Novelty Assessment Report

Paper: Overthinking Reduction with Decoupled Rewards and Curriculum Data Scheduling
PDF URL: https://openreview.net/pdf?id=kdeiRledV6
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

While large reasoning models trained with critic-free reinforcement learning and verifiable rewards (RLVR) represent the state-of-the-art, their practical utility is hampered by ``overthinking'', a critical issue where models generate excessively long reasoning paths without any performance benefit. Existing solutions that penalize length often fail, inducing performance degradation due to a fundamental misalignment between trajectory-level rewards and token-level optimization. In this work, we introduce a novel framework, DECS, built on our theoretical discovery of two previously unaddressed flaws in current length rewards: (1) the erroneous penalization of essential exploratory tokens and (2) the inadvertent rewarding of partial redundancy. Our framework's innovations include (i) a first-of-its-kind decoupled token-level reward mechanism that surgically distinguishes and penalizes redundant tokens, and (ii) a novel curriculum batch scheduling strategy to master the efficiency-efficacy equilibrium. Experimental results show DECS can achieve a dramatic reduction in reasoning tokens by over 50\% across seven benchmarks while simultaneously maintaining or even improving performance. It demonstrates conclusively that substantial gains in reasoning efficiency can be achieved without compromising a model's underlying reasoning power.

Core Task Landscape

This paper addresses: **reducing overthinking in large reasoning models**

A total of **50 papers** were analyzed and organized into a taxonomy with **25 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Overthinking Detection and Analysis**
- **Adaptive Reasoning Control**
- **Training-Based Overthinking Mitigation**
- **Inference-Time Optimization**
- **Efficiency Enhancement via Model Architecture**
- **Context and Input Optimization**
- **System-Level Inference Optimization**
- **Domain-Specific and Application-Oriented Efficiency**
- **Prompting and In-Context Learning for Efficiency**
- **Comprehensive Surveys and Frameworks**
- ... and 3 more categories

Complete Taxonomy Tree

- reducing overthinking in large reasoning models Survey Taxonomy
- Overthinking Detection and Analysis
  - Overthinking Characterization and Measurement (5 papers)
  - [5] Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models (Sui Yang, 2025) View paper
  - [10] Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms (Su, 2025) View paper
  - [16] Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs (Chen, 2024) View paper
  - [38] The Price of a Second Thought: On the Evaluation of Reasoning Efficiency in Large Language Models (Fan Siqi, 2025) View paper
  - [44] Optimalthinkingbench: Evaluating over and underthinking in llms (Aggarwal, 2025) View paper
  - Mechanistic Analysis of Overthinking (3 papers)
  - [21] Mitigating Overthinking in Large Reasoning Models via Manifold Steering (Huang Yao, 2025) View paper
  - [29] Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification (Zhang Anqi, 2025) View paper
  - [37] The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks (Li DaCheng, 2025) View paper
- Adaptive Reasoning Control
  - Difficulty-Adaptive Reasoning Allocation (4 papers)
  - [3] DAST: Difficulty-Adaptive Slow-Thinking for Large Reasoning Models (Shen Yi, 2025) View paper
  - [14] ARM: Adaptive Reasoning Model (Wu, 2025) View paper
  - [26] MUR: Momentum Uncertainty guided Reasoning for Large Language Models (Yan Hang, 2025) View paper
  - [31] AutoL2S: Auto Long-Short Reasoning for Efficient Large Language Models (Luo Feng, 2025) View paper
  - Early Exit Mechanisms (3 papers)
  - [6] Dynamic Early Exit in Reasoning Models (Yang Chenxu, 2025) View paper
  - [15] S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models (Yang Chenxu, 2025) View paper

- [34] Stop spinning wheels: Mitigating llm overthinking via mining patterns for early reasoning exit (Wei, 2025) View paper
- Training-Based Overthinking Mitigation
  - Reward Engineering for Efficiency ★ (4 papers)
  - [0] Overthinking Reduction with Decoupled Rewards and Curriculum Data Scheduling (Anon et al., 2026) View paper
  - [19] REA-RL: Reflection-Aware Online Reinforcement Learning for Efficient Large Reasoning Models (Deng, 2025) View paper
  - [32] Mitigating Overthinking through Reasoning Shaping (Song, 2025) View paper
  - [45] SmartThinker: Learning to Compress and Preserve Reasoning by Step-Level Length Control (He Xing-yang, 2025) View paper
  - Data-Centric Training Strategies (3 papers)
  - [2] Self-training elicits concise reasoning in large language models (Ho, 2025) View paper
  - [27] O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning (Luo Haotian, 2025) View paper
  - [28] Long-Short Chain-of-Thought Mixture Supervised Fine-Tuning Eliciting Efficient Reasoning in Large Language Models (Yu Bin, 2025) View paper
  - Reasoning Pattern Guidance (2 papers)
  - [12] Don't Think Longer, Think Wisely: Optimizing Thinking Dynamics for Large Reasoning Models (Anï¼ So-Hyun, 2025) View paper
  - [20] Let LLMs Break Free from Overthinking via Self-Braking Tuning (H Zhao, 2025) View paper
- Inference-Time Optimization
  - Reasoning Compression and Pruning (2 papers)
  - [17] Do Thinking Tokens Help or Trap? Towards More Efficient Large Reasoning Model (Ding, 2025) View paper
  - [49] ThinkLess: A Training-Free Inference-Efficient Method for Reducing Reasoning Redundancy (Gao Yifeng, 2025) View paper
  - Reasoning Intervention and Steering (1 papers)
  - [41] Effectively Controlling Reasoning Models through Thinking Intervention (Wu, 2025) View paper
  - Fast-Slow Reasoning Mode Switching (2 papers)
  - [11] Fast Thinking for Large Language Models (Zheng Haoyu, 2025) View paper
  - [46] OThink-R1: Intrinsic Fast/Slow Thinking Mode Switching for Over-Reasoning Mitigation (Zhang Sheng-jia, 2025) View paper
- Efficiency Enhancement via Model Architecture
  - Layer and Parameter Efficiency (1 papers)
  - [13] Shortgpt: Layers in large language models are more redundant than you expect (Xin Men, 2025) View paper
- Context and Input Optimization
  - Input Context Compression (1 papers)
  - [1] Compressing context to enhance inference efficiency of large language models (Yucheng LI, 2023) View paper
- System-Level Inference Optimization
  - Multi-Model and Tiered Inference Systems (2 papers)
  - [4] Tabi: An efficient multi-level inference system for large language models (Yiding Wang, 2023) View paper
  - [22] Ce-collm: Efficient and adaptive large language models through cloud-edge collaboration (Hongpeng Jin, 2025) View paper
  - Batching and Scheduling Optimization (2 papers)
  - [9] Baton: Enhancing batch-wise inference efficiency for large language models via dynamic re-batching (Peizhuang Cong, 2025) View paper
  - [23] Inference without interference: Disaggregate llm inference for mixed downstream workloads (Cunchen Hu, 2024) View paper
  - Memory and Hardware Acceleration (2 papers)
  - [18] Llm in a flash: Efficient large language model inference with limited memory (Keivan Alizadeh, 2024) View paper
  - [43] MCBP: A memory-compute efficient LLM inference accelerator leveraging bit-slice-enabled sparsity and repetitiveness (Wang Hui-zheng, 2025) View paper
  - Speculative and Collaborative Inference (1 papers)
  - [40] Collaborative speculative inference for efficient llm inference serving (Gao Luyao, 2025) View paper
- Domain-Specific and Application-Oriented Efficiency
  - Retrieval-Augmented Reasoning Efficiency (1 papers)
  - [35] ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation (Cao, 2025) View paper
  - Task-Specific Reasoning Optimization (2 papers)
  - [25] Don't" Overthink" Passage Reranking: Is Reasoning Truly Necessary? (Chuang, 2025) View paper
  - [50] Optimizing llm queries in relational workloads (Shuming Liu, 2024) View paper
  - Multimodal Reasoning Efficiency (2 papers)
  - [7] Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models (Huang Wenxuan, 2025) View paper
  - [39] Mitigating Visual Knowledge Forgetting in MLLM Instruction-tuning via Modality-decoupled Gradient Descent (Junda Wu, 2025) View paper
- Prompting and In-Context Learning for Efficiency
  - Prompting Strategy Impact on Reasoning Efficiency (1 papers)
  - [36] Innate Reasoning is Not Enough: In-Context Learning Enhances Reasoning Large Language Models with Less Overthinking (Ge Yuyao, 2025) View paper
- Comprehensive Surveys and Frameworks
  - Reasoning Efficiency Surveys (3 papers)
  - [24] Harnessing the reasoning economy: A survey of efficient reasoning for large language models (Wang Rui, 2025) View paper
  - [42] Efficient inference for large reasoning models: A survey (Liu Yu-e, 2025) View paper
  - [48] Efficient reasoning models: A survey (Feng Sicheng, 2025) View paper
- Emerging and Cross-Cutting Approaches
  - Inherent Efficiency Exploration (1 papers)
  - [47] Exploring and Exploiting the Inherent Efficiency within Large Reasoning Models for Self-Guided Efficiency Enhancement (Zhao WeiXiang, 2025) View paper
  - General Adaptive Compute Frameworks (1 papers)
  - [8] An adaptive compute approach to optimize inference efficiency in large language models (James Lesatod, 2024) View paper
- Safety and Robustness Considerations
  - Safety in Efficient Reasoning Models (1 papers)

- [33] Safety in large reasoning models: A survey (Cheng Wang, 2025) View paper
- Auxiliary System Components
  - Data Drift and Model Maintenance (1 papers)
  - [30] Matchmaker: Data drift mitigation in machine learning for large-scale systems (A Mallick, 2022) View paper

Narrative

Core task: reducing overthinking in large reasoning models. The field has organized itself around a diverse set of strategies that span detection, control, training, and inference-time optimization. At the highest level, one branch focuses on Overthinking Detection and Analysis, identifying when models expend unnecessary computation, while Adaptive Reasoning Control and Training-Based Overthinking Mitigation develop mechanisms to modulate reasoning depth either during training or via reward engineering. Inference-Time Optimization and System-Level Inference Optimization address runtime efficiency through techniques like early exit and dynamic batching, and Efficiency Enhancement via Model Architecture explores structural changes such as layer pruning. Additional branches cover Context and Input Optimization, Domain-Specific and Application-Oriented Efficiency, Prompting and In-Context Learning for Efficiency, and cross-cutting themes in Comprehensive Surveys and Frameworks, Emerging and Cross-Cutting Approaches, Safety and Robustness Considerations, and Auxiliary System Components. Representative works illustrate these directions: Stop Overthinking Survey[5] and Reasoning Economy Survey[24] provide broad overviews, while Difficulty-Adaptive Slow-Thinking[3] and Dynamic Early Exit[6] exemplify adaptive control and inference-time methods.

Within this landscape, a particularly active line of work centers on training-based reward engineering, where models learn to balance reasoning depth against computational cost. Overthinking Reduction[0] sits squarely in this cluster, emphasizing reward signals that discourage excessive deliberation during training. Nearby, REA-RL Reflection-Aware[19] and Reasoning Shaping[32] also leverage reinforcement learning to guide reasoning efficiency, while SmartThinker Step-Level Control[45] introduces finer-grained step-level interventions. These approaches contrast with inference-time methods like Dynamic Early Exit[6] or training-free techniques such as ThinkLess Training-Free[49], which avoid modifying the training objective. The central trade-off across these branches is whether to bake efficiency into the model's learned behavior or to impose it dynamically at test time. Overthinking Reduction[0] aligns closely with the former philosophy, sharing the reward-engineering emphasis of works like REA-RL[19] but differing in the specific signals used to penalize overthinking, thus contributing a distinct perspective on how to shape reasoning economy during the learning phase.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. REA-RL: Reflection-Aware Online Reinforcement Learning for Efficient Large Reasoning Models

Authors: Deng, Hexuan, Jiao, Wenxiang, Hexuan Deng, et al. (13 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Large Reasoning Models (LRMs) demonstrate strong performance in complex tasks but often face the challenge of overthinking, leading to substantially high inference costs. Existing approaches synthesize shorter reasoning responses for LRMs to learn, but are inefficient for online usage due to the time-consuming data generation and filtering processes. Meanwhile, online reinforcement learning mainly adopts a length reward to encourage short reasoning responses, but tends to lose the reflection abi...

Relationship Analysis

Both papers belong to the Reward Engineering for Efficiency category, designing novel reward mechanisms to reduce overthinking in large reasoning models while maintaining performance. They overlap in addressing the fundamental challenge of balancing reasoning quality with efficiency through reward-based approaches in reinforcement learning frameworks. The key difference is that the original paper (DECS) focuses on decoupled token-level rewards that distinguish necessary reasoning prefixes from redundant tokens combined with curriculum data scheduling, while the candidate paper (REA-RL) introduces a reflection-aware approach using a small reflection model and reflection rewards to maintain reflection ability during online training.

---

2. Mitigating Overthinking through Reasoning Shaping

Authors: Song, Feifan, Feifan Song, Gao, Bofei, et al. (26 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large reasoning models (LRMs) boosted by Reinforcement Learning from Verifier Reward (RLVR) have shown great power in problem solving, yet they often cause overthinking: excessive, meandering reasoning that inflates computational cost. Prior designs of penalization in RLVR manage to reduce token consumption while often harming model performance, which arises from the oversimplicity of token-level supervision. In this paper, we argue that the granularity of supervision plays a crucial role in bal...

Relationship Analysis

Both papers belong to the Reward Engineering for Efficiency category, designing novel reward mechanisms to reduce overthinking in large reasoning models while maintaining performance. They share the approach of addressing the misalignment between sequence-level rewards and token-level optimization in RLVR frameworks like GRPO. The key difference is that the original paper (DECS) introduces decoupled token-level rewards that distinguish necessary reasoning prefix (NRP) from redundant tokens with curriculum data scheduling, while the candidate paper (GRSP) proposes step-level segment penalization with length-aware weighting across reasoning segment clusters, operating at a coarser granularity than token-level supervision.

---

3. SmartThinker: Learning to Compress and Preserve Reasoning by Step-Level Length Control

Authors: He Xing-yang, Ling Xiao, Xingyang He, Liu Jie, Xiao Ling, et al. (6 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Large reasoning models (LRMs) have exhibited remarkable reasoning capabilities through inference-time scaling, but this progress has also introduced considerable redundancy and inefficiency into their reasoning processes, resulting in substantial computational waste. Previous work has attempted to mitigate this issue by penalizing the overall length of generated samples during reinforcement learning (RL), with the goal of encouraging a more concise chains of thought. However, we observe that suc...

Relationship Analysis

Both papers belong to the Reward Engineering for Efficiency category, focusing on designing novel reward functions to reduce overthinking in large reasoning models while maintaining performance. They overlap in addressing the misalignment between sequence-level length penalties and token-level optimization, and both propose reward mechanisms that differentiate between necessary and redundant reasoning tokens. The key difference is that the original paper (DECS) uses a decoupled token-level reward based on detecting the Necessary Reasoning Prefix (NRP) combined with curriculum data scheduling, while SmartThinker employs a two-stage approach with step-level length control policy optimization (SCPO) that uses an online importance estimator to allocate length budgets differentially across reasoning steps based on their importance.

Contributions Analysis

Overall novelty summary. The paper introduces DECS, a framework addressing overthinking in large reasoning models through decoupled token-level rewards and curriculum batch scheduling. It resides in the 'Reward Engineering for Efficiency' leaf under 'Training-Based Overthinking Mitigation', alongside three sibling papers (REA-RL Reflection-Aware, Reasoning Shaping, and one other). This leaf represents a moderately populated research direction within a taxonomy of 50 papers across approximately 36 topics, indicating focused but not overcrowded attention to reward-based training solutions for reasoning efficiency.

The taxonomy reveals that reward engineering sits within a broader training-based mitigation branch, distinct from inference-time methods (e.g., early exit mechanisms, reasoning compression) and adaptive control approaches (e.g., difficulty-adaptive allocation). Neighboring leaves include 'Data-Centric Training Strategies' and 'Reasoning Pattern Guidance', which address efficiency through curated datasets and modular reasoning supervision respectively. DECS diverges from these by focusing on token-level reward decomposition rather than data curation or pattern-level guidance, positioning it as a training objective innovation rather than an architectural or data-driven solution.

Among 25 candidates examined across three contributions, no clearly refuting prior work was identified. The theoretical analysis of length-based reward misalignment examined 10 candidates with zero refutations, suggesting this specific framing may be novel within the limited search scope. The DECS framework and curriculum scheduling contributions each examined 5 and 10 candidates respectively, also without refutation. These statistics indicate that within the top-K semantic matches explored, the paper's specific combination of token-level reward decoupling and curriculum strategies appears distinct from existing reward engineering approaches.

Based on the limited literature search of 25 candidates, the work appears to offer a fresh perspective within reward engineering for reasoning efficiency. However, the analysis does not cover exhaustive prior work beyond top-K semantic matches and citation expansion. The taxonomy context suggests the paper contributes to an active but not saturated research direction, with its token-level reward decomposition distinguishing it from sibling works that may employ trajectory-level or step-level reward mechanisms.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Theoretical analysis of misalignment in length-based rewards

Description: The authors provide a theoretical analysis revealing two critical flaws in existing length reward mechanisms: the erroneous penalization of essential exploratory high-entropy tokens and the inadvertent rewarding of partial redundancy. This misalignment between trajectory-level rewards and token-level optimization is shown to degrade reasoning performance and limit efficiency gains.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Towards Flash Thinking via Decoupled Advantage Policy Optimization

URL: View paper

Brief Assessment

Flash Thinking[53] focuses on decoupled advantage policy optimization for reducing overthinking, but does not provide the same theoretical analysis of misalignment between trajectory-level rewards and token-level optimization that characterizes the original contribution.

---

2. From Uniform to Heterogeneous: Tailoring Policy Optimization to Every Token's Nature

URL: View paper

Brief Assessment

Uniform to Heterogeneous[57] focuses on token-level heterogeneity in policy optimization based on entropy, not on the misalignment between sequence-level length penalties and token-level optimization that the original paper analyzes.

---

3. Soft Adaptive Policy Optimization

URL: View paper

Brief Assessment

Soft Adaptive Policy[58] addresses token-level variance and stability in policy optimization through soft gating mechanisms, not the misalignment between sequence-level length penalties and token-level optimization that the original paper analyzes.

---

4. Hierarchical Budget Policy Optimization for Adaptive Reasoning

URL: View paper

Brief Assessment

Hierarchical Budget Policy[55] addresses length penalties through hierarchical budget constraints and reward structures, but does not provide theoretical analysis of token-level vs. trajectory-level misalignment or the specific flaws (penalization of exploratory tokens and rewarding of partial redundancy) identified in the original paper.

---

5. L1: Controlling how long a reasoning model thinks with reinforcement learning

URL: View paper

Brief Assessment

L1 Controlling Thinking[51] focuses on controllable length constraints via reinforcement learning for test-time compute allocation, not on analyzing misalignment between trajectory-level rewards and token-level optimization in existing length penalty mechanisms.

---

6. Tlcr: Token-level continuous reward for fine-grained reinforcement learning from human feedback

URL: View paper

Brief Assessment

TLCR Token-level Reward[59] focuses on token-level continuous rewards for RLHF alignment with human preferences, not on analyzing misalignment between sequence-level length penalties and token-level policy optimization in reasoning models.

---

7. LAPO: Internalizing Reasoning Efficiency via Length-Adaptive Policy Optimization

URL: View paper

Brief Assessment

LAPO Length-Adaptive[54] focuses on learning natural reasoning patterns through statistical distribution of solution lengths and meta-cognitive guidance, rather than analyzing misalignment between trajectory-level rewards and token-level optimization in existing length penalty mechanisms.

---

8. GTPO and GRPO-S: Token and Sequence-Level Reward Shaping with Policy Entropy

URL: View paper

Brief Assessment

GTPO Token-Level Shaping[56] focuses on entropy-weighted token-level rewards for credit assignment in RL, not on analyzing misalignment between sequence-level length penalties and token-level optimization as described in the original contribution.

---

9. Scaling laws for reward model overoptimization in direct alignment algorithms

URL: View paper

Brief Assessment

Reward Model Overoptimization[52] focuses on over-optimization in direct alignment algorithms (DAAs) like DPO, not on length-based reward mechanisms in critic-free RL frameworks like GRPO. The technical contexts are fundamentally different.

---

10. Delve into PPO: Implementation matters for stable RLHF

URL: View paper

Brief Assessment

Delve into PPO[60] focuses on PPO implementation details for RLHF stability (reward model quality, policy constraints, initialization methods) rather than analyzing misalignment between sequence-level length penalties and token-level optimization in reasoning models.

Contribution 2: DECS framework with decoupled token-level rewards

Description: The authors propose DECS, a framework featuring a first-of-its-kind decoupled token-level reward mechanism that surgically distinguishes and penalizes redundant tokens generated after the necessary reasoning prefix, while preserving rewards for essential reasoning steps. This addresses the identified misalignment by operating at the token level rather than sequence level.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Arithmetic accuracy and stability as a function of token reinforcement

URL: View paper

Brief Assessment

Arithmetic Token Reinforcement[75] examines token reinforcement in educational psychology for arithmetic learning, not decoupled token-level reward mechanisms in large language model reasoning optimization.

---

2. DRPO: Efficient Reasoning via Decoupled Reward Policy Optimization

URL: View paper

Brief Assessment

DRPO Decoupled Reward[73] addresses a similar problem (overthinking in reasoning models) but uses a fundamentally different approach: decoupling rewards between correct and incorrect rollouts at the group level, rather than surgically distinguishing redundant tokens after the necessary reasoning prefix at the token level as DECS does.

---

3. On the Current Landscape of Language Model Reward Modeling for Alignment

URL: View paper

Brief Assessment

Reward Modeling Landscape[72] is a thesis on reward modeling for alignment, not a framework for token-level reward mechanisms in reasoning models. The candidate focuses on reward modeling broadly, while the original proposes a specific decoupled token-level reward system for reducing overthinking in reasoning models.

---

4. StepSearch: Igniting LLMs Search Ability via Step-Wise Proximal Policy Optimization

URL: View paper

Brief Assessment

StepSearch Step-Wise PPO[71] focuses on multi-hop reasoning with search-based document retrieval using step-wise rewards for search actions, not on reducing overthinking in reasoning models through decoupled token-level rewards that distinguish necessary reasoning prefixes from redundant tokens.

---

5. UpSkill: Mutual Information Skill Learning for Structured Response Diversity in LLMs

URL: View paper

Brief Assessment

UpSkill Mutual Information[74] focuses on inducing structured response diversity through mutual information skill learning for multi-attempt accuracy (pass@k), not on reducing overthinking or penalizing redundant tokens after reasoning prefixes. The technical approaches and objectives are fundamentally different.

Contribution 3: Curriculum batch scheduling strategy

Description: The authors introduce a dynamic batching strategy that adaptively adjusts the proportion of easy prompts in training batches based on the current NRP ratio. This curriculum approach mitigates over-penalization of exploratory behavior and maintains the balance between reasoning efficiency and model capability throughout training.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Curriculum learning by optimizing learning dynamics

URL: View paper

Brief Assessment

Optimizing Learning Dynamics[70] focuses on selecting training samples based on residual and linear temporal dynamics to maximize learning progress, not on dynamically adjusting the proportion of easy prompts based on NRP ratio as in the original paper's curriculum batch scheduling strategy.

---

2. Deep curriculum learning for polsar image classification

URL: View paper

Brief Assessment

Deep Curriculum PolSAR[68] focuses on PolSAR image classification using entropy-alpha decomposition to order training patches by complexity, not on balancing efficiency and efficacy in reasoning model training through dynamic NRP-based batch scheduling.

---

3. When do curricula work?

URL: View paper

Brief Assessment

When Curricula Work[69] studies curriculum learning for image classification with fixed datasets, focusing on sample ordering by difficulty. The original paper's dynamic batch scheduling adjusts easy prompt proportions based on NRP ratios during RL training for reasoning models—a fundamentally different application domain and mechanism.

---

4. Reinforcement learning for the adaptive scheduling of educational activities

URL: View paper

Brief Assessment

Adaptive Scheduling Activities[67] focuses on scheduling educational activities for online courses using RL to maximize learning gains, not on curriculum batch scheduling for training language models with dynamic prompt difficulty adjustment.

---

5. Prompt Curriculum Learning for Efficient LLM Post-Training

URL: View paper

Brief Assessment

Prompt Curriculum Learning[62] focuses on selecting intermediate-difficulty prompts using a value model for training efficiency, not on dynamically adjusting the proportion of easy prompts based on NRP ratios to balance reasoning efficiency and model capability as in the original paper.

---

6. Curriculum guided reinforcement learning for efficient multi hop retrieval augmented generation

URL: View paper

Brief Assessment

Curriculum Guided Retrieval[65] applies curriculum learning to multi-hop retrieval-augmented generation tasks with dynamic reward scheduling, not to reasoning model training with batch composition adjustments based on NRP ratios.

---

7. Dump: Automated distribution-level curriculum learning for rl-based llm post-training

URL: View paper

Brief Assessment

Dump Distribution-level Curriculum[63] focuses on distribution-level curriculum learning across diverse data sources and difficulties using UCB-based sampling, while the original paper's curriculum batch scheduling adaptively adjusts easy prompt proportions based on NRP ratios within batches to balance efficiency and capability during training.

---

8. Research and Implementation of Education Resource Scheduling Algorithm Based on Machine Learning

URL: View paper

Brief Assessment

Education Resource Scheduling[66] focuses on educational resource allocation and classroom scheduling using machine learning, not on reinforcement learning training dynamics or adaptive batch composition strategies for reasoning models.

---

9. Curriculum learning: A survey

URL: View paper

Brief Assessment

Curriculum Learning Survey[61] discusses general curriculum learning principles for ordering training samples from easy to hard. The original paper's contribution is a specific dynamic batching strategy that adaptively adjusts easy prompt proportions based on NRP ratio during RL training for reasoning models, which is a distinct technical approach not addressed in this survey.

---

10. Emergent mechanisms for long timescales depend on training curriculum and affect performance in memory tasks

URL: View paper

Brief Assessment

Emergent Mechanisms Timescales[64] focuses on curriculum learning for training RNNs on memory tasks by adjusting neuron timescales and network connectivity, not on batch scheduling strategies for balancing efficiency and efficacy in large language model training with reinforcement learning.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Overthinking Reduction with Decoupled Rewards and Curriculum Data Scheduling View paper
- [1] Compressing context to enhance inference efficiency of large language models View paper
- [2] Self-training elicits concise reasoning in large language models View paper
- [3] DAST: Difficulty-Adaptive Slow-Thinking for Large Reasoning Models View paper
- [4] Tabi: An efficient multi-level inference system for large language models View paper
- [5] Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models View paper
- [6] Dynamic Early Exit in Reasoning Models View paper
- [7] Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models View paper
- [8] An adaptive compute approach to optimize inference efficiency in large language models View paper
- [9] Baton: Enhancing batch-wise inference efficiency for large language models via dynamic re-batching View paper
- [10] Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms View paper
- [11] Fast Thinking for Large Language Models View paper
- [12] Don't Think Longer, Think Wisely: Optimizing Thinking Dynamics for Large Reasoning Models View paper
- [13] Shortgpt: Layers in large language models are more redundant than you expect View paper
- [14] ARM: Adaptive Reasoning Model View paper
- [15] S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models View paper
- [16] Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs View paper

- [17] Do Thinking Tokens Help or Trap? Towards More Efficient Large Reasoning Model View paper
- [18] Llm in a flash: Efficient large language model inference with limited memory View paper
- [19] REA-RL: Reflection-Aware Online Reinforcement Learning for Efficient Large Reasoning Models View paper
- [20] Let LLMs Break Free from Overthinking via Self-Braking Tuning View paper
- [21] Mitigating Overthinking in Large Reasoning Models via Manifold Steering View paper
- [22] Ce-collm: Efficient and adaptive large language models through cloud-edge collaboration View paper
- [23] Inference without interference: Disaggregate llm inference for mixed downstream workloads View paper
- [24] Harnessing the reasoning economy: A survey of efficient reasoning for large language models View paper
- [25] Don't" Overthink" Passage Reranking: Is Reasoning Truly Necessary? View paper
- [26] MUR: Momentum Uncertainty guided Reasoning for Large Language Models View paper
- [27] O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning View paper
- [28] Long-Short Chain-of-Thought Mixture Supervised Fine-Tuning Eliciting Efficient Reasoning in Large Language Models View paper
- [29] Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification View paper
- [30] Matchmaker: Data drift mitigation in machine learning for large-scale systems View paper
- [31] AutoL2S: Auto Long-Short Reasoning for Efficient Large Language Models View paper
- [32] Mitigating Overthinking through Reasoning Shaping View paper
- [33] Safety in large reasoning models: A survey View paper
- [34] Stop spinning wheels: Mitigating llm overthinking via mining patterns for early reasoning exit View paper
- [35] ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation View paper
- [36] Innate Reasoning is Not Enough: In-Context Learning Enhances Reasoning Large Language Models with Less Overthinking View paper
- [37] The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks View paper
- [38] The Price of a Second Thought: On the Evaluation of Reasoning Efficiency in Large Language Models View paper
- [39] Mitigating Visual Knowledge Forgetting in MLLM Instruction-tuning via Modality-decoupled Gradient Descent View paper
- [40] Collaborative speculative inference for efficient llm inference serving View paper
- [41] Effectively Controlling Reasoning Models through Thinking Intervention View paper
- [42] Efficient inference for large reasoning models: A survey View paper
- [43] MCBP: A memory-compute efficient LLM inference accelerator leveraging bit-slice-enabled sparsity and repetitiveness View paper
- [44] Optimalthinkingbench: Evaluating over and underthinking in llms View paper
- [45] SmartThinker: Learning to Compress and Preserve Reasoning by Step-Level Length Control View paper
- [46] OThink-R1: Intrinsic Fast/Slow Thinking Mode Switching for Over-Reasoning Mitigation View paper
- [47] Exploring and Exploiting the Inherent Efficiency within Large Reasoning Models for Self-Guided Efficiency Enhancement View paper
- [48] Efficient reasoning models: A survey View paper
- [49] ThinkLess: A Training-Free Inference-Efficient Method for Reducing Reasoning Redundancy View paper
- [50] Optimizing llm queries in relational workloads View paper
- [51] L1: Controlling how long a reasoning model thinks with reinforcement learning View paper
- [52] Scaling laws for reward model overoptimization in direct alignment algorithms View paper
- [53] Towards Flash Thinking via Decoupled Advantage Policy Optimization View paper
- [54] LAPO: Internalizing Reasoning Efficiency via Length-Adaptive Policy Optimization View paper
- [55] Hierarchical Budget Policy Optimization for Adaptive Reasoning View paper
- [56] GTPO and GRPO-S: Token and Sequence-Level Reward Shaping with Policy Entropy View paper
- [57] From Uniform to Heterogeneous: Tailoring Policy Optimization to Every Token's Nature View paper
- [58] Soft Adaptive Policy Optimization View paper
- [59] Tlcr: Token-level continuous reward for fine-grained reinforcement learning from human feedback View paper
- [60] Delve into PPO: Implementation matters for stable RLHF View paper
- [61] Curriculum learning: A survey View paper
- [62] Prompt Curriculum Learning for Efficient LLM Post-Training View paper
- [63] Dump: Automated distribution-level curriculum learning for rl-based llm post-training View paper
- [64] Emergent mechanisms for long timescales depend on training curriculum and affect performance in memory tasks View paper
- [65] Curriculum guided reinforcement learning for efficient multi hop retrieval augmented generation View paper
- [66] Research and Implementation of Education Resource Scheduling Algorithm Based on Machine Learning View paper
- [67] Reinforcement learning for the adaptive scheduling of educational activities View paper
- [68] Deep curriculum learning for polsar image classification View paper
- [69] When do curricula work? View paper
- [70] Curriculum learning by optimizing learning dynamics View paper
- [71] StepSearch: Igniting LLMs Search Ability via Step-Wise Proximal Policy Optimization View paper
- [72] On the Current Landscape of Language Model Reward Modeling for Alignment View paper
- [73] DRPO: Efficient Reasoning via Decoupled Reward Policy Optimization View paper
- [74] UpSkill: Mutual Information Skill Learning for Structured Response Diversity in LLMs View paper
- [75] Arithmetic accuracy and stability as a function of token reinforcement View paper