Towards Universal Thinking: A Domain-Adaptive Reinforcement Learning Framework for Language Models

Rani and Gemini

September 25, 2025

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities, particularly with the advent of Chain-of-Thought (CoT) reasoning and Reinforcement Learning (RL) techniques like RL from Human Feedback (RLHF) and RL with Verifiable Rewards (RLVR). Recent work, such as RL with Model-rewarded Thinking (RLMT) (Bhaskar et al., 2025), has shown significant improvements in general-purpose chat and creative writing by explicitly training models to "think" before responding. However, a key limitation remains the inconsistent generalization of these thinking pipelines across highly diverse domains, with performance sometimes lagging in structured tasks like instruction following or logical puzzles. This paper proposes a novel **Domain-Adaptive Reinforcement Learning Framework** for LLMs designed to overcome this challenge. Our framework introduces mechanisms for dynamic CoT generation tailored to inferred task domains, coupled with a hybrid and adaptive reward system that integrates preference-based, verifier-based, and critically, CoT-quality-based reward signals. We mathematically formalize this approach and outline conceptual experimental considerations, aiming to foster more robust, versatile, and universally "thinking" LLMs.

1 Introduction

The rapid advancements in Large Language Models (LLMs) have transformed the landscape of Artificial Intelligence, enabling machines to perform complex language understanding and generation tasks with unprecedented fluency. A pivotal development in enhancing LLM reasoning capabilities has been the integration of Chain-of-Thought (CoT) prompting [9], which encourages models to articulate intermediate reasoning steps before arriving at a final answer. This explicit reasoning process has been further refined and optimized through various reinforcement learning (RL) paradigms.

Reinforcement Learning from Human Feedback (RLHF) [6] has become a cornerstone for aligning LLMs with human preferences, allowing models to gen-

erate more helpful, harmless, and honest responses. Complementary to this, Reinforcement Learning with Verifiable Rewards (RLVR) [5] has proven highly effective in domains where ground-truth verification is possible, such as mathematics and code, by optimizing models against rule-based rewards.

Building upon these foundations, Bhaskar et al. (2025) introduced RL with Model-rewarded Thinking (RLMT) [1], demonstrating that training LLMs to generate a long CoT reasoning sequence before a response, and then optimizing this process with an online RL algorithm against a preference-based reward model, significantly boosts performance in general-purpose chat and creative writing. Remarkably, their 8B model even rivaled or surpassed much larger state-of-the-art models in these open-ended tasks.

Despite these successes, Bhaskar et al. (2025) noted that the benefits of RLMT were less pronounced in certain structured domains, indicating a limitation in the universal generalization of their "thinking" pipeline. This observation highlights a critical research gap: how can LLMs be trained to adapt their internal reasoning processes and leverage appropriate evaluation signals across an entire spectrum of tasks, ranging from highly subjective creative endeavors to objectively verifiable problem-solving?

This paper proposes a novel **Domain-Adaptive Reinforcement Learning Framework** for LLMs to address this challenge. Our framework aims to equip LLMs with the ability to dynamically infer the nature of a given task and consequently adjust their CoT generation strategy and the composition of their reward signals. The ultimate goal is to enable truly universal thinking capabilities, ensuring robust performance across all domains.

The remainder of this paper is structured as follows: Section 2 provides background on CoT, RLHF, RLVR, and RLMT. Section 3 details our proposed Domain-Adaptive Thinking Framework, including its architectural components and mathematical formalism. Section 4 outlines conceptual experimental considerations. Finally, Section 5 concludes the paper and suggests future research directions.

2 Background and Related Work

2.1 Chain-of-Thought (CoT) Reasoning

Chain-of-Thought (CoT) prompting [9] is a technique that enables LLMs to decompose complex problems into intermediate steps and articulate these steps explicitly. This process has been shown to significantly improve reasoning abilities in tasks requiring multi-step problem-solving, such as arithmetic, common sense reasoning, and symbolic manipulation. By generating a coherent sequence of thoughts, LLMs can often achieve more accurate and interpretable results, mimicking human-like deliberation.

2.2 Reinforcement Learning from Human Feedback (RLHF)

RLHF [2, 6] is a powerful paradigm for aligning LLMs with human preferences and values. It typically involves three steps: (1) pre-training a language model on a large text corpus, (2) fine-tuning the model using supervised data (SFT), and (3) further optimizing the SFT model using RL. In the RL step, a reward model is trained on human preference data (e.g., pairwise comparisons of model outputs). This reward model then provides a scalar reward signal to an RL algorithm (e.g., PPO [7]) to fine-tune the LLM policy, maximizing the expected human-aligned reward.

2.3 Reinforcement Learning with Verifiable Rewards (RLVR)

RLVR extends the RL framework to domains where the correctness of an LLM's output can be objectively verified. This is particularly relevant for tasks like mathematical problem-solving [4] or code generation [3]. Instead of relying on a subjective reward model, RLVR often uses rule-based verifiers or unit tests to provide a binary (correct/incorrect) or scalar reward. DeepSeek-R1 [3], for instance, combined RLVR with long CoT reasoning, where models generate traces that are then stripped before verification. While highly effective in its target domains, RLVR's reliance on strict verifiability limits its applicability to open-ended tasks.

2.4 RL with Model-rewarded Thinking (RLMT)

Bhaskar et al. (2025) introduced RLMT [1] as a method to extend the benefits of CoT reasoning, typically seen in verifiable domains, to general-purpose chat. RLMT mandates that LLMs generate a detailed Chain of Thought (z) before producing a final response (y). This entire process is then optimized using online RL algorithms (like GRPO [8]) against a preference-based reward model, similar to RLHF. The key insight is that by explicitly training for "thinking" using human-aligned rewards, models can develop more sophisticated reasoning strategies for open-ended tasks. RLMT demonstrated significant performance gains on chat and creative writing benchmarks, even outperforming larger models and frontier thinking models like Claude-3.7-Sonnet on some metrics. However, the authors acknowledged that RLMT's generalization to domains requiring strict instruction following or precise logical puzzles was not as strong, pointing to the need for a more universally adaptable framework.

3 The Domain-Adaptive Thinking Framework

To address the limitations of existing RL approaches in generalizing "thinking" across all domains, we propose a **Domain-Adaptive Reinforcement Learning Framework**. This framework is designed to enable LLMs to dynamically adjust their internal reasoning strategies and leverage domain-specific reward signals, thus achieving more robust and versatile performance.

3.1 Problem Formulation

Let $x \in X$ be an input prompt. Our goal is to train a language model π_θ with parameters θ to generate a Chain of Thought z and a final response y. Unlike previous methods, we introduce a mechanism to infer the domain $\mathcal{D}(x)$ of the input prompt. The model's policy for generating z and y will then be conditioned on this inferred domain, denoted as $\pi_\theta(z, y|x, \mathcal{D}(x))$. The objective is to maximize an expected domain-adaptive reward:

$$\max_\theta \mathbb{E}_{x\sim X,(z,y)\sim\pi_\theta(\cdot|x,\mathcal{D}(x))} \left[R(x, z, y, \mathcal{D}(x))\right]$$

where $R(x, z, y, \mathcal{D}(x))$ is a hybrid reward function that adapts its components based on the inferred domain.

3.2 Architectural Overview

Our proposed framework consists of the following conceptual components:

1. **Domain Inferrer:** A module that analyzes the input prompt x and predicts its most relevant domain $\mathcal{D}(x)$. This could be a separate classifier or an intrinsic capability of the LLM itself, trained to identify task characteristics (e.g., "mathematical problem," "creative writing request," "strict instruction following").

2. **Adaptive CoT Policy ($\pi_\theta(z|x, \mathcal{D}(x))$):** This policy is responsible for generating the internal Chain of Thought z. Crucially, it adapts its strategy based on the inferred domain. For instance, for a mathematical problem, it might prioritize formal logical steps, while for a creative writing task, it might engage in brainstorming or narrative planning.

3. **Response Generation Policy ($\pi_\theta(y|x, z)$):** This policy generates the final response y, conditioned on both the input prompt x and the generated CoT z.

4. **Hybrid Reward Model ($R(x, z, y, \mathcal{D}(x))$):** This module computes the reward signal, dynamically combining different reward components based on the inferred domain $\mathcal{D}(x)$.

3.3 Dynamic Chain-of-Thought Generation

The key to universal thinking lies in the ability to generate a CoT that is optimal for the specific task at hand. Instead of a fixed CoT format, our framework envisions a dynamic approach:

- **Task-conditioned CoT Structures:** For verifiable tasks (e.g., math, logic), the CoT policy might generate a highly structured, step-by-step deduction. For creative tasks, it might produce a more fluid brainstorming process, exploring multiple ideas before converging. For instruction following, the CoT could involve explicit constraint identification and checking.

- **Meta-Strategies for CoT:** The LLM could learn meta-strategies for generating CoT, such as "plan-and-execute," "critique-and-refine," or "explore-and-select," and apply these based on the domain. This could be facilitated by explicit prompting for CoT generation that includes domain-specific guidance.

3.4 Hybrid and Adaptive Reward Mechanisms

A single reward model, whether preference-based or rule-based, is insufficient for universal generalization. Our framework proposes a hybrid reward system:

$$R(x, z, y, \mathcal{D}(x)) = w_1(\mathcal{D}(x)) \cdot r_{preference}(x, y) + w_2(\mathcal{D}(x)) \cdot r_{verifier}(x, y) + w_3(\mathcal{D}(x)) \cdot r_{CoT\_quality}(x, z)$$

where $w_1(\mathcal{D}(x))$, $w_2(\mathcal{D}(x))$, and $w_3(\mathcal{D}(x))$ are domain-dependent weights such that $w_1 + w_2 + w_3 = 1$. These weights are learned or explicitly set based on the inferred domain $\mathcal{D}(x)$.

- **Preference-based Rewards ($r_{preference}(x, y)$):** Similar to RLHF and RLMT, these rewards capture subjective quality, coherence, and helpfulness, particularly important for open-ended tasks like chat and creative writing.
- **Verifier-based Rewards ($r_{verifier}(x, y)$):** For domains where objective correctness is paramount (e.g., math, code, factual questions, strict instruction following), these rewards are derived from external verifiers or rule-based checks.
- **CoT Quality Rewards ($r_{CoT\_quality}(x, z)$):** This is a crucial novel component. It aims to directly evaluate the quality of the internal reasoning process z, independent of the final answer y. This could involve:
  - A separate "thought-evaluator" LLM trained on annotated CoT examples for coherence, logical consistency, completeness, and relevance.
  - Rule-based metrics for specific CoT structures (e.g., checking if all constraints were addressed in an instruction-following CoT).
  - Human feedback specifically on the quality of the thought process.

The weights $w_i(\mathcal{D}(x))$ ensure that the reward signal is appropriately composed for each domain. For example, in a math problem, w_2 would be high, and w_1 low. In a creative writing task, w_1 would dominate. w_3 (CoT quality) would be significant across all domains, encouraging robust internal reasoning regardless of the task type.

3.5 Meta-Learning for Generalization

To facilitate rapid adaptation to new or unseen domains, meta-learning techniques could be integrated. This would involve training the LLM to learn how to learn domain-specific CoT strategies and reward interpretations. The domain inferrer itself could be a meta-learner, dynamically adjusting its confidence in domain classification and influencing the adaptive weights.

4 Conceptual Experimental Considerations

To validate the proposed framework, a comprehensive experimental setup would be required:

- **Diverse Datasets:** A benchmark suite spanning a wide range of domains, including open-ended tasks (chat, creative writing), verifiable tasks (math, code, logical puzzles), and instruction-following tasks with varying complexity and constraint types.
- **Baselines:** Comparison against strong baselines, including standard RLHF models, RLVR models (where applicable), and the original RLMT approach.
- **Evaluation Metrics:** Domain-specific metrics (e.g., human preference scores for creativity, exact match accuracy for math, instruction compliance scores for IF tasks) and aggregate performance metrics to assess overall versatility.
- **Ablation Studies:** To quantify the contribution of each novel component: the dynamic CoT generation, the hybrid reward model, and specifically the CoT quality reward.
- **Qualitative Analysis:** In-depth analysis of generated CoTs across different domains to demonstrate the adaptive nature of the thinking process.

5 Conclusion and Future Work

This paper has introduced a **Domain-Adaptive Reinforcement Learning Framework** for Language Models, addressing the critical challenge of generalizing "thinking" capabilities across a diverse range of tasks. By proposing dynamic Chain-of-Thought generation strategies, a hybrid reward system incorporating preference-based, verifier-based, and CoT-quality rewards, and the integration of meta-learning principles, we aim to pave the way for LLMs that can truly adapt their internal reasoning to any given domain.

Our work builds upon the significant advancements of RLMT [1] and extends the vision of universally capable LLMs. The proposed framework offers a path towards more robust, versatile, and inherently interpretable AI systems.

Future work will focus on the practical implementation and empirical validation of this framework. Specific directions include:

- Developing robust and scalable methods for training the CoT quality reward model, potentially leveraging self-supervised techniques or advanced human-in-the-loop annotation.
- Exploring different architectural designs for the domain inferrer and adaptive CoT policy, including end-to-end differentiable approaches.
- Investigating the theoretical properties of domain generalization in RL for LLMs, particularly concerning the interplay between different reward signals.
- Extending the framework to multimodal generative tasks, where adaptive thinking could involve integrating information from various modalities.

Through these efforts, we envision a future where LLMs can not only "think" but also adapt their thinking processes intelligently to meet the unique demands of any task, thereby achieving truly universal intelligence.

References

[1] Bhaskar, A., Ye, X., & Chen, D. (2025). *Language Models that Think, Chat Better*. arXiv preprint arXiv:2509.20357. [PDF]

[2] Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences*. Advances in Neural Information Processing Systems, 30.

[3] DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. Available at: `https://arxiv.org/abs/2504.13837`.

[4] Kazemnejad, A., Aghajohari, M., Portelance, E., Sordoni, A., Reddy, S., Courville, A., & Le Roux, N. (2025). *VinePPO: Refining credit assignment in RL training of LLMs*. In Forty-second International Conference on Machine Learning.

[5] Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Miranda, L. J. V., ... & Hajishirzi, H. (2025). *Tulu 3: Pushing frontiers in open language model post-training*. In Second Conference on Language Modeling.

[6] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. Advances in Neural Information Processing Systems, 35.

[7] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. arXiv preprint arXiv:1707.06347.

[8] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., ... & Guo, D. (2024). *DeepSeekMath: Pushing the limits of mathematical reasoning in open language models*. arXiv preprint arXiv:2407.04271.

[9] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). *Chain of thought prompting elicits reasoning in large language models*. Advances in Neural Information Processing Systems, 35.