Novelty Assessment Report

Paper: The Serial Scaling Hypothesis
PDF URL: https://openreview.net/pdf?id=ObXB7KJn0B
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

While machine learning has advanced through massive parallelization, we identify a critical blind spot: some problems are fundamentally sequential. These "inherently serial" problems—from mathematical reasoning to physical simulations to sequential decision-making—require sequentially dependent computational steps that cannot be efficiently parallelized. We formalize this distinction in complexity theory, and demonstrate that current parallel-centric architectures face fundamental limitations on such tasks. Then, we show for first time that diffusion models despite their sequential nature are incapable of solving inherently serial problems. We argue that recognizing the serial nature of computation holds profound implications on machine learning, model design, and hardware development.

Core Task Landscape

This paper addresses: **Computational Limitations of Parallel versus Serial Computation in Machine Learning**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Theoretical Foundations and Complexity Analysis**
- **Parallelization Techniques for Sequential Tasks**
- **Distributed Training Frameworks and Optimization**
- **Hardware Acceleration and Implementation Strategies**
- **Application-Specific Parallel and Serial Computation**
- **Neural Architecture Design and Training Dynamics**

Complete Taxonomy Tree

- Computational Limitations of Parallel versus Serial Computation in Machine Learning Survey Taxonomy
- Theoretical Foundations and Complexity Analysis
  - Inherently Serial Problem Characterization ★ (2 papers)
  - [0] The Serial Scaling Hypothesis (Anon et al., 2026) View paper
  - [3] Chain of thought empowers transformers to solve inherently serial problems (Li, 2024) View paper
  - Convergence Bounds and Optimization Limits (2 papers)
  - [20] Theoretical limits of pipeline parallel optimization and application to distributed deep learning (Colin, 2019) View paper
  - [49] Graph oracle models, lower bounds, and gaps for parallel stochastic optimization (Woodworth, 2018) View paper
  - Expressiveness and Representational Capacity (2 papers)
  - [17] Promises and Pitfalls of Generative Masked Language Modeling: Theoretical Framework and Practical Guidelines (Li, 2024) View paper
  - [42] An exponentially-growing family of universal quantum circuits (Mohammad Kordzanganeh, 2022) View paper
- Parallelization Techniques for Sequential Tasks
  - Parallel Reasoning and Inference Strategies (4 papers)
  - [4] Parallel-r1: Towards parallel thinking via reinforcement learning (Zheng Tong, 2025) View paper
  - [5] Break the Sequential Dependency of LLM Inference Using Lookahead Decoding (Fu, 2024) View paper
  - [6] Learning adaptive parallel reasoning with language models (Pan Jiayi, 2025) View paper
  - [7] An LLM Compiler for Parallel Function Calling (Kim Se-Hoon, 2023) View paper
  - Recurrent and Sequential Model Parallelization (2 papers)
  - [39] Parallelizing linear recurrent neural nets over sequence length (Eric Martin, 2017) View paper
  - [47] Concurrent meta reinforcement learning (Parisotto, 2019) View paper
  - Concurrent and Asynchronous Execution Frameworks (3 papers)
  - [8] DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models (Yuzhi Zhang, 2020) View paper
  - [11] Pathways: Asynchronous distributed dataflow for ml (Barham, 2022) View paper
  - [24] Highly concurrent solutions to graph and image processing problems (James, 2022) View paper
- Distributed Training Frameworks and Optimization
  - Data and Model Parallelism Strategies (4 papers)
  - [15] Parallel machine learning algorithms (Saba Abdulbaqi Salman, 2023) View paper
  - [21] Demystifying parallel and distributed deep learning: An in-depth concurrency analysis (Ben-Nun, 2019) View paper
  - [28] Parallel Machine Learning Algorithm (Saba Abdulbaqi Salman, 2023) View paper
  - [29] Strads: A distributed framework for scheduled model parallel machine learning (Jin Kyu Kim, 2016) View paper

- Federated and Distributed Learning (4 papers)
  - [26] Black-Box Parallelization for Machine Learning. (Kamp, 2019) View paper
  - [27] Communication-Efficient Generalized Neuron Matching for Federated Learning (Sixu Hu, 2023) View paper
  - [36] Split Federated Learning Over Heterogeneous Edge Devices: Algorithm and Optimization (Yunrui Sun, 2024) View paper
  - [41] Multi-Model Wireless Federated Learning with Downlink Beamforming (Chong Zhang, 2023) View paper
  - Checkpointing and Fault Tolerance (1 papers)
  - [32] PCcheck: Persistent Concurrent Checkpointing for ML (Foteini Strati, 2025) View paper
- Hardware Acceleration and Implementation Strategies
  - Specialized Hardware Architectures (4 papers)
  - [2] Quantum parallel information exchange (QPIE) hybrid network with transfer learning (Ziqing Guo, 2025) View paper
  - [9] A 818â□4094 TOPS/W Capacitor-Reconfigured Analog CIM for Unified Acceleration of CNNs and Transformers (Kentaro Yoshioka, 2025) View paper
  - [30] Multi-Task Quantum Annealing for Rapid Multi-Class Classification (Jargalsaikhan Artag, 2024) View paper
  - [44] Efficient Mux-Based Multiplier for MAC Unit (Huruy Tesfai, 2023) View paper
  - GPU and Accelerator Optimization (4 papers)
  - [14] Machine Learning-Based Kernel Selector for SpMV Optimization in Graph Analysis (Guoqing Xiao, 2024) View paper
  - [40] vbeam: a Fast and Differentiable Beamformer for Optimizing Ultrasound Imaging (Magnus Dalen KvalevÃ¥g, 2023) View paper
  - [46] Performance and power prediction for concurrent execution on gpus (Diksha Moolchandani, 2022) View paper
  - [50] DTSpMV: An Adaptive SpMV Framework for Graph Analysis on GPUs (Guoqing Xiao, 2022) View paper
  - CPU-Based Parallel Computing (5 papers)
  - [19] Leveraging Parallel Computing for Enhanced Stock Movement Forecasting Using Machine Learning (Shahd Aleissa, 2024) View paper
  - [31] Matrix Multiplication Analysis on Sequential and Parallel Computation using CUDA (Robertus Hudi, 2022) View paper
  - [35] Unsupervised-based distributed machine learning for efficient data clustering and prediction (Fathi Amsaad, 2023) View paper
  - [43] A parallel Fortran framework for neural networks and deep learning (Milan Curcic, 2019) View paper
  - [45] Parallel Computing Techniques for Accelerating Machine Learning Algorithms on Big Data (Rahul Mishra, 2023) View paper
  - Bit-Level and Approximate Computing (2 papers)
  - [13] Enhancing Computation-Efficiency of Deep Neural Network Processing on Edge Devices through Serial/Parallel Systolic Computing (Iraj Moghaddasi, 2024) View paper
  - [48] Optimization of stochastic computing based deep learning systems with parallel finite state machine implementation (Jinjie Liu, 2020) View paper
- Application-Specific Parallel and Serial Computation
  - Sequential Pattern Mining and Stream Processing (2 papers)
  - [1] A survey of parallel sequential pattern mining (Wensheng Gan, 2019) View paper
  - [18] Cuckoo Heavy Keeper and the balancing act of maintaining heavy hitters in stream processing (Vinh Quang Ngo, 2024) View paper
  - Optimization and Search Problems (4 papers)
  - [10] Nonlinear system identification using a semi concurrent sequential niching framework (Sun Yanxia, 2023) View paper
  - [12] A deep reinforcement learning-guided multimodal multi-objective evolutionary algorithm with a serial-parallel mechanism (Ying Huang, 2026) View paper
  - [33] Optimizing Well Trajectory Using Sequential, Hybrid Sequential, and Fully Concurrent Method Utilizing Machine Learning: A Case Study of a Tight Limestone â□¦ (R Rizkiaputra, 2023) View paper
  - [34] Optimizing Well Trajectory Using Sequential, Hybrid Sequential, and Fully Concurrent Method Utilizing Machine Learning: A Case Study of a Tight Limestone Reservoir (Ricko Rizkiaputra, 2023) View paper
  - Activity Recognition and Sequential Modeling (1 papers)
  - [22] RecurrentHAR: A novel transfer learning-based deep learning model for sequential, complex, concurrent, interleaved, and heterogeneous type human activity â□¦ (P Kumar, 2023) View paper
  - Scientific Computing and Process Modeling (1 papers)
  - [25] A hybrid scienceâ□guided machine learning approach for modeling chemical processes: A review (Niket Sharma, 2022) View paper
  - Code Generation and Programming Assistance (1 papers)
  - [16] ChatGPT for Programming Numerical Methods (Kashefi, 2023) View paper
  - Security and Anomaly Detection (1 papers)
  - [37] Phishing URLs Detection Using Sequential and Parallel ML Techniques: Comparative Analysis (Naya Nagy, 2023) View paper
- Neural Architecture Design and Training Dynamics
  - Gradient-Free and Alternative Training Methods (1 papers)
  - [38] Neuro-distributed cognitive adaptive optimization for training neural networks in a parallel and asynchronous manner (Panagiotis Michailidis, 2023) View paper
  - Parallel Support Vector Machines (1 papers)
  - [23] Parallel computing of support vector machines: a survey (Tavara, 2019) View paper

Narrative

Core task: understanding the computational limitations of parallel versus serial computation in machine learning. The field structure reflects a multi-layered investigation spanning theoretical foundations, practical implementation strategies, and application domains. At the highest level, the taxonomy distinguishes between theoretical complexity analysis that characterizes which problems resist parallelization, techniques for parallelizing inherently sequential tasks, distributed training frameworks that coordinate large-scale optimization, hardware acceleration strategies that exploit specialized architectures, application-specific studies across diverse domains, and neural architecture design considerations that influence training dynamics. Works such as Pathways[11] and Demystifying Distributed Deep Learning[21] illustrate how distributed frameworks address scalability, while hardware-focused studies like Capacitor-Reconfigured Analog CIM[9] and Matrix Multiplication CUDA[31] explore acceleration at the silicon and system levels. Meanwhile, theoretical branches examine fundamental limits, asking which computational patterns can be efficiently parallelized and which remain bottlenecked by serial dependencies.

A particularly active line of inquiry centers on reasoning and sequential decision-making, where the tension between parallel and serial computation becomes acute. Chain of Thought[3] and related works such as Lookahead Decoding[5] and Parallel Thinking Reinforcement[4] explore whether multi-step reasoning can be accelerated through parallelism or whether it fundamentally requires

sequential token generation. The Serial Scaling Hypothesis[0] sits squarely within this debate, examining the inherent serial nature of certain problem classes and proposing that some tasks may resist parallelization due to intrinsic computational structure rather than algorithmic limitations. This contrasts with Chain of Thought[3], which investigates how sequential reasoning emerges in language models, and with Lookahead Decoding[5], which attempts to mitigate serial bottlenecks through speculative execution. Together, these works highlight an open question: to what extent can we overcome serial dependencies through clever parallelization, and where do fundamental complexity barriers force us to accept sequential computation as unavoidable?

Related Works in Same Category

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

1. Chain of thought empowers transformers to solve inherently serial problems

Authors: Li, Zhiyuan, Zhiyuan Li, Liu Hong, Hong Liu, et al. (10 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Instructing the model to generate a sequence of intermediate steps, a.k.a., a chain of thought (CoT), is a highly effective method to improve the accuracy of large language models (LLMs) on arithmetics and symbolic reasoning tasks. However, the mechanism behind CoT remains unclear. This work provides a theoretical understanding of the power of CoT for decoder-only transformers through the lens of expressiveness. Conceptually, CoT empowers the model with the ability to perform inherently serial c...

Relationship Analysis

Both papers belong to the Inherently Serial Problem Characterization category, examining problems requiring sequential dependencies that resist parallelization through complexity-theoretic foundations. They overlap in analyzing how constant-depth transformers are limited to TC^0 without chain-of-thought (CoT) and demonstrating that CoT enables solving inherently serial problems by adding sequential computation steps. The key difference is that the original paper (Serial Scaling Hypothesis) provides a broader survey across multiple domains (cellular automata, physical systems, MDPs, diffusion models) and proposes implications for hardware and model design, while the candidate paper focuses specifically on the theoretical expressiveness analysis of transformers with CoT, proving tighter bounds (AC^0 vs TC^0) and providing targeted empirical validation on circuit problems.

Contributions Analysis

Overall novelty summary. The paper formalizes the distinction between inherently serial and parallelizable problems in machine learning, proposing the Serial Scaling Hypothesis and analyzing diffusion models' limitations on sequential tasks. It resides in the 'Inherently Serial Problem Characterization' leaf, which contains only two papers total within the 'Theoretical Foundations and Complexity Analysis' branch. This sparse positioning suggests the paper addresses a relatively underexplored theoretical direction within a 50-paper taxonomy spanning 21 leaf nodes. The leaf's scope explicitly focuses on formalizing problems requiring sequential dependencies that resist parallelization, distinguishing it from implementation-focused branches.

The taxonomy reveals neighboring work in 'Convergence Bounds and Optimization Limits' and 'Expressiveness and Representational Capacity' within the same theoretical branch, alongside a substantial 'Parallelization Techniques for Sequential Tasks' branch containing methods like Lookahead Decoding and Parallel Thinking Reinforcement. The paper's theoretical stance contrasts with these parallelization attempts: while neighboring leaves examine optimization complexity or architectural expressiveness, this work argues certain problems fundamentally resist the parallelization strategies explored in adjacent branches. The taxonomy's structure highlights this tension between theoretical impossibility results and practical parallelization efforts.

Among 30 candidates examined across three contributions, none yielded clear refutations. The Serial Scaling Hypothesis examined 10 candidates with zero refutable matches, as did the complexity-theoretic characterization of diffusion models and the theoretical framework for inherently serial problems. This absence of overlapping prior work within the limited search scope suggests these specific formalizations—particularly the SSH and the diffusion model analysis—may represent novel theoretical angles. However, the search scale is modest: 30 candidates from semantic search cannot comprehensively cover all relevant complexity theory or diffusion model literature.

The analysis indicates apparent novelty within the examined scope, particularly for the SSH and diffusion model characterization, though the 30-candidate search cannot rule out relevant prior work in broader complexity theory or generative modeling literature. The sparse taxonomy leaf and zero refutations across contributions suggest the paper occupies a relatively unexplored theoretical niche, but definitive novelty assessment would require examining complexity theory venues and diffusion model theory beyond the top-30 semantic matches analyzed here.

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

Contribution 1: Serial Scaling Hypothesis (SSH)

Description: The authors propose the Serial Scaling Hypothesis, which states that for many important ML problems such as reasoning, decision making, and modeling dynamic systems, increasing parallel computation alone is insufficient and progress requires scaling the amount of serial computation. This hypothesis is grounded in complexity theory and connects theoretical limitations to practical ML challenges.

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. Multi-model machine learning inference serving with gpu spatial partitioning

URL: View paper

Brief Assessment

GPU Spatial Partitioning[56] focuses on GPU resource partitioning for multi-model ML inference serving with SLO constraints, not on the fundamental distinction between serial versus parallel computation scaling in ML problems.

---

2. ASPD: Unlocking Adaptive Serial-Parallel Decoding by Exploring Intrinsic Parallelism in LLMs

URL: View paper

Brief Assessment

ASPD[54] focuses on optimizing LLM inference through adaptive serial-parallel decoding mechanisms, not on theoretical complexity analysis of serial versus parallel computation scaling. The candidate addresses engineering optimization of existing autoregressive models rather than fundamental computational complexity theory.

---

3. ML Inference Scheduling with Predictable Latency

URL: View paper

Brief Assessment

ML Inference Scheduling[58] focuses on inference serving systems and latency optimization for ML workloads, not on the theoretical distinction between serial versus parallel computation scaling in learning algorithms or the computational complexity of ML problems.

---

4. S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models

URL: View paper

Brief Assessment

S-GRPO[51] focuses on optimizing reasoning efficiency through early exit mechanisms in chain-of-thought generation, not on the fundamental distinction between serial versus parallel computation scaling in ML architectures or the theoretical complexity framework that SSH establishes.

---

5. Learning adaptive parallel reasoning with language models

URL: View paper

Brief Assessment

Adaptive Parallel Reasoning[6] focuses on optimizing inference-time computation through adaptive parallelization in language models for reasoning tasks, not on the fundamental theoretical distinction between serial and parallel computation scaling in ML.

---

6. Parallel Test-Time Scaling for Latent Reasoning Models

URL: View paper

Brief Assessment

Parallel Test-Time Scaling[57] focuses on parallel inference strategies for latent reasoning models in continuous spaces, not on the fundamental serial vs. parallel computation dichotomy in ML architectures that SSH addresses.

---

7. Don't Overthink it. Preferring Shorter Thinking Chains for Improved LLM Reasoning

URL: View paper

Brief Assessment

Shorter Thinking Chains[52] focuses on optimizing test-time compute efficiency by using shorter reasoning chains in LLM inference, not on the fundamental distinction between serial versus parallel computation scaling in ML architectures and problem complexity.

---

8. SSR: Speculative Parallel Scaling Reasoning in Test-time

URL: View paper

Brief Assessment

SSR[55] focuses on test-time inference acceleration through speculative decoding and parallel path selection for mathematical reasoning, not on the fundamental distinction between serial versus parallel computation scaling in ML architectures or the theoretical complexity framework proposed in SSH.

---

9. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities?

URL: View paper

Brief Assessment

Test-Time Scaling Revisited[53] focuses on empirical analysis of chain-of-thought length and self-revision in specific o1-like models (QWQ, DeepSeek-R1, LIMO), comparing sequential versus parallel scaling strategies. It does not address the theoretical complexity-theoretic framework distinguishing serial versus parallel computation that forms the core of SSH.

---

10. Break the Sequential Dependency of LLM Inference Using Lookahead Decoding

URL: View paper

Brief Assessment

Lookahead Decoding[5] focuses on parallelizing LLM inference through a specific decoding algorithm, not on the fundamental distinction between serial and parallel computation scaling in ML problems like reasoning and decision-making.

Contribution 2: Complexity-theoretic characterization of diffusion models

Description: The authors prove that diffusion models with a TC0 backbone remain in TC0 even with infinitely many sampling steps, demonstrating that despite their stepwise structure, diffusion models are incapable of solving inherently serial problems. This is the first formal characterization showing diffusion models do not provide scalable serial computation.

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. Denoising diffusion restoration models

URL: View paper

Brief Assessment

Denoising Diffusion Restoration[60] focuses on using diffusion models for image restoration tasks (super-resolution, deblurring, inpainting) rather than analyzing their computational complexity or serial computation capabilities. The paper does not address complexity-theoretic characterizations of diffusion models' sampling steps.

---

2. Diffir: Efficient diffusion model for image restoration

URL: View paper

Brief Assessment

Diffir[63] focuses on efficient diffusion models for image restoration tasks, not on complexity-theoretic characterization of diffusion models' computational capabilities or their inability to solve inherently serial problems.

---

3. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps

URL: View paper

Brief Assessment

DPM-Solver[64] focuses on developing fast ODE solvers for diffusion model sampling efficiency, not on complexity-theoretic characterization of their computational capabilities or serial computation limits.

---

4. Denoising task routing for diffusion models

URL: View paper

Brief Assessment

Denoising Task Routing[66] focuses on architectural improvements for diffusion models through task routing mechanisms, not on complexity-theoretic characterization of their computational capabilities or serial computation limits.

---

5. Noise estimation for generative diffusion models

URL: View paper

Brief Assessment

Noise Estimation Diffusion[61] focuses on adjusting noise schedules during inference to improve sample quality with fewer steps, not on complexity-theoretic characterization of diffusion models' computational capabilities or their inability to solve inherently serial problems.

---

6. Practical and Asymptotically Exact Conditional Sampling in Diffusion Models

URL: View paper

Brief Assessment

Conditional Sampling Diffusion[59] focuses on conditional sampling algorithms for diffusion models using sequential Monte Carlo methods, not on complexity-theoretic characterization of their computational capabilities or serial computation limits.

---

7. Invertible diffusion models for compressed sensing

URL: View paper

Brief Assessment

Invertible Diffusion Models[62] focuses on compressed sensing applications and memory-efficient training of diffusion models, not on complexity-theoretic characterization of their computational capabilities or serial computation limits.

---

8. Accelerating diffusion models via early stop of the diffusion process

URL: View paper

Brief Assessment

Early Stop Diffusion[68] focuses on accelerating diffusion models by stopping the diffusion process early and combining with pre-trained generative models (GANs/VAEs). It does not address complexity-theoretic characterization of diffusion models' computational capabilities or their relationship to TC0 circuits.

---

9. Sequential Posterior Sampling with Diffusion Models

URL: View paper

Brief Assessment

Sequential Posterior Sampling[67] focuses on accelerating diffusion inference for sequential ultrasound imaging through transition dynamics modeling, not on computational complexity characterization of diffusion models' iterative sampling steps.

---

10. Denoising diffusion implicit models

URL: View paper

Brief Assessment

Denoising Diffusion Implicit[65] focuses on accelerating sampling in diffusion models through non-Markovian processes, not on complexity-theoretic characterization of their computational capabilities or serial computation limits.

Contribution 3: Theoretical framework for inherently serial problems in ML

Description: The authors formalize the distinction between parallel and inherently serial problems using complexity theory (specifically the TC class), identify real-world ML problems that are inherently serial (cellular automata, many-body mechanics, sequential decision-making, mathematical QA), and prove that modern architectures like MLPs, Transformers, SSMs, and diffusion models with TC0 backbones cannot solve general instances of these problems.

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. Multigraph learning for parallelism discovery in sequential programs

URL: View paper

Brief Assessment

Multigraph Parallelism Discovery[74] focuses on discovering parallelizable code segments in sequential programs using neural networks, not on complexity-theoretic limitations of parallel architectures on inherently serial problems.

---

2. 60 years of mastering concurrent computing through sequential thinking

URL: View paper

Brief Assessment

Mastering Concurrent Computing[72] focuses on concurrent/distributed computing systems and synchronization primitives (mutual exclusion, consensus, state machine replication), not on ML architectures or complexity-theoretic limitations of neural networks on serial problems.

---

3. Concurrent Scanning through Adaptive Task Distribution for Simultaneous Lexing on Multi-Core Platforms

URL: View paper

Brief Assessment

Concurrent Scanning Adaptive[75] focuses on parallel lexical analysis in compiler design for multi-core systems, not on complexity theory limitations of ML architectures on inherently serial problems. The domains and research questions are entirely distinct.

---

4. A complexity theory of efficient parallel algorithms

URL: View paper

Brief Assessment

Complexity Theory Parallel[69] focuses on general parallel algorithm complexity theory and P-completeness, not specifically on ML architectures (Transformers, SSMs, diffusion models) or ML problems (cellular automata, sequential decision-making, mathematical QA) as formalized in the original paper.

---

5. High-level synthesis of parallel specifications coupling static and dynamic controllers

URL: View paper

Brief Assessment

High-Level Synthesis Parallel[76] focuses on hardware synthesis for parallel task execution in FPGA/ASIC design, not on complexity-theoretic analysis of ML architectures' limitations on serial problems. The candidate discusses FSM controllers being 'inherently serial' in a hardware implementation context, which is unrelated to the original paper's TC complexity class framework for ML models.

6. Reducing concurrent analysis under a context bound to sequential analysis

URL: View paper

Brief Assessment

Concurrent to Sequential Analysis[77] addresses reducing concurrent program analysis to sequential analysis under context bounds in formal verification, not complexity-theoretic limitations of ML architectures on inherently serial problems.

7. Bounded model checking of concurrent programs

URL: View paper

Brief Assessment

Bounded Model Checking[78] focuses on verifying bounded sequential C programs with limited loop executions and recursion depth. This is a program verification technique, not a theoretical framework for analyzing ML architectures' complexity-theoretic limitations on inherently serial problems.

8. Toward parallel intelligence: An interdisciplinary solution for complex systems

URL: View paper

Brief Assessment

Parallel Intelligence[71] discusses parallel systems methods for complex systems but does not address complexity theory classifications (TC classes) or formal limitations of parallel architectures on inherently serial computational problems.

9. Attractor dynamics and parallelism in a connectionist sequential machine

URL: View paper

Brief Assessment

Attractor Dynamics Parallelism[70] focuses on connectionist sequential machines for speech production with parallel distributed representations, not on complexity-theoretic formalization of serial vs. parallel problems in ML architectures.

10. The theory and practice of concurrency

URL: View paper

Brief Assessment

Theory Practice Concurrency[73] is a textbook on concurrency theory and CSP (Communicating Sequential Processes), not a machine learning paper. It does not address ML architectures, complexity classes in the context of neural networks, or the specific problems (cellular automata prediction, many-body mechanics, sequential decision-making, mathematical QA) identified in the original paper.

Appendix: Text Similarity Detection

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

References

- [0] The Serial Scaling Hypothesis View paper
- [1] A survey of parallel sequential pattern mining View paper
- [2] Quantum parallel information exchange (QPIE) hybrid network with transfer learning View paper
- [3] Chain of thought empowers transformers to solve inherently serial problems View paper
- [4] Parallel-r1: Towards parallel thinking via reinforcement learning View paper
- [5] Break the Sequential Dependency of LLM Inference Using Lookahead Decoding View paper
- [6] Learning adaptive parallel reasoning with language models View paper
- [7] An LLM Compiler for Parallel Function Calling View paper
- [8] DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models View paper
- [9] A 818â□4094 TOPS/W Capacitor-Reconfigured Analog CIM for Unified Acceleration of CNNs and Transformers View paper
- [10] Nonlinear system identification using a semi concurrent sequential niching framework View paper
- [11] Pathways: Asynchronous distributed dataflow for ml View paper
- [12] A deep reinforcement learning-guided multimodal multi-objective evolutionary algorithm with a serial-parallel mechanism View paper
- [13] Enhancing Computation-Efficiency of Deep Neural Network Processing on Edge Devices through Serial/Parallel Systolic Computing View paper
- [14] Machine Learning-Based Kernel Selector for SpMV Optimization in Graph Analysis View paper
- [15] Parallel machine learning algorithms View paper
- [16] ChatGPT for Programming Numerical Methods View paper
- [17] Promises and Pitfalls of Generative Masked Language Modeling: Theoretical Framework and Practical Guidelines View paper
- [18] Cuckoo Heavy Keeper and the balancing act of maintaining heavy hitters in stream processing View paper
- [19] Leveraging Parallel Computing for Enhanced Stock Movement Forecasting Using Machine Learning View paper
- [20] Theoretical limits of pipeline parallel optimization and application to distributed deep learning View paper
- [21] Demystifying parallel and distributed deep learning: An in-depth concurrency analysis View paper
- [22] RecurrentHAR: A novel transfer learning-based deep learning model for sequential, complex, concurrent, interleaved, and heterogeneous type human activity â□¦ View paper
- [23] Parallel computing of support vector machines: a survey View paper
- [24] Highly concurrent solutions to graph and image processing problems View paper
- [25] A hybrid scienceâ□guided machine learning approach for modeling chemical processes: A review View paper
- [26] Black-Box Parallelization for Machine Learning. View paper
- [27] Communication-Efficient Generalized Neuron Matching for Federated Learning View paper
- [28] Parallel Machine Learning Algorithm View paper
- [29] Strads: A distributed framework for scheduled model parallel machine learning View paper
- [30] Multi-Task Quantum Annealing for Rapid Multi-Class Classification View paper
- [31] Matrix Multiplication Analysis on Sequential and Parallel Computation using CUDA View paper
- [32] PCcheck: Persistent Concurrent Checkpointing for ML View paper

- [33] Optimizing Well Trajectory Using Sequential, Hybrid Sequential, and Fully Concurrent Method Utilizing Machine Learning: A Case Study of a Tight Limestone â¦ View paper
- [34] Optimizing Well Trajectory Using Sequential, Hybrid Sequential, and Fully Concurrent Method Utilizing Machine Learning: A Case Study of a Tight Limestone Reservoir View paper
- [35] Unsupervised-based distributed machine learning for efficient data clustering and prediction View paper
- [36] Split Federated Learning Over Heterogeneous Edge Devices: Algorithm and Optimization View paper
- [37] Phishing URLs Detection Using Sequential and Parallel ML Techniques: Comparative Analysis View paper
- [38] Neuro-distributed cognitive adaptive optimization for training neural networks in a parallel and asynchronous manner View paper
- [39] Parallelizing linear recurrent neural nets over sequence length View paper
- [40] vbeam: a Fast and Differentiable Beamformer for Optimizing Ultrasound Imaging View paper
- [41] Multi-Model Wireless Federated Learning with Downlink Beamforming View paper
- [42] An exponentially-growing family of universal quantum circuits View paper
- [43] A parallel Fortran framework for neural networks and deep learning View paper
- [44] Efficient Mux-Based Multiplier for MAC Unit View paper
- [45] Parallel Computing Techniques for Accelerating Machine Learning Algorithms on Big Data View paper
- [46] Performance and power prediction for concurrent execution on gpus View paper
- [47] Concurrent meta reinforcement learning View paper
- [48] Optimization of stochastic computing based deep learning systems with parallel finite state machine implementation View paper
- [49] Graph oracle models, lower bounds, and gaps for parallel stochastic optimization View paper
- [50] DTSpMV: An Adaptive SpMV Framework for Graph Analysis on GPUs View paper
- [51] S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models View paper
- [52] Don't Overthink it. Preferring Shorter Thinking Chains for Improved LLM Reasoning View paper
- [53] Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? View paper
- [54] ASPD: Unlocking Adaptive Serial-Parallel Decoding by Exploring Intrinsic Parallelism in LLMs View paper
- [55] SSR: Speculative Parallel Scaling Reasoning in Test-time View paper
- [56] Multi-model machine learning inference serving with gpu spatial partitioning View paper
- [57] Parallel Test-Time Scaling for Latent Reasoning Models View paper
- [58] ML Inference Scheduling with Predictable Latency View paper
- [59] Practical and Asymptotically Exact Conditional Sampling in Diffusion Models View paper
- [60] Denoising diffusion restoration models View paper
- [61] Noise estimation for generative diffusion models View paper
- [62] Invertible diffusion models for compressed sensing View paper
- [63] Diffir: Efficient diffusion model for image restoration View paper
- [64] Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps View paper
- [65] Denoising diffusion implicit models View paper
- [66] Denoising task routing for diffusion models View paper
- [67] Sequential Posterior Sampling with Diffusion Models View paper
- [68] Accelerating diffusion models via early stop of the diffusion process View paper
- [69] A complexity theory of efficient parallel algorithms View paper
- [70] Attractor dynamics and parallelism in a connectionist sequential machine View paper
- [71] Toward parallel intelligence: An interdisciplinary solution for complex systems View paper
- [72] 60 years of mastering concurrent computing through sequential thinking View paper
- [73] The theory and practice of concurrency View paper
- [74] Multigraph learning for parallelism discovery in sequential programs View paper
- [75] Concurrent Scanning through Adaptive Task Distribution for Simultaneous Lexing on Multi-Core Platforms View paper
- [76] High-level synthesis of parallel specifications coupling static and dynamic controllers View paper
- [77] Reducing concurrent analysis under a context bound to sequential analysis View paper
- [78] Bounded model checking of concurrent programs View paper