Novelty Assessment Report

Paper: The Markovian Thinker

PDF URL: https://openreview.net/pdf?id=3As6AQ9ELI

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-05

Abstract

Reasoning LLMs suffer from quadratic compute growth as their context length increases, making reinforcement learning with verifiable rewards (RLVR) and test-time scaling prohibitively expensive. Prior work has tried to lighten the computational burden by shortening reasoning traces through pruning, summarization, or multi-stage training, but these methods remain bound to quadratic costs. We introduce Delethink, a thinking algorithm that realizes the Markovian Thinking Paradigm. Instead of producing one long monolithic reasoning trace, Delethink thinks in a sequence of chunks, the Delethink trace. Each chunk continues reasoning by referring only to a fixed number of prior tokens, which functions as a Markovian state sufficient for progressing reasoning, while deleting the rest. This preserves continuity without carrying the quadratic baggage. As a result, compute scales linearly and peak memory remains constant. In experiments, we show that Delethink can be applied directly to off-the-shelf reasoning models ranging from $1.5\textnormal{B}$ to $30\textnormal{B}$ parameters, with no loss in performance. Extended reasoning becomes possible under fixed memory and linear compute, while enabling efficient RL training on new tasks. On the DeepScaleR dataset, Delethink trains R1DistillQwen1.5B to the same benchmark performance as a standard long chain-of-thought (LongCoT) approach, where both models generate up to $24\textnormal{k}$ thinking tokens. The difference is efficiency. Delethink reasons 40% faster with 70% less memory footprint. By decoupling reasoning length from context length, the Markovian Thinking paradigm opens the door to next-generation reasoning LLMs that can scale to millions of tokens with linear compute and constant memory.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Efficient Reasoning with Linear Compute and Constant Memory**

A total of **50 papers** were analyzed and organized into a taxonomy with **29 categories**.

Taxonomy Overview

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

- **Linear-Complexity Attention Mechanisms**
- **State-Space and Recurrent Architectures**
- **Memory-Augmented Architectures**
- **Markovian and Chunked Reasoning Paradigms**
- **Algorithmic and Data Structure Foundations**
- **Hardware and Systems Optimization**
- **Specialized Application Domains**

Complete Taxonomy Tree

- Efficient Reasoning with Linear Compute and Constant Memory Survey Taxonomy
- Linear-Complexity Attention Mechanisms
  - Low-Rank and Projection-Based Attention (2 papers)
  - [1] Linformer: Self-Attention with Linear Complexity (Wang Si-nong, 2025) View paper
  - [15] Linearizing Transformer with Key-Value Memory (Cai, 2022) View paper
  - Alternative Similarity Functions (1 papers)
  - [9] Cottention: Linear Transformers With Cosine Attention (Gabriel Mongaras, 2024) View paper
  - Kernel-Based Linear Attention (1 papers)
  - [28] Sub-linear memory: How to make performers slim (Valerii Likhosherstov, 2021) View paper
  - Sparse and Hierarchical Attention Patterns (2 papers)
  - [35] Memory-efficient Transformers via Top- Attention (A Gupta, 2021) View paper
  - [43] Hardware-aligned Hierarchical Sparse Attention for Efficient Long-term Memory Access (Hu Xiang, 2025) View paper
  - Domain-Specific Linear Attention Applications (3 papers)
  - [4] FLASepformer: Efficient Speech Separation with Gated Focused Linear Attention Transformer (Wang Hao-Xu, 2025) View paper
  - [7] JEDI-linear: Fast and Efficient Graph Neural Networks for Jet Tagging on FPGAs (Que, 2025) View paper
  - [13] GLULA: Linear attention-based model for efficient human activity recognition from wearable sensors (Aldiyar Bolatov, 2024) View paper
  - Reduced-Parameter Attention Variants (1 papers)
  - [22] Cost-Effective Attention Mechanisms for Low Resource Settings: Necessity&Sufficiency of Linear Transformations (Hosseini, 2024) View paper
- State-Space and Recurrent Architectures
  - Pure State-Space Models (3 papers)
  - [8] IMSSA: Deploying modern state-space models on memristive in-memory compute hardware (Siegel, 2024) View paper
  - [10] Apriel-H1: Towards Efficient Enterprise Reasoning Models (Ostapenko, 2025) View paper

  - [45] LongMamba: Enhancing Mamba's Long-Context Capabilities via Training-Free Receptive Field Enlargement (Ye, 2025) View paper
  - Hybrid State-Space and Attention Models (1 papers)
  - [2] RWKV-X: A Linear Complexity Hybrid Language Model (Hou Haowen, 2025) View paper
  - Recurrent Models with Linear Complexity (2 papers)
  - [19] Real Time Recurrent Learning with Complex-Valued Trace Units (Elelimy, 2023) View paper
  - [49] xLSTM 7B: A Recurrent LLM for Fast and Efficient Inference (Beck Maximilian, 2025) View paper
  - Prefix-Scannable and Dual-Mode Models (1 papers)
  - [5] Sequential-Parallel Duality in Prefix Scannable Models (Yau, 2025) View paper
  - Kalman Filter-Based Architectures (1 papers)
  - [48] Gated KalmaNet: A Fading Memory Layer Through Test-Time Ridge Regression (Liangzu Peng, 2025) View paper
- Memory-Augmented Architectures
  - External Dynamic Memory Systems (1 papers)
  - [29] Memformer: A memory-augmented transformer for sequence modeling (Qingyang Wu, 2022) View paper
  - Bounded Memory Buffers (1 papers)
  - [21] Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks (Toshniwal, 2020) View paper
  - Infinite Memory Transformers (1 papers)
  - [6] â□□-former: Infinite memory transformer (Marinho, 2022) View paper
  - Recurrent Memory Transformers (1 papers)
  - [44] Diagonal Batching Unlocks Parallelism in Recurrent Memory Transformers for Long Contexts (Kuzmin, 2025) View paper
  - Vector-Quantized Memory Structures (1 papers)
  - [20] LongVQ: Long Sequence Modeling with Vector Quantization on Structured Memory (Liu, 2024) View paper
- Markovian and Chunked Reasoning Paradigms ★ (2 papers)
  - [0] The Markovian Thinker (Anon et al., 2026) View paper
  - [12] Working-Memory-Correct Long-Horizon Expert-Retrieval TTT Dialogue (Ghosh, 2025) View paper
- Algorithmic and Data Structure Foundations
  - Streaming and Online Algorithms (3 papers)
  - [24] Contraction Clustering (RASTER): A Very Fast Big Data Algorithm for Sequential and Parallel Density-Based Clustering in Linear Time, Constant Memory, and a Single Pass (Gregor Ulm, 2022) View paper
  - [27] Fully Online Grammar Compression in Constant Space (Maruyama, 2022) View paper
  - [41] Sequential Nonparametric Testing with the Law of the Iterated Logarithm (Akshay Balsubramani, 2022) View paper
  - Batch Processing Algorithms (3 papers)
  - [31] pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree (Matsen, 2010) View paper
  - [33] Computing Wasserstein-P Distance Between Images with Linear Cost (Yidong Chen, 2022) View paper
  - [39] Efficient Sparseness-Enforcing Projections (Thom, 2022) View paper
  - Specialized Linear-Time Algorithms (4 papers)
  - [18] A linear complexity phasing method for thousands of genomes (Olivier Delaneau, 2012) View paper
  - [23] A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements (MarÃa Luisa MicÃ³, 1994) View paper
  - [38] Internal Shortest Absent Word Queries in Constant Time and Linear Space (Badkobeh, 2022) View paper
  - [40] Childrenâ□□s Basic Etiquette Learning in Mobile Application (Istiono, 2022) View paper
- Hardware and Systems Optimization
  - Memory Compression Techniques (1 papers)
  - [25] Linearly compressed pages: A low-complexity, low-latency main memory compression framework (Gennady Pekhimenko, 2013) View paper
  - Hardware-Accelerated Inference (1 papers)
  - [46] LogSumExp: Efficient Approximate Logarithm Acceleration for Embedded Tractable Probabilistic Reasoning (Lingyun Yao, 2025) View paper
  - Modular and Localized Computation (1 papers)
  - [3] Leveraging mesh modularization to lower the computational cost of localized updates to regional 2D hydrodynamic model outputs (M. Garcia, 2023) View paper
- Specialized Application Domains
  - Energy and Infrastructure Systems (1 papers)
  - [11] Quantifying the Impact of Energy System Model Resolution on Siting, Cost, Reliability, and Emissions (Mauzerall, 2024) View paper
  - Control Theory and Dynamical Systems (1 papers)
  - [26] Finiteâ□□time output regulation by bounded linear timeâ□□varying controls with applications to the satellite formation flying (Kai Zhang, 2021) View paper
  - Scientific Computing and Simulation (2 papers)
  - [30] A fast computational framework for the linear bond-based peridynamic model (Chenguang Liu, 2023) View paper
  - [42] Cinematic-L1 Video Stabilization with a Log-Homography Model (Arwen Bradley, 2021) View paper
  - Combinatorial Optimization Problems (2 papers)
  - [14] Bottleneck spanning tree interdiction problem with fixed and linear costs (Amir Abdolahzadeh, 2023) View paper
  - [34] High Memory Masked Convolutional Codes for PQC (Meir Ariel, 2025) View paper
  - Software Verification and Testing (2 papers)
  - [17] Predictive Monitoring against Pattern Regular Languages (Zhendong Ang, 2024) View paper
  - [50] Enhanced Data Race Prediction Through Modular Reasoning (Farzan, 2025) View paper
  - Learning and Inference Systems (5 papers)
  - [16] A Solution for Large Scale Nonlinear Regression with High Rank and Degree at Constant Memory Complexity via Latent Tensor Reconstruction (SzedmÃ¡k SÃ¡ndor, 2022) View paper
  - [32] Approximating Median Absolute Deviation with Bounded Error (Zhiwei Chen, 2021) View paper

    - [36] A Learning-Based Framework for Memory-Bounded Heuristic Search: First Results (Carlos HernÃ¡ndez Ulloa, 2021) View paper
    - [37] Semi-Parametric Deep Neural Networks in Linear Time and Memory (Rastogi, 2022) View paper
    - [47] Fast linear-space computations of longest common subsequences (Alberto Apostolico, 1992) View paper

Narrative

Core task: efficient reasoning with linear compute and constant memory. The field addresses the fundamental challenge of scaling neural architectures without incurring quadratic attention costs or unbounded memory growth. The taxonomy reveals several complementary strategies: Linear-Complexity Attention Mechanisms explore approximations and sparse patterns to reduce computational overhead (e.g., Linformer[1], Hierarchical Sparse Attention[43]); State-Space and Recurrent Architectures revive sequential models with modern parameterizations (RWKV-X[2], xLSTM[49], LongMamba[45]); Memory-Augmented Architectures introduce external storage to decouple context length from parameter count (Infinite Memory Transformer[6], Memformer[29]); and Markovian and Chunked Reasoning Paradigms impose structured constraints on information flow to maintain bounded state. Additional branches cover algorithmic foundations, hardware optimization, and specialized domains, reflecting the breadth of efficiency concerns across theory and practice.

A particularly active tension emerges between architectures that preserve full attention expressiveness through clever approximations versus those that embrace explicit memory constraints or Markovian assumptions. Works like Sequential Parallel Duality[5] and JEDI Linear[7] seek to retain transformer-like flexibility while achieving linear scaling, whereas approaches in the Markovian branch accept bounded context windows in exchange for provable constant-memory guarantees. Markovian Thinker[0] sits squarely within this latter paradigm, emphasizing structured reasoning under strict memory limits—a philosophy it shares with Working Memory Dialogue[12], which similarly explores bounded-state dialogue systems. Compared to memory-augmented methods like Infinite Memory Transformer[6] that dynamically manage external storage, Markovian Thinker[0] opts for a more constrained, predictable footprint, trading off potential long-range recall for deterministic efficiency. This positioning highlights an ongoing debate: whether future scalability lies in smarter approximations of full attention or in fundamentally rethinking what information must be retained.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Working-Memory-Correct Long-Horizon Expert-Retrieval TTT Dialogue

Authors: DP Ghosh | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ form to improve long-range modeling with bounded memory growth (Rae et al., 2019). â¦ Our DST/TST lineage claims that skewed architectures achieve linear complexity and can â¦

Relationship Analysis

Both papers belong to the Markovian and Chunked Reasoning Paradigms category, focusing on methods that decompose reasoning into chunks with Markovian state transitions to achieve linear compute and constant memory. While the original paper (The Markovian Thinker) introduces Delethink, which explicitly deletes past reasoning tokens and maintains only a fixed-size markovian state across chunks for extended reasoning, the candidate paper (Working-Memory-Correct Long-Horizon Expert-Retrieval TTT Dialogue) appears to focus on dialogue systems with expert retrieval and working memory mechanisms. The key difference is that the original paper targets general reasoning tasks with a deletion-based chunking approach, whereas the candidate emphasizes dialogue-specific architectures with retrieval components and bounded memory growth through different architectural choices.

Contributions Analysis

Overall novelty summary. The paper introduces Delethink, a thinking algorithm implementing the Markovian Thinking Paradigm for reasoning LLMs. It decomposes reasoning into sequential chunks where each chunk references only a fixed number of prior tokens as Markovian state, deleting the rest to achieve linear compute and constant memory. Within the taxonomy, this work resides in the 'Markovian and Chunked Reasoning Paradigms' leaf, which contains only two papers total. This represents a sparse, emerging research direction focused specifically on structured reasoning under strict memory constraints, contrasting with the more populated branches addressing attention mechanisms or state-space models.

The taxonomy reveals that neighboring approaches pursue efficiency through different mechanisms. The 'State-Space and Recurrent Architectures' branch (containing models like Mamba and xLSTM across five subcategories) achieves linear complexity through sequential state updates but maintains implicit memory in parameters. The 'Memory-Augmented Architectures' branch (five subcategories including Infinite Memory Transformer and Memformer) uses external storage to extend context dynamically. Delethink diverges by imposing explicit Markovian constraints on reasoning traces rather than approximating full attention or managing external memory, positioning it as a fundamentally different paradigm for bounded-state reasoning.

Among thirty candidates examined, the contribution-level analysis shows varied novelty. The core Markovian Thinking Paradigm and zero-shot Delethink inference examined ten candidates each with zero refutations, suggesting these contributions occupy relatively unexplored territory within the limited search scope. However, the Delethink training for reinforcement learning examined ten candidates and found one refutable overlap, indicating more substantial prior work in efficient RL training methods. This pattern suggests the paradigm itself is novel while its application to RL training connects to existing efficiency techniques in that domain.

Based on the limited search of thirty semantically similar papers, Delethink appears to introduce a distinctive approach to reasoning efficiency. The sparse population of its taxonomy leaf and low refutation rates for core contributions suggest novelty, though the analysis cannot claim exhaustiveness. The single refutation in RL training highlights that while the Markovian paradigm is fresh, its integration with established training methods naturally encounters existing work in that intersection.

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

Contribution 1: Markovian Thinking Paradigm and Delethink Algorithm

Description: The authors propose a new reasoning paradigm where models think in fixed-size chunks, retaining only a minimal markovian state from prior reasoning. This enables linear compute scaling and constant memory usage during both training and inference, in contrast to the quadratic costs of standard long chain-of-thought approaches.

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. JanusVLN: Decoupling Semantics and Spatiality with Dual Implicit Memory for Vision-Language Navigation

URL: View paper

Brief Assessment

JanusVLN[53] addresses vision-language navigation with implicit neural memory for spatial-geometric and visual-semantic representations, not reasoning in fixed-size chunks with markovian states for LLM thinking tokens.

2. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads

URL: View paper

Brief Assessment

DuoAttention[57] focuses on efficient KV cache management for long-context LLM inference through attention head optimization, not on reasoning paradigms or fixed-size chunk-based thinking with markovian states during training and inference.

3. Lserve: Efficient long-sequence llm serving with unified sparse attention

URL: View paper

Brief Assessment

Lserve[55] focuses on sparse attention mechanisms for efficient LLM serving in long-sequence contexts, not on reasoning paradigms or training methods. The candidate addresses system-level optimizations for attention computation, while the original contribution concerns a novel reasoning approach with fixed-size chunks and constant memory during reasoning.

4. Titans: Learning to memorize at test time

URL: View paper

Brief Assessment

Titans[51] focuses on learning to memorize historical context through a neural long-term memory module that stores information in network parameters, not on reasoning in fixed-size chunks with markovian states during inference as proposed in the original paper.

5. Trellis: Learning to Compress Key-Value Memory in Attention Models

URL: View paper

Brief Assessment

Trellis[59] addresses memory compression in attention mechanisms for transformers, focusing on bounded KV cache size. The original paper proposes a reasoning paradigm where models think in fixed-size chunks with markovian state retention for extended reasoning tasks, which is a fundamentally different problem domain than architectural memory compression.

6. Artificial hippocampus networks for efficient long-context modeling

URL: View paper

Brief Assessment

Artificial Hippocampus[56] focuses on compressing historical information in transformers using RNN-like architectures for long-context modeling, not on reasoning in fixed-size chunks during chain-of-thought generation. The candidate addresses memory management in attention mechanisms, while the original addresses reasoning trace generation with constant memory.

7. Lococo: Dropping in convolutions for long context compression

URL: View paper

Brief Assessment

Lococo[58] focuses on KV cache compression for long-context processing in LLMs using convolutional kernels, not on reasoning paradigms or fixed-size chunk-based thinking with markovian states for chain-of-thought generation.

8. MEDCnet: A Memory Efficient Approach for Processing Highâ□□Resolution Fundus Images for Diabetic Retinopathy Classification Using CNN

URL: View paper

Brief Assessment

MEDCnet[60] addresses memory-efficient processing of high-resolution medical images for diabetic retinopathy classification using a divide-and-conquer approach with CNNs. This is fundamentally different from the original paper's focus on reasoning in language models with fixed-size chunks and markovian states for linear compute scaling.

9. Random-access infinite context length for transformers

URL: View paper

Brief Assessment

Random Access Context[54] focuses on extending transformer context length through landmark tokens for block retrieval, not on reasoning in fixed-size chunks with markovian states. The candidate addresses memory efficiency in attention mechanisms, while the original proposes a reasoning paradigm for chain-of-thought generation.

10. Scaling Reasoning without Attention

URL: View paper

Brief Assessment

Scaling Reasoning[52] focuses on architectural efficiency through state space models (Mamba-2) that eliminate attention mechanisms entirely, not on chunked reasoning with markovian states. The candidate addresses memory efficiency through architecture replacement rather than reasoning trace restructuring.

Contribution 2: Delethink Inference for Zero-Shot Markovian Thinking

Description: The authors present an inference method that can be applied directly to existing reasoning models without additional training or prompting, allowing them to function as Markovian thinkers by reasoning in chunks while maintaining fixed context size.

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. Reasoning with large language models, a survey

URL: View paper

Brief Assessment

Reasoning Survey[79] is a broad survey paper covering multi-step reasoning approaches in LLMs, focusing on chain-of-thought methods and their taxonomy. It does not describe zero-shot inference methods for chunked reasoning with fixed context sizes like Delethink.

2. A comprehensive survey of prompt engineering techniques in large language models

URL: View paper

Brief Assessment

Prompt Engineering Survey[73] focuses on prompt engineering techniques for LLMs, not on zero-shot inference methods for chunked reasoning or Markovian thinking paradigms. The candidate's limited context mentions logical reasoning and query decomposition but does not address the specific contribution of enabling models to function as Markovian thinkers through fixed-context chunked reasoning without additional training.

3. SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning

URL: View paper

Brief Assessment

SelfCheck[78] focuses on zero-shot verification of step-by-step reasoning chains through regeneration and comparison, not on chunked reasoning with fixed context sizes. The candidate addresses error detection in reasoning steps, while the original contribution concerns inference methods that maintain constant memory by reasoning in chunks.

4. Decomposed Prompting: A Modular Approach for Solving Complex Tasks

URL: View paper

Brief Assessment

Decomposed Prompting[80] focuses on decomposing complex tasks into simpler sub-tasks using modular prompts for different components, not on zero-shot inference methods that enable chunked reasoning with fixed context size in existing models without additional training.

5. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models

URL: View paper

Brief Assessment

Least to Most[71] focuses on decomposing complex problems into simpler subproblems through prompting, not on zero-shot inference methods for chunked reasoning with fixed context size. The candidate addresses compositional generalization through problem decomposition, while the original contribution concerns memory-efficient reasoning via Markovian state management.

6. Recursive decomposition of logical thoughts: Framework for superior reasoning and knowledge propagation in large language models

URL: View paper

Brief Assessment

Recursive Decomposition[75] focuses on decomposing reasoning tasks into hierarchical complexity levels (easy, intermediate, final) with a knowledge propagation module that tracks selected and rejected thoughts. This differs fundamentally from the original paper's zero-shot inference method that enables chunked reasoning with fixed context size through Markovian state management.

7. PAL: Program-aided Language Models

URL: View paper

Brief Assessment

PAL[76] focuses on generating Python programs to solve mathematical and symbolic reasoning problems by offloading computation to an interpreter. It does not address zero-shot chunked reasoning or Markovian thinking paradigms for managing context length in reasoning models.

8. A systematic survey of prompt engineering in large language models: Techniques and applications

URL: View paper

Brief Assessment

Prompt Engineering Techniques[77] is a survey paper cataloging various prompting methods (zero-shot, few-shot, CoT, etc.) for LLMs. It does not describe any inference method for chunked reasoning or Markovian thinking paradigms. The candidate focuses on prompt design techniques, not architectural or inference-level modifications for managing context length during reasoning.

9. Lisa: Reasoning segmentation via large language model

URL: View paper

Brief Assessment

Lisa[74] focuses on reasoning segmentation tasks using multimodal LLMs to generate segmentation masks from implicit queries. It does not address zero-shot inference methods for chunked reasoning or Markovian thinking paradigms in language models.

10. Art: Automatic multi-step reasoning and tool-use for large language models

URL: View paper

Brief Assessment

ART[72] focuses on multi-step reasoning with tool-use through structured prompting and task libraries, not on zero-shot inference methods for chunked reasoning that maintain fixed context size during generation.

Contribution 3: Delethink Training for Efficient Reinforcement Learning

Description: The authors develop a reinforcement learning training procedure that explicitly trains models to reason in the Markovian manner. This approach achieves comparable performance to standard long chain-of-thought training while requiring significantly less compute and memory resources.

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. Training language models to reason efficiently

URL: View paper

Brief Assessment

Training Efficient Reasoning[65] focuses on reducing inference costs by training models to produce shorter chain-of-thoughts while maintaining accuracy, using a length-penalized reward function. The original paper's Delethink training addresses a different problem: enabling linear-time RL training through Markovian chunking of reasoning traces, not primarily shortening them.

2. LLM-Guided Reinforcement Learning for Interactive Environments

URL: View paper

Brief Assessment

LLM Guided RL[69] focuses on using LLMs for high-level planning and subgoal decomposition in interactive environments, not on efficient chain-of-thought training with reduced compute through Markovian reasoning paradigms.

3. GRPO-MA: Multi-Answer Generation in GRPO for Stable and Efficient Chain-of-Thought Training

URL: View paper

Brief Assessment

GRPO-MA[70] focuses on improving the GRPO algorithm through multi-answer generation to address gradient coupling and variance issues, not on reducing compute through Markovian reasoning or memory-efficient training architectures.

4. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning

URL: View paper

Brief Assessment

L1[64] focuses on length-controlled policy optimization to control reasoning length via user-specified constraints, not on Markovian thinking with chunk-based deletion. The original paper's delethink training explicitly trains models to reason in chunks with fixed context windows, fundamentally different from L1's length control approach.

5. VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks

URL: View paper

Brief Assessment

VAPO[61] focuses on value-based RL methods for reasoning tasks, addressing challenges like value model bias and reward sparsity. The original paper's Delethink training uses policy gradient methods with chunked reasoning traces to achieve linear compute scaling, which is a fundamentally different approach from VAPO's value-based framework.

6. Latent chain-of-thought for visual reasoning

URL: View paper

Brief Assessment

Latent Chain of Thought[67] focuses on amortized variational inference for visual reasoning in LVLMs, not on Markovian thinking paradigms or linear compute scaling for general RL frameworks. The candidate addresses visual chain-of-thought through probabilistic inference rather than the original's approach of chunked reasoning with fixed context windows.

7. When More is Less: Understanding Chain-of-Thought Length in LLMs

URL: View paper

Brief Assessment

Chain of Thought Length[62] focuses on optimal CoT length and simplicity bias in RL training, not on memory-efficient RL training procedures like Delethink's chunked reasoning approach.

8. Reinforcement Learning-Guided Chain-of-Draft for Token-Efficient Code Generation

URL: View paper

Brief Assessment

Chain of Draft[68] focuses on reinforcement learning for selecting optimal code generation solutions from multiple candidates, not on training models to reason in a Markovian manner with reduced compute. The candidate addresses code generation efficiency through multi-candidate selection, while the original contribution concerns training procedures for reasoning models with linear compute scaling.

9. Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning

URL: View paper

Prior Art Analysis

Thinkprune[63] demonstrates prior work on training models via reinforcement learning to reduce reasoning length while maintaining performance. Both papers use RL to train models that generate shorter reasoning traces with comparable accuracy. Thinkprune[63] explicitly trains models with length constraints during RL (clipping outputs beyond a token limit), achieving 40% faster generation with 70% less memory on similar benchmarks. The original paper's claim of being first to 'explicitly train models to reason in the Markovian manner' is refuted by Thinkprune[63]'s earlier demonstration of RL-based training for efficient reasoning with reduced compute requirements.

Evidence

Evidence 1 - **Rationale**: Both papers describe RL training methods that reduce reasoning length while maintaining performance. Thinkprune[63] explicitly uses RL with length constraints, directly comparable to the original paper's approach. - **Original**: we introduce delethink training, that trains these markovian thinkers through rl. we show it matches the performance of the standard long chain-of-thought in much less compute. - **Candidate**: we present think prune , a simple yet effective method for pruning the thinking length for long-thinking llms, which have been found to often produce inefficient and redundant thinking processes. [...] think prune offers a simple solution that continuously trains the long-thinking llms via reinforce...

Evidence 2 - **Rationale**: Thinkprune[63] describes an RL training procedure with length constraints that achieves efficiency gains, demonstrating prior work on the core contribution of RL-based efficient reasoning training. - **Original**: we introduce delethink training, that trains these markovian thinkers through rl. we show it matches the performance of the standard long chain-of-thought in much less compute. - **Candidate**: think prune adopts a similar rl scheme to the deepseek-r1 model (deepseek-ai, 2025) while reducing the generation length. specifically, we adopt the group relative policy optimization (grpo) algorithm shao et al. (2024). the reward function is almost the same as the deepseek-r1 framework, except tha...

10. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning

URL: View paper

Brief Assessment

Video RTS[66] focuses on video reasoning tasks using pure-RL training on video-question pairs, while the original paper addresses general reasoning with Markovian thinking paradigm for text-based chain-of-thought. The domains and technical approaches differ fundamentally.

Appendix: Text Similarity Detection

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

References

- [0] The Markovian Thinker View paper
- [1] Linformer: Self-Attention with Linear Complexity View paper
- [2] RWKV-X: A Linear Complexity Hybrid Language Model View paper
- [3] Leveraging mesh modularization to lower the computational cost of localized updates to regional 2D hydrodynamic model outputs View paper
- [4] FLASepformer: Efficient Speech Separation with Gated Focused Linear Attention Transformer View paper
- [5] Sequential-Parallel Duality in Prefix Scannable Models View paper
- [6] â□□-former: Infinite memory transformer View paper
- [7] JEDI-linear: Fast and Efficient Graph Neural Networks for Jet Tagging on FPGAs View paper
- [8] IMSSA: Deploying modern state-space models on memristive in-memory compute hardware View paper
- [9] Cottention: Linear Transformers With Cosine Attention View paper
- [10] Apriel-H1: Towards Efficient Enterprise Reasoning Models View paper
- [11] Quantifying the Impact of Energy System Model Resolution on Siting, Cost, Reliability, and Emissions View paper
- [12] Working-Memory-Correct Long-Horizon Expert-Retrieval TTT Dialogue View paper
- [13] GLULA: Linear attention-based model for efficient human activity recognition from wearable sensors View paper
- [14] Bottleneck spanning tree interdiction problem with fixed and linear costs View paper
- [15] Linearizing Transformer with Key-Value Memory View paper
- [16] A Solution for Large Scale Nonlinear Regression with High Rank and Degree at Constant Memory Complexity via Latent Tensor Reconstruction View paper
- [17] Predictive Monitoring against Pattern Regular Languages View paper
- [18] A linear complexity phasing method for thousands of genomes View paper
- [19] Real Time Recurrent Learning with Complex-Valued Trace Units View paper
- [20] LongVQ: Long Sequence Modeling with Vector Quantization on Structured Memory View paper
- [21] Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks View paper
- [22] Cost-Effective Attention Mechanisms for Low Resource Settings: Necessity&Sufficiency of Linear Transformations View paper
- [23] A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements View paper
- [24] Contraction Clustering (RASTER): A Very Fast Big Data Algorithm for Sequential and Parallel Density-Based Clustering in Linear Time, Constant Memory, and a Single Pass View paper
- [25] Linearly compressed pages: A low-complexity, low-latency main memory compression framework View paper
- [26] Finiteâ□□time output regulation by bounded linear timeâ□□varying controls with applications to the satellite formation flying View paper
- [27] Fully Online Grammar Compression in Constant Space View paper
- [28] Sub-linear memory: How to make performers slim View paper
- [29] Memformer: A memory-augmented transformer for sequence modeling View paper
- [30] A fast computational framework for the linear bond-based peridynamic model View paper
- [31] pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree View paper
- [32] Approximating Median Absolute Deviation with Bounded Error View paper
- [33] Computing Wasserstein-P Distance Between Images with Linear Cost View paper
- [34] High Memory Masked Convolutional Codes for PQC View paper
- [35] Memory-efficient Transformers via Top- Attention View paper
- [36] A Learning-Based Framework for Memory-Bounded Heuristic Search: First Results View paper
- [37] Semi-Parametric Deep Neural Networks in Linear Time and Memory View paper
- [38] Internal Shortest Absent Word Queries in Constant Time and Linear Space View paper
- [39] Efficient Sparseness-Enforcing Projections View paper
- [40] Childrenâ□□s Basic Etiquette Learning in Mobile Application View paper
- [41] Sequential Nonparametric Testing with the Law of the Iterated Logarithm View paper
- [42] Cinematic-L1 Video Stabilization with a Log-Homography Model View paper
- [43] Hardware-aligned Hierarchical Sparse Attention for Efficient Long-term Memory Access View paper
- [44] Diagonal Batching Unlocks Parallelism in Recurrent Memory Transformers for Long Contexts View paper
- [45] LongMamba: Enhancing Mamba's Long-Context Capabilities via Training-Free Receptive Field Enlargement View paper
- [46] LogSumExp: Efficient Approximate Logarithm Acceleration for Embedded Tractable Probabilistic Reasoning View paper
- [47] Fast linear-space computations of longest common subsequences View paper
- [48] Gated KalmaNet: A Fading Memory Layer Through Test-Time Ridge Regression View paper
- [49] xLSTM 7B: A Recurrent LLM for Fast and Efficient Inference View paper
- [50] Enhanced Data Race Prediction Through Modular Reasoning View paper
- [51] Titans: Learning to memorize at test time View paper
- [52] Scaling Reasoning without Attention View paper
- [53] JanusVLN: Decoupling Semantics and Spatiality with Dual Implicit Memory for Vision-Language Navigation View paper
- [54] Random-access infinite context length for transformers View paper
- [55] Lserve: Efficient long-sequence llm serving with unified sparse attention View paper
- [56] Artificial hippocampus networks for efficient long-context modeling View paper
- [57] DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads View paper
- [58] Lococo: Dropping in convolutions for long context compression View paper
- [59] Trellis: Learning to Compress Key-Value Memory in Attention Models View paper
- [60] MEDCnet: A Memory Efficient Approach for Processing Highâ□□Resolution Fundus Images for Diabetic Retinopathy Classification Using CNN View paper
- [61] VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks View paper
- [62] When More is Less: Understanding Chain-of-Thought Length in LLMs View paper
- [63] Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning View paper

- [64] L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning View paper
- [65] Training language models to reason efficiently View paper
- [66] Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning View paper
- [67] Latent chain-of-thought for visual reasoning View paper
- [68] Reinforcement Learning-Guided Chain-of-Draft for Token-Efficient Code Generation View paper
- [69] LLM-Guided Reinforcement Learning for Interactive Environments View paper
- [70] GRPO-MA: Multi-Answer Generation in GRPO for Stable and Efficient Chain-of-Thought Training View paper
- [71] Least-to-Most Prompting Enables Complex Reasoning in Large Language Models View paper
- [72] Art: Automatic multi-step reasoning and tool-use for large language models View paper
- [73] A comprehensive survey of prompt engineering techniques in large language models View paper
- [74] Lisa: Reasoning segmentation via large language model View paper
- [75] Recursive decomposition of logical thoughts: Framework for superior reasoning and knowledge propagation in large language models View paper
- [76] PAL: Program-aided Language Models View paper
- [77] A systematic survey of prompt engineering in large language models: Techniques and applications View paper
- [78] SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning View paper
- [79] Reasoning with large language models, a survey View paper
- [80] Decomposed Prompting: A Modular Approach for Solving Complex Tasks View paper