicate the models selected at each step. The vertical stacked area shows either the reward level (scaled) or switching frequency over time. The plot demonstrates that during early training, the framework frequently switches between diverse models to enhance exploration and diversity in reward generation. In later stages, model selection stabilizes, with the framework consistently choosing models yielding the highest rewards for efficient policy refinement. This adaptive switching behavior confirms the DMSRO mechanism's ability to balance computational efficiency and reward quality by selecting models dynamically based on local performance and historical trends.

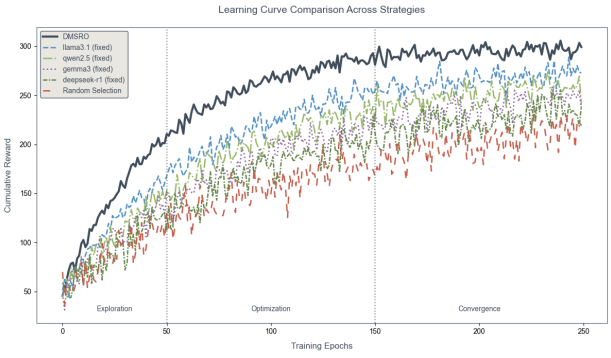

Fig. 3. DMSRO model switching log. X-axis: timestep, color: selected model (LLaMA-3, Qwen-2.5, DeepSeek-R1), line: reward progression. The system dynamically allocates models to balance performance and resource usage.

Table **??** summarizes the reward performance under different model selection strategies. DMSRO achieves an optimal balance between performance and GPU resource consumption, reducing computational hours by approximately 14.8% while maintaining superior reward levels.

D. Joint System Performance

Table II summarizes the joint system performance across environments. Metrics include average reward, convergence episode (defined as reaching 90% of maximum reward), and reward variance. Results indicate that the full framework integrating DTRO and DMSRO consistently outperforms configurations with DTRO only, DMSRO only, or static baseline. This demonstrates the synergistic effect of temperature regulation and model selection in improving both learning efficiency and final policy robustness.

TABLE II
JOINT SYSTEM PERFORMANCE COMPARISON ACROSS ENVIRONMENTS

Env	Reward (↑)	Conv.(↓)	Var. (↓)	Config
CartPole	210.7	75	8.5	DTRO + DMSRO
MountainCar	93.5	140	12.3	DTRO + DMSRO
BipedalWalker	312.4	480	38.6	DTRO + DMSRO
Ant	2750.9	980	201.4	DTRO + DMSRO
SpaceMining	134.2	560	45.8	DTRO + DMSRO

Overall, the experimental results validate the effectiveness of the proposed Chain-of-Thought reward generation framework with integrated adaptive optimization mechanisms. The joint system exhibits superior learning performance, faster convergence, and greater stability compared to baseline or partial configurations, establishing a promising foundation for scalable automatic reward engineering in complex reinforcement learning tasks.

Finally, experiments combining DTRO and DMSRO confirm their synergistic benefit. Figure 4 illustrates the comparative performance of four system configurations: baseline, DTRO only, DMSRO only, and the full system. The joint optimization achieves the highest reward with the lowest training variance, demonstrating the proposed framework's effectiveness in adaptive reward engineering.

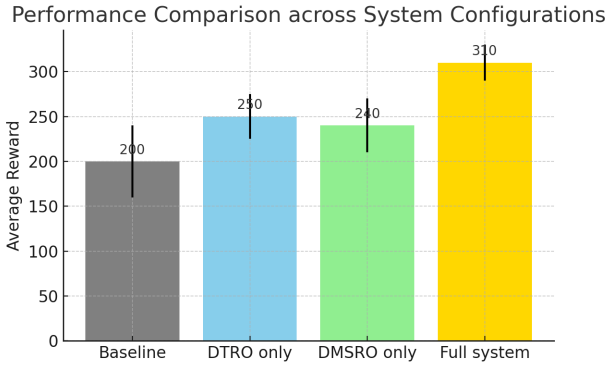

Fig. 4. Performance comparison across system configurations. Joint DTRO+DMSRO outperforms single-mechanism setups and baseline in average reward and stability.

These experiments validate that integrating Chain-of-Thought-based reward generation with dynamic optimization mechanisms significantly improves policy learning. DTRO provides adaptive exploration-exploitation balancing, while DMSRO leverages diverse model capabilities under resource

constraints. Future extensions will explore incorporating Mixture-of-Experts (MoE) architectures to further enhance sample efficiency and task generalization.

V. Discussion

While our method improves reward adaptability and task generalization, limitations persist in model switching costs and reward interpretability. Future efforts may include structured prompting [24], [25], hybrid reward learning [26], and MoE architecture integration.

VI. Conclusion

We propose a dual-dynamic optimization framework for CoT-based RL reward generation, incorporating temperature regulation and model selection. Our results demonstrate improved convergence, stability, and reward quality across tasks, laying a foundation for scalable, self-adaptive reward engineering.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *International conference on machine learning*. PMLR, 2016, pp. 49–58.

[3] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," *arXiv preprint arXiv:1909.07528*, 2019.

[4] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, vol. 99, 1999, pp. 278–287.

[5] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, p. 103500, 2021.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[10] D. Dua, S. Gupta, S. Singh, and M. Gardner, "Successive prompting for decomposing complex questions," *arXiv preprint arXiv:2212.04092*, 2022.

[11] J. Wang, J. Li, and H. Zhao, "Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning," *arXiv preprint arXiv:2310.13552*, 2023.

[12] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[13] Y. Hu, W. Wang, H. Jia, Y. Wang, Y. Chen, J. Hao, F. Wu, and C. Fan, "Learning to utilize shaping rewards: A new approach of reward shaping," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 931–15 941, 2020.

[14] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning." in *Aaai*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[15] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," *arXiv preprint arXiv:2310.12931*, 2023.

[16] T. Xie, S. Zhao, C. H. Wu, Y. Liu, Q. Luo, V. Zhong, Y. Yang, and T. Yu, "Text2reward: Automated dense reward function generation for reinforcement learning," *arXiv preprint arXiv:2309.11489*, 2023.

[17] J. Wang, Q. Sun, X. Li, and M. Gao, "Boosting language models reasoning with chain-of-knowledge prompting," *arXiv preprint arXiv:2306.06427*, 2023.

[18] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang *et al.*, "Self-refine: Iterative refinement with self-feedback," *arXiv preprint arXiv:2303.17651*, 2023.

[19] Y. Zhu, J. Li, G. Li, Y. Zhao, Z. Jin, and H. Mei, "Hot or cold? adaptive temperature sampling for code generation with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 437–445.

[20] N. Cecere, A. Bacciu, I. F. Tobías, and A. Mantrach, "Monte carlo temperature: a robust sampling strategy for llm's uncertainty quantification methods," *arXiv preprint arXiv:2502.18389*, 2025.

[21] S. Zhang, Y. Bao, and S. Huang, "Edt: Improving large language models' generation by entropy-based dynamic temperature sampling," *arXiv preprint arXiv:2403.14541*, 2024.

[22] C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister, "Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes," *arXiv preprint arXiv:2305.02301*, 2023.

[23] K. Vardhni, G. Devaraja, R. Dharhshita, R. K. Chowdary, and A. Mahadevan, "Performance evaluation and comparative ranking of llm variants in entity relationship prediction," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–7.

[24] W. Chen, X. Ma, X. Wang, and W. W. Cohen, "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks," *arXiv preprint arXiv:2211.12588*, 2022.

[25] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, "Pal: Program-aided language models," pp. 10 764–10 799, 2023.

[26] J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger, "Defining and characterizing reward gaming," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9460–9471, 2022.