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:

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

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

Experiments are conducted on four OpenAI Gym-style environments and a custom single-agent mining task. Each trial logs reward, convergence iterations, and stability.

TABLE I
COMPARISON OF LLMS ON REWARD GENERATION PERFORMANCE

Model	Avg. Reward	Max	Min	Std. Dev.
LLaMA3.1	208.6	312.4	-33.7	139.5
Qwen2.5	202.5	316.3	-41.6	138.9
Gemma3	121.6	305.4	-34.2	140.0
Deepseek-R1	61.2	318.2	-96.2	150.3

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.