Novelty Assessment Report

Paper: WSM: Decay-Free Learning Rate Schedule via Checkpoint Merging for LLM Pre-training

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

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Recent advances in learning rate~(LR) scheduling have demonstrated the effectiveness of decay-free approaches that eliminate the traditional decay phase while maintaining competitive performance. Model merging techniques have emerged as particularly promising solutions in this domain. We present Warmup-Stable and Merge (WSM), a general framework that establishes a formal connection between learning rate decay and model merging. WSM provides a unified theoretical foundation for emulating various decay strategies—including cosine decay, linear decay and inverse square root decay—as principled model averaging schemes, while remaining fully compatible with diverse optimization methods. Through extensive experiments, we identify merge duration—the training window for checkpoint aggregation—as the most critical factor influencing model performance, surpassing the importance of both checkpoint interval and merge quantity. Our framework consistently outperforms the widely-adopted Warmup-Stable-Decay (WSD) approach across multiple benchmarks, achieving significant improvements of +3.5\% on MATH, +2.9\% on HumanEval, and +5.5\% on MMLU-Pro. The performance advantages extend to supervised fine-tuning scenarios, highlighting WSM's potential for long-term model refinement.

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: **Learning Rate Scheduling for Large Language Model Pre-training**

A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

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

- **Learning Rate Schedule Design and Optimization**
- **Continual and Incremental Pre-training**
- **Training Efficiency and Acceleration Methods**
- **Specialized Learning Rate Tuning Approaches**
- **Empirical Analysis and Comparative Studies**
- **Infrastructure and Systems Perspectives**
- **Model Compression and Efficiency**
- **Specialized Applications and Extensions**

Complete Taxonomy Tree

- Learning Rate Scheduling for Large Language Model Pre-training Survey Taxonomy
- Learning Rate Schedule Design and Optimization
  - Novel Schedule Architectures ★ (4 papers)
  - [0] WSM: Decay-Free Learning Rate Schedule via Checkpoint Merging for LLM Pre-training (Anon et al., 2026) View paper
  - [8] Straight to zero: Why linearly decaying the learning rate to zero works best for LLMs (Bergsma, 2025) View paper
  - [10] Beyond Cosine Decay: On the effectiveness of Infinite Learning Rate Schedule for Continual Pre-training (Singh, 2025) View paper
  - [23] Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective (Wen, 2024) View paper
  - Theoretical Foundations and Scaling Laws (4 papers)
  - [12] The Surprising Agreement Between Convex Optimization Theory and Learning-Rate Scheduling for Large Model Training (Schaipp, 2025) View paper
  - [30] Scaling law with learning rate annealing (Lu, 2024) View paper
  - [43] Optimization Hyper-parameter Laws for Large Language Models (Xie Xingyu, 2024) View paper
  - [45] Scaling and Transferability of Annealing Strategies in Large Language Model Training (Siqi Wang, 2025) View paper
  - Adaptive and Dynamic Schedule Optimization (3 papers)
  - [13] AdaLRS: Loss-Guided Adaptive Learning Rate Search for Efficient Foundation Model Pretraining (Dong Hongyuan, 2025) View paper
  - [28] Learning to schedule learning rate with graph neural networks (Y Xiong, 2022) View paper
  - [36] Learning rate decay scheduler using the loss operator as generalization (Wenbo An, 2024) View paper
  - Optimal Schedule Analysis and Refinement (2 papers)
  - [37] Optimal linear decay learning rate schedules and further refinements (Defazio, 2023) View paper
  - [46] When, why and how much? adaptive learning rate scheduling by refinement (Aaron Defazio, 2023) View paper
- Continual and Incremental Pre-training
  - Learning Rate Re-warming and Re-decay Strategies (2 papers)
  - [7] Continual pre-training of large language models: How to (re) warm your model? (Gupta, 2023) View paper
  - [9] Simple and Scalable Strategies to Continually Pre-train Large Language Models (Ibrahim, 2024) View paper
  - Version Update Training Paradigms (2 papers)

    - [2] A learning rate path switching training paradigm for version updates of large language models (Huang Jian-Heng, 2024) View paper
    - [3] Reuse, don't retrain: A recipe for continued pretraining of language models (Parmar, 2024) View paper
    - Learning Dynamics in Continual Settings (1 papers)
    - [11] Learning Dynamics in Continual Pre-Training for Large Language Models (Wang Xingjin, 2025) View paper
- Training Efficiency and Acceleration Methods
    - Checkpoint Averaging and Model Merging (2 papers)
    - [4] Hop, skip, jump to convergence: Dynamics of learning rate transitions for improved training of large language models (Shreyas Subramanian, 2024) View paper
    - [5] Early Weight Averaging meets High Learning Rates for LLM Pre-training (Sanyal, 2023) View paper
    - Curriculum Learning and Data Scheduling (3 papers)
    - [6] Length-based curriculum learning for efficient pre-training of language models (Koichi Nagatsuka, 2023) View paper
    - [24] Dataset decomposition: Faster llm training with variable sequence length curriculum (Jen-Hao Chang, 2024) View paper
    - [44] How Learning Rate Decay Wastes Your Best Data in Curriculum-Based LLM Pretraining (Kairong Luo, 2025) View paper
    - Batch Size and Compute Scaling (2 papers)
    - [15] How Does Critical Batch Size Scale in Pre-training? (Zhang Hanlin, 2024) View paper
    - [42] Surge phenomenon in optimal learning rate and batch size scaling (Bin Cui, 2024) View paper
- Specialized Learning Rate Tuning Approaches
    - Fine-tuning and Transfer Learning (3 papers)
    - [1] Rethinking learning rate tuning in the era of large language models (Hongpeng Jin, 2023) View paper
    - [21] Large Language Model Empowered Recommendation Meets All-domain Continual Pre-Training (Ma, 2025) View paper
    - [29] Layer-Wise Learning Rate Optimization for Task-Dependent Fine-Tuning of Pre-Trained Models: An Evolutionary Approach (Chenyang Bu, 2024) View paper
    - Layer-wise and Component-specific Scheduling (3 papers)
    - [16] The Sharpness Disparity Principle in Transformers for Accelerating Language Model Pre-Training (Wang Jin-bo, 2025) View paper
    - [19] No more adam: Learning rate scaling at initialization is all you need (Xu Minghao, 2024) View paper
    - [50] Optimal Embedding Learning Rate in LLMs: The Effect of Vocabulary Size (Hayou, 2025) View paper
- Empirical Analysis and Comparative Studies
    - Compute-optimal Training Analysis (2 papers)
    - [17] An empirical analysis of compute-optimal large language model training (J Hoffmann, 2022) View paper
    - [18] Scaling laws and compute-optimal training beyond fixed training durations (Elie Bakouch, 2024) View paper
    - Training Stability and Gradient Dynamics (3 papers)
    - [22] Neural Thermodynamic Laws for Large Language Model Training (Liu, 2025) View paper
    - [25] Why Gradients Rapidly Increase Near the End of Training (Defazio, 2025) View paper
    - [34] Methods of improving LLM training stability (Rybakov Oleg, 2024) View paper
    - Cross-domain and Multi-modal Applications (2 papers)
    - [33] Scaling Laws for Robust Comparison of Open Foundation Language-Vision Models and Datasets (Nezhurina, 2025) View paper
    - [48] InLegalLLaMA: Indian Legal Knowledge Enhanced Large Language Model (S Ghosh, 2024) View paper
- Infrastructure and Systems Perspectives
    - Distributed and Parallel Training (2 papers)
    - [20] Evaluation of pre-training large language models on leadership-class supercomputers: J. Yin et al. (J Yin, 2023) View paper
    - [26] DreamDDP: Accelerating Data Parallel Distributed LLM Training with Layer-wise Scheduled Partial Synchronization (Tang, 2025) View paper
    - Resource Scheduling and Job Management (4 papers)
    - [14] A Survey on the Scheduling of DL and LLM Training Jobs in GPU Clusters (Tianhao Fu, 2025) View paper
    - [38] Coflow Scheduling for LLM Training (Xinchen Wan, 2025) View paper
    - [39] Optimizing the Utilization of Large Language Models via Schedule Optimization: An Exploratory Study (Yueyue Liu, 2024) View paper
    - [49] Spaced Scheduling for Large Language Model Training (Amine El Hattami, 2025) View paper
    - Hardware-specific Implementations (1 papers)
    - [27] Hlat: High-quality large language model pre-trained on aws trainium (Haozheng Fan, 2024) View paper
- Model Compression and Efficiency
    - Pruning and Structured Compression (2 papers)
    - [40] IDEA Prune: An Integrated Enlarge-and-Prune Pipeline in Generative Language Model Pretraining (Li, 2025) View paper
    - [47] Structured Representation Compression for Large Language Models through Hierarchical Tensor Partitioning (Penelope Tifantome, 2025) View paper
    - Low-precision and Quantized Training (1 papers)
    - [41] Bitnet: 1-bit pre-training for large language models (H Wang, 2025) View paper
- Specialized Applications and Extensions
    - Reasoning and Reflection-based Training (1 papers)
    - [31] CyclicReflex: Improving Large Reasoning Models via Cyclical Reflection Token Scheduling (Zhang Yi-hua, 2025) View paper
    - Multi-level and Hierarchical Representations (1 papers)
    - [32] Enhancing large language models with stochastic multi-level embedding fusion: An experimental approach on open-source llm (Hayden Raines, 2024) View paper
    - Survey and Review Studies (1 papers)
    - [35] Mid-Training of Large Language Models: A Survey (Shi Yuxin, 2025) View paper

Narrative

Core task: learning rate scheduling for large language model pre-training. The field has evolved into a rich landscape of interconnected research directions. At the highest level, the taxonomy distinguishes between foundational schedule design (exploring novel architectures and optimization principles), continual and incremental pre-training (addressing how to adapt schedules when extending or updating models), training efficiency and acceleration methods (focusing on computational trade-offs and system-level speedups), and specialized tuning approaches (including layer-wise or adaptive strategies). Additional branches cover empirical comparisons, infrastructure perspectives, model compression, and domain-specific extensions. Works such as Rethinking Learning Rate[1] and Beyond Cosine Decay[10] exemplify efforts to rethink classical decay patterns, while Continual Pretraining Warmup[7] and Scalable Continual Pretraining[9] illustrate the growing interest in schedule adjustments for ongoing training. Meanwhile, studies like Critical Batch Size[15] and Compute Optimal Training[17] bridge schedule design with resource allocation, and Sharpness Disparity Principle[16] connects learning rate choices to loss landscape geometry.

Several active lines of work highlight contrasting priorities and open questions. One thread investigates whether traditional cosine or linear decay remains optimal, with papers like Straight to Zero[8] and Optimal Linear Decay[37] proposing alternatives that challenge conventional wisdom. Another thread explores checkpoint reuse and model merging strategies—such as Early Weight Averaging[5] and Reuse Dont Retrain[3]—to reduce redundant computation. Within this landscape, WSM Checkpoint Merging[0] sits naturally among novel schedule architectures, sharing thematic ties with Learning Rate Path Switching[2] and River Valley Landscape[23], all of which examine how to navigate or combine different training trajectories. Compared to Reuse Dont Retrain[3], which emphasizes recycling existing checkpoints, WSM Checkpoint Merging[0] appears to focus more directly on merging strategies as a schedule design primitive. The broader tension between discovering fundamentally new schedules versus refining existing ones for continual or resource-constrained settings remains a central question across these branches.

Related Works in Same Category

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

1. Straight to zero: Why linearly decaying the learning rate to zero works best for LLMs

Authors: Bergsma, Shane, Dey, Nolan, Gosal, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

LLMs are commonly trained with a learning rate (LR) warmup, followed by cosine decay to 10% of the maximum (10x decay). In a large-scale empirical study, we show that under an optimal peak LR, a simple linear decay-to-zero (D2Z) schedule consistently outperforms other schedules when training at compute-optimal dataset sizes. D2Z is superior across a range of model sizes, batch sizes, datasets, and vocabularies. Benefits increase as dataset size increases. Leveraging a novel interpretation of Ada...

Relationship Analysis

Both papers belong to the Novel Schedule Architectures category, introducing new learning rate schedule structures for LLM pre-training. While the original paper (WSM) proposes a decay-free approach that uses checkpoint merging to emulate various decay strategies (cosine, linear, inverse square root), the candidate paper advocates for a simple linear decay-to-zero schedule as superior to the standard cosine decay with 10% minimum. The key difference is that WSM eliminates the decay phase entirely through model merging, whereas the candidate paper retains traditional decay but argues for linear decay-to-zero as the optimal decay function.

2. Beyond Cosine Decay: On the effectiveness of Infinite Learning Rate Schedule for Continual Pre-training

Authors: Singh, Vaibhav, Janson, Paul, Ibrahim, et al. (12 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

The ever-growing availability of unlabeled data presents both opportunities and challenges for training artificial intelligence systems. While self-supervised learning (SSL) has emerged as a powerful paradigm for extracting meaningful representations from vast amounts of unlabeled data, existing methods still struggle to adapt to the non-stationary, non-IID nature of real-world data streams without forgetting previously learned knowledge. Recent works have adopted a repeated cosine annealing sch...

Relationship Analysis

Both papers belong to the Novel Schedule Architectures category by proposing alternatives to standard cosine/linear decay schedules for LLM pre-training. They overlap in addressing the limitations of traditional decay schedules and exploring decay-free or modified approaches to maintain training flexibility. However, the original paper (WSM) introduces a checkpoint merging framework that emulates various decay strategies through model averaging, while the candidate paper focuses on comparing infinite learning rate schedules against repeated cosine annealing for continual pre-training scenarios.

3. Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective

Authors: Wen, Kaiyue, Li, Zhiyuan, Wang, et al. (10 authors total) | **Year/Venue**: 2024 • arXiv.org | **URL**: View paper

Abstract

Training language models currently requires pre-determining a fixed compute budget because the typical cosine learning rate schedule depends on the total number of steps. In contrast, the Warmup-Stable-Decay (WSD) schedule uses a constant learning rate to produce a main branch of iterates that can in principle continue indefinitely without a pre-specified compute budget. Then, given any compute budget, one can branch out from the main branch at a proper time with a rapidly decaying learning rate...

Relationship Analysis

Both papers belong to the Novel Schedule Architectures category, introducing new learning rate scheduling approaches that deviate from standard cosine/linear decay strategies for LLM pre-training. They share overlapping focus on the Warmup-Stable-Decay (WSD) framework and aim to improve upon it through different mechanisms. The key difference is that the original paper (WSM) proposes eliminating the decay phase entirely by using checkpoint merging to simulate decay effects, while the candidate paper provides theoretical analysis of WSD's river valley loss landscape and introduces WSD-S, which optimizes the decay phase reuse strategy while maintaining the traditional learning rate decay approach.

Contributions Analysis

Overall novelty summary. The paper proposes WSM, a framework that connects learning rate decay strategies to checkpoint merging by deriving merge weights from decay schedules. It resides in the 'Novel Schedule Architectures' leaf, which contains four papers exploring alternatives to standard cosine or linear decay. This leaf sits within the broader 'Learning Rate Schedule Design and Optimization' branch, indicating a moderately populated research direction focused on rethinking fundamental schedule structures. The taxonomy shows that while schedule design is an active area, this specific leaf is not overcrowded, suggesting room for novel architectural contributions.

The taxonomy reveals neighboring work in 'Checkpoint Averaging and Model Merging' (two papers) and 'Theoretical Foundations and Scaling Laws' (four papers), both closely related to WSM's dual focus on merging mechanics and theoretical grounding. The 'Adaptive and Dynamic Schedule Optimization' leaf (three papers) explores online tuning, contrasting with WSM's fixed framework approach. The taxonomy's scope notes clarify that WSM's theoretical connection between decay and merging distinguishes it from purely empirical checkpoint reuse methods, positioning it at the intersection of schedule design and training efficiency.

Among seventeen candidates examined, no contribution was clearly refuted. The core WSM framework examined ten candidates with zero refutations, the theorem derivation examined one candidate with no overlap, and the merge duration finding examined six candidates without refutation. This limited search scope—seventeen papers from semantic retrieval—means the analysis captures nearby work but cannot claim exhaustive coverage. The absence of refutations among examined candidates suggests the specific framing of decay-to-merging equivalence may be underexplored, though the small sample size limits confidence in this assessment.

Based on top-seventeen semantic matches, the work appears to occupy a relatively sparse intersection between schedule architecture and checkpoint merging theory. The taxonomy structure confirms that while both schedule design and merging are active areas, their formal unification is less densely studied. However, the limited search scope means potentially relevant work in adjacent leaves or outside the top-K results may not be reflected here.

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

Contribution 1: WSM framework connecting LR decay and checkpoint merging

Description: The authors introduce WSM, a framework that theoretically connects learning rate decay strategies to checkpoint merging operations. This framework provides a unified foundation for emulating various decay strategies (cosine, linear, inverse square root) as principled model averaging schemes while remaining compatible with diverse optimization methods.

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. Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective

URL: View paper

Brief Assessment

River Valley Landscape[23] focuses on the WSD (warmup-stable-decay) schedule and analyzes its loss landscape properties, introducing WSD-S as a variant. It does not propose a framework that connects LR decay to checkpoint merging operations or demonstrate how to emulate various decay strategies through model averaging schemes.

2. How to Merge Your Multimodal Models Over Time?

URL: View paper

Brief Assessment

Merge Multimodal Over Time[58] focuses on temporal merging of multimodal expert models across tasks, not on learning rate scheduling or checkpoint merging as a replacement for LR decay in single-model training.

3. Hop, skip, jump to convergence: Dynamics of learning rate transitions for improved training of large language models

URL: View paper

Brief Assessment

Hop Skip Jump[4] focuses on analyzing the dynamics of abrupt learning rate transitions during training (the 'SkipLR' phenomenon), not on establishing a framework that connects learning rate decay strategies to checkpoint merging operations. The candidate studies how switching learning rates causes loss curves to contract, while the original proposes WSM as a practical training methodology that replaces decay phases with checkpoint merging.

4. How to Merge Multimodal Models Over Time?

URL: View paper

Brief Assessment

Merge Multimodal Models[59] focuses on temporal merging of multimodal expert models across tasks over time, not on learning rate scheduling or connecting LR decay to checkpoint merging for LLM pre-training.

5. Improved Cotton Leaf Disease Classification Using Parameter-Efficient Deep Learning Framework

URL: View paper

Brief Assessment

Cotton Leaf Disease[61] focuses on image classification for agricultural disease detection using MobileNet with frozen layers and standard callbacks (ModelCheckpoint, EarlyStopping, ReduceLROnPlateau). It does not address learning rate scheduling frameworks, checkpoint merging strategies, or the theoretical connection between LR decay and model averaging that characterizes the original paper's WSM contribution.

6. Surge phenomenon in optimal learning rate and batch size scaling

URL: View paper

Brief Assessment

Surge Phenomenon[42] focuses on the relationship between optimal learning rates and batch sizes for Adam-style optimizers, not on connecting learning rate decay strategies to checkpoint merging operations as a unified framework.

7. How Learning Rate Decay Wastes Your Best Data in Curriculum-Based LLM Pretraining

URL: View paper

Brief Assessment

Decay Wastes Best Data[44] focuses on the incompatibility between learning rate decay and curriculum-based data ordering in LLM pretraining, not on establishing a general framework connecting LR decay to checkpoint merging operations across diverse optimization methods.

8. JaColBERTv2. 5: Optimising Multi-Vector Retrievers to Create State-of-the-Art Japanese Retrievers with Constrained Resources

URL: View paper

Brief Assessment

JaColBERTv2.5[60] focuses on optimizing multi-vector retrievers for Japanese text retrieval tasks, not on learning rate scheduling frameworks or checkpoint merging methods for LLM pre-training.

9. Multimodal automl on structured tables with text fields

URL: View paper

Brief Assessment

Multimodal AutoML Tables[62] focuses on automated machine learning for multimodal data tables containing text, numeric, and categorical fields. It does not address learning rate scheduling, checkpoint merging, or LLM pre-training optimization strategies.

10. When, Where and Why to Average Weights?

URL: View paper

Brief Assessment

When Where Why Average[57] focuses on weight averaging techniques (LAWA, EMA) for improving training efficiency and generalization, but does not establish a theoretical framework connecting learning rate decay strategies to checkpoint merging operations as WSM does.

Contribution 2: Theorem for deriving checkpoint weights from decay schedules

Description: The authors formalize Theorem 3.1, which provides a principled method to derive checkpoint merging weights from any desired gradient decay schedule. This theorem enables the conversion of LR decay methods into theoretically approximate model averaging implementations.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. A Simple Baseline for Bayesian Uncertainty in Deep Learning

URL: View paper

Brief Assessment

Bayesian Uncertainty Baseline[51] focuses on uncertainty quantification using Gaussian approximations of SGD iterates for Bayesian model averaging, not on deriving checkpoint merging weights from learning rate decay schedules.

Contribution 3: Identification of merge duration as critical performance factor

Description: Through systematic experiments, the authors identify merge duration (the training window for checkpoint aggregation) as the most important factor affecting model performance in their framework, surpassing the importance of checkpoint saving intervals and the number of checkpoints merged.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Instability in Downstream Task Performance During LLM Pretraining

URL: View paper

Brief Assessment

Downstream Task Instability[55] focuses on checkpoint integration methods to stabilize fluctuating downstream task performance during pretraining, not on identifying merge duration as a critical factor for checkpoint averaging performance in decay-free learning rate schedules.

2. Window-based Model Averaging Improves Generalization in Heterogeneous Federated Learning

URL: View paper

Brief Assessment

Window Model Averaging[56] focuses on federated learning scenarios with heterogeneous data distributions across clients, not on checkpoint merging during centralized LLM pre-training. The 'window' in their work refers to aggregating global models across federated rounds, not training duration for checkpoint aggregation in a single training run.

3. Scaling laws and compute-optimal training beyond fixed training durations

URL: View paper

Brief Assessment

Scaling Laws Beyond Fixed[18] focuses on training duration flexibility and compute-optimal training with constant learning rates and cooldowns, not on checkpoint merging or merge duration as a performance factor. The candidate's investigation centers on learning rate schedules and their alternatives, which is a different technical focus from the original paper's checkpoint aggregation framework.

4. COVID-era forecasting: Google trends and window and model averaging

URL: View paper

Brief Assessment

Google Trends Forecasting[54] focuses on window averaging for time-series forecasting with Google Trends data, not on checkpoint merging or training window duration in neural network optimization. The domains are fundamentally different (tourism demand forecasting vs. LLM pre-training).

5. LLM circuit analyses are consistent across training and scale

URL: View paper

Brief Assessment

Circuit Analyses Consistent[52] focuses on the stability of circuit components and algorithms across training checkpoints in LLMs, not on checkpoint merging strategies or merge duration parameters. The candidate studies how mechanistic circuits remain consistent over time, while the original contribution concerns optimizing training schedules through checkpoint aggregation windows.

6. Sequential manifold regularization for large language model contextual stability

URL: View paper

Brief Assessment

Sequential Manifold Regularization[53] focuses on contextual stability through manifold regularization techniques, not on systematic analysis of checkpoint averaging parameters like merge duration, checkpoint intervals, or merge quantity in training frameworks.

Appendix: Text Similarity Detection

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

References

- [0] WSM: Decay-Free Learning Rate Schedule via Checkpoint Merging for LLM Pre-training View paper
- [1] Rethinking learning rate tuning in the era of large language models View paper
- [2] A learning rate path switching training paradigm for version updates of large language models View paper
- [3] Reuse, don't retrain: A recipe for continued pretraining of language models View paper

- [4] Hop, skip, jump to convergence: Dynamics of learning rate transitions for improved training of large language models View paper
- [5] Early Weight Averaging meets High Learning Rates for LLM Pre-training View paper
- [6] Length-based curriculum learning for efficient pre-training of language models View paper
- [7] Continual pre-training of large language models: How to (re) warm your model? View paper
- [8] Straight to zero: Why linearly decaying the learning rate to zero works best for LLMs View paper
- [9] Simple and Scalable Strategies to Continually Pre-train Large Language Models View paper
- [10] Beyond Cosine Decay: On the effectiveness of Infinite Learning Rate Schedule for Continual Pre-training View paper
- [11] Learning Dynamics in Continual Pre-Training for Large Language Models View paper
- [12] The Surprising Agreement Between Convex Optimization Theory and Learning-Rate Scheduling for Large Model Training View paper
- [13] AdaLRS: Loss-Guided Adaptive Learning Rate Search for Efficient Foundation Model Pretraining View paper
- [14] A Survey on the Scheduling of DL and LLM Training Jobs in GPU Clusters View paper
- [15] How Does Critical Batch Size Scale in Pre-training? View paper
- [16] The Sharpness Disparity Principle in Transformers for Accelerating Language Model Pre-Training View paper
- [17] An empirical analysis of compute-optimal large language model training View paper
- [18] Scaling laws and compute-optimal training beyond fixed training durations View paper
- [19] No more adam: Learning rate scaling at initialization is all you need View paper
- [20] Evaluation of pre-training large language models on leadership-class supercomputers: J. Yin et al. View paper
- [21] Large Language Model Empowered Recommendation Meets All-domain Continual Pre-Training View paper
- [22] Neural Thermodynamic Laws for Large Language Model Training View paper
- [23] Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective View paper
- [24] Dataset decomposition: Faster llm training with variable sequence length curriculum View paper
- [25] Why Gradients Rapidly Increase Near the End of Training View paper
- [26] DreamDDP: Accelerating Data Parallel Distributed LLM Training with Layer-wise Scheduled Partial Synchronization View paper
- [27] Hlat: High-quality large language model pre-trained on aws trainium View paper
- [28] Learning to schedule learning rate with graph neural networks View paper
- [29] Layer-Wise Learning Rate Optimization for Task-Dependent Fine-Tuning of Pre-Trained Models: An Evolutionary Approach View paper
- [30] Scaling law with learning rate annealing View paper
- [31] CyclicReflex: Improving Large Reasoning Models via Cyclical Reflection Token Scheduling View paper
- [32] Enhancing large language models with stochastic multi-level embedding fusion: An experimental approach on open-source llm View paper
- [33] Scaling Laws for Robust Comparison of Open Foundation Language-Vision Models and Datasets View paper
- [34] Methods of improving LLM training stability View paper
- [35] Mid-Training of Large Language Models: A Survey View paper
- [36] Learning rate decay scheduler using the loss operator as generalization View paper
- [37] Optimal linear decay learning rate schedules and further refinements View paper
- [38] Coflow Scheduling for LLM Training View paper
- [39] Optimizing the Utilization of Large Language Models via Schedule Optimization: An Exploratory Study View paper
- [40] IDEA Prune: An Integrated Enlarge-and-Prune Pipeline in Generative Language Model Pretraining View paper
- [41] Bitnet: 1-bit pre-training for large language models View paper
- [42] Surge phenomenon in optimal learning rate and batch size scaling View paper
- [43] Optimization Hyper-parameter Laws for Large Language Models View paper
- [44] How Learning Rate Decay Wastes Your Best Data in Curriculum-Based LLM Pretraining View paper
- [45] Scaling and Transferability of Annealing Strategies in Large Language Model Training View paper
- [46] When, why and how much? adaptive learning rate scheduling by refinement View paper
- [47] Structured Representation Compression for Large Language Models through Hierarchical Tensor Partitioning View paper
- [48] InLegalLLaMA: Indian Legal Knowledge Enhanced Large Language Model View paper
- [49] Spaced Scheduling for Large Language Model Training View paper
- [50] Optimal Embedding Learning Rate in LLMs: The Effect of Vocabulary Size View paper
- [51] A Simple Baseline for Bayesian Uncertainty in Deep Learning View paper
- [52] LLM circuit analyses are consistent across training and scale View paper
- [53] Sequential manifold regularization for large language model contextual stability View paper
- [54] COVID-era forecasting: Google trends and window and model averaging View paper
- [55] Instability in Downstream Task Performance During LLM Pretraining View paper
- [56] Window-based Model Averaging Improves Generalization in Heterogeneous Federated Learning View paper
- [57] When, Where and Why to Average Weights? View paper
- [58] How to Merge Your Multimodal Models Over Time? View paper
- [59] How to Merge Multimodal Models Over Time? View paper
- [60] JaColBERTv2. 5: Optimising Multi-Vector Retrievers to Create State-of-the-Art Japanese Retrievers with Constrained Resources View paper
- [61] Improved Cotton Leaf Disease Classification Using Parameter-Efficient Deep Learning Framework View paper
- [62] Multimodal automl on structured tables with text fields View paper