Novelty Assessment Report

Paper: Scaling with Collapse: Efficient and Predictable Training of LLM Families
PDF URL: https://openreview.net/pdf?id=3YKeB9R1g9
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Effective LLM training relies on consistency, meaning that key quantities—such as final losses and optimal hyperparameters—scale predictably across model sizes. Qiu et al. (2025) recently showed that this consistency extends beyond scalars: whole training loss curves can collapse onto a universal trajectory after a simple normalization. What remains unclear is whether this phenomenon holds for LLM families trained under practical scaling recipes, where width, depth, learning rate, batch size, and weight decay are scaled jointly. We show that it does: loss curves collapse across scales precisely when optimization hyperparameters are set optimally for the given data budget, in accordance with recent empirical scaling laws. Collapse thus emerges as a signature of compute-efficient training. We demonstrate two applications at scale: (1) deviation-from-collapse provides a sensitive, early diagnostic of training pathologies, and (2) the predictability of collapsed curves enables early stopping in large-scale hyperparameter tuning. Finally, we train a competitive LLM family, Celerity, using these insights, highlighting collapse as an effective tool for developing efficient LLMs.

Core Task Landscape

This paper addresses: **Training Loss Curve Prediction and Collapse Across Model Scales**

A total of **36 papers** were analyzed and organized into a taxonomy with **30 categories**.

Taxonomy Overview

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

- **Scaling Laws and Loss Prediction**
- **Loss Curve Collapse and Universality**
- **Training Dynamics and Phase Transitions**
- **Training Instabilities and Pathologies**
- **Optimization and Hyperparameter Scaling**
- **Theoretical Foundations and Mechanistic Models**
- **Data and Training Efficiency**
- **Model Compression and Efficiency**
- **Implicit Bias and Downstream Performance**
- **Domain-Specific Applications**

Complete Taxonomy Tree

- Training Loss Curve Prediction and Collapse Across Model Scales Survey Taxonomy
- Scaling Laws and Loss Prediction
  - Foundational Scaling Law Formulations (3 papers)
  - [1] Scaling laws for neural language models (Kaplan, 2020) View paper
  - [2] An empirical analysis of compute-optimal large language model training (J Hoffmann, 2022) View paper
  - [11] Unraveling the Mystery of Scaling Laws: Part I (Su Hui, 2024) View paper
  - Cross-Distribution and Transfer Prediction (1 papers)
  - [9] Loss-to-loss prediction: Scaling laws for all datasets (Brandfonbrener, 2024) View paper
  - Architecture-Specific Scaling Laws (3 papers)
  - [7] Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models (Chen Zheng-yu, 2024) View paper
  - [14] Are protein language models compute optimal? (Molina, 2024) View paper
  - [15] Efficient training of self-supervised speech foundation models on a compute budget (Andy T. Liu, 2024) View paper
  - Precision and Quantization Effects on Scaling (1 papers)
  - [4] Scaling Laws for Precision (Kumar, 2024) View paper
- Loss Curve Collapse and Universality
  - Compute-Optimal Collapse Phenomena ★ (2 papers)
  - [0] Scaling with Collapse: Efficient and Predictable Training of LLM Families (Anon et al., 2026) View paper
  - [27] Scaling Collapse Reveals Universal Dynamics in Compute-Optimally Trained Neural Networks (Qiu, 2025) View paper
  - Width-Invariant Feature Learning Consistency (1 papers)
  - [8] Feature-Learning Networks Are Consistent Across Widths At Realistic Scales (Vyas, 2023) View paper
- Training Dynamics and Phase Transitions
  - Loss Deceleration and Zero-Sum Learning (1 papers)
  - [21] Training Dynamics Underlying Language Model Scaling Laws: Loss Deceleration and Zero-Sum Learning (Mircea, 2025) View paper
  - Grokking and Critical Data Size Transitions (1 papers)

- [10] Critical data size of language models from a grokking perspective (Zhu, 2024) View paper
- Training Trajectory Analysis Across Scales (1 papers)
- [19] Training Trajectories of Language Models Across Scales (Artetxe, 2023) View paper
- Epochal and Oscillatory Loss Patterns (1 papers)
- [29] The Epochal Sawtooth Phenomenon: Unveiling Training Loss Oscillations in Adam and Other Optimizers: Q. Liu, W. Ma (Q Liu, 2025) View paper

- Training Instabilities and Pathologies
  - Large-Scale Instability Reproduction at Small Scale (1 papers)
  - [6] Small-scale proxies for large-scale transformer training instabilities (Wortsman, 2023) View paper
  - Architecture-Specific Training Collapse (1 papers)
  - [18] Learning rate collapse prevents training recurrent neural networks at scale (B Kurtkaya, 2025) View paper
  - Inverse Scaling and Performance Degradation (1 papers)
  - [5] Inverse Scaling: When Bigger Isn't Better (McKenzie, 2023) View paper
- Optimization and Hyperparameter Scaling
  - Learning Rate Scaling and Adaptive Search (1 papers)
  - [28] AdaLRS: Loss-Guided Adaptive Learning Rate Search for Efficient Foundation Model Pretraining (Dong Hongyuan, 2025) View paper
  - Learning Rate Schedules and Functional Scaling (1 papers)
  - [20] Functional Scaling Laws in Kernel Regression: Loss Dynamics and Learning Rate Schedules (Li, 2025) View paper
  - Batch Size Scaling and Data Parallelism (1 papers)
  - [17] An empirical model of large-batch training (McCandlish, 2018) View paper
  - Loss Curvature and Conditioning Effects (1 papers)
  - [13] A loss curvature perspective on training instabilities of deep learning models (J Gilmer, 2022) View paper
- Theoretical Foundations and Mechanistic Models
  - Random Feature and Kernel Models of Scaling (1 papers)
  - [31] A Dynamical Model of Neural Scaling Laws (Bordelon, 2024) View paper
  - Zipf's Law and Power-Law Task Structure (1 papers)
  - [16] AlphaZero Neural Scaling and Zipf's Law: a Tale of Board Games and Power Laws (Neumann, 2024) View paper
  - Interpolation Regime and Generalization Theory (2 papers)
  - [34] Generalization and Optimization in the Interpolation Regime: From Linear Models to Neural Networks (Hossein, 2024) View paper
  - [36] A Universal Trade-off Between the Model Size, Test Loss, and Training Loss of Linear Predictors (Nikhil Ghosh, 2023) View paper
  - Infinite-Width and Convergence Limits (2 papers)
  - [12] Infinite limits of multi-head transformer dynamics (Blake Bordelon, 2024) View paper
  - [30] Learning in Large Neural Networks (Davide Anguita, 2025) View paper
- Data and Training Efficiency
  - Loss-Based Sample Reweighting and Selection (1 papers)
  - [23] Dynamic Loss-Based Sample Reweighting for Improved Large Language Model Pretraining (Sow, 2025) View paper
  - Synthetic Data Mixing and Training Effects (1 papers)
  - [24] Characterizing Model Behavior Under Synthetic Data Training: An Empirical Study Across Scales and Mixing Ratios (Du Y, 2025) View paper
  - Vocabulary Size and Tokenization Effects (1 papers)
  - [22] Exploiting Vocabulary Frequency Imbalance in Language Model Pre-training (Chung Woojin, 2025) View paper
  - Early Stopping and Hyperparameter Tuning via Loss Prediction (1 papers)
  - [35] nanoLM: an Affordable LLM Pre-training Benchmark via Accurate Loss Prediction across Scales (Yao Yi-qun, 2023) View paper
- Model Compression and Efficiency
  - Low-Rank Factorization and SVD-Based Compression (1 papers)
  - [32] Integrating Independent Layer-Wise Rank Selection with Low-Rank SVD Training for Model Compression: A Theory-Driven Approach (Yifan Guo, 2024) View paper
  - Dense Layer Replacement with Efficient Structures (1 papers)
  - [33] Exploration of replacing Dense Layers with Higher Efficiency Structures (Li, 2025) View paper
  - Unlearning Geometry and Loss Dynamics (1 papers)
  - [25] The Geometry of Forgetting: Analyzing Machine Unlearning through Local Learning Coefficients (A Muhamed, 2025) View paper
- Implicit Bias and Downstream Performance (1 papers)
  - [3] Same pre-training loss, better downstream: Implicit bias matters for language models (Liu Hong, 2023) View paper
- Domain-Specific Applications (1 papers)
  - [26] Towards Real-Time Monitoring of High-Voltage Insulators: Progressive Flashover Classification Using Quantized Deep Learning (Khan, 2025) View paper

Narrative

Core task: training loss curve prediction and collapse across model scales. The field investigates how neural network training loss evolves as a function of model size, data, and compute, seeking predictable patterns that generalize across scales. The taxonomy organizes this landscape into several major branches. Scaling Laws and Loss Prediction focuses on empirical power-law relationships that forecast final performance from resource budgets, exemplified by foundational work like Scaling Laws[1] and compute-optimal studies such as Chinchilla[2]. Loss Curve Collapse and Universality examines whether training curves from different model sizes can be mapped onto a single master curve, revealing universal structure in optimization dynamics. Training Dynamics and Phase Transitions studies abrupt changes in learning behavior, while Training Instabilities and Pathologies addresses phenomena like loss spikes and divergence. Optimization and Hyperparameter Scaling explores how learning rates and batch sizes should adapt with model scale, and Theoretical Foundations seeks mechanistic explanations for observed regularities. Additional branches cover data efficiency, model compression, implicit bias effects on downstream tasks, and domain-specific applications ranging from protein language models to reinforcement learning.

Recent work has intensified focus on whether loss curves truly collapse in a universal manner and what this implies for predicting large-scale behavior from small-scale proxies. Studies like Small-scale Proxies[6] and Loss-to-loss Prediction[9] explore whether cheaper pilot

runs can reliably forecast expensive training outcomes, while Feature-Learning Consistency[8] investigates whether internal representations evolve similarly across scales. The original paper, Scaling with Collapse[0], sits squarely within the Loss Curve Collapse and Universality branch, specifically addressing compute-optimal collapse phenomena. It shares thematic ground with Scaling Collapse Universal[27], which also examines universal collapse properties, but Scaling with Collapse[0] emphasizes how collapse behavior manifests under compute-optimal training regimes where model size and data are jointly scaled. This contrasts with earlier scaling law studies like Scaling Laws Precision[4] that focused primarily on predictive accuracy of power laws rather than the geometric structure of curve families, highlighting an evolving interest in deeper invariances beyond simple extrapolation formulas.

Related Works in Same Category

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

1. Scaling Collapse Reveals Universal Dynamics in Compute-Optimally Trained Neural Networks

Authors: Qiu, Shikai, Xiao, Lechao, Wilson, et al. (10 authors total) | **Year/Venue**: 2025 • arXiv (Cornell University) | **URL**: View paper

Abstract

What scaling limits govern neural network training dynamics when model size and training time grow in tandem? We show that despite the complex interactions between architecture, training algorithms, and data, compute-optimally trained models exhibit a remarkably precise universality. Specifically, loss curves from models of varying sizes collapse onto a single universal curve when training compute and loss are normalized to unity at the end of training. With learning rate decay, the collapse bec...

Relationship Analysis

Both papers belong to the Compute-Optimal Collapse Phenomena category, investigating how training loss curves collapse under compute-optimal conditions and using this collapse as a diagnostic tool. They overlap in demonstrating that normalized loss curves collapse across model scales when trained compute-optimally, and both explore collapse as a signature of efficient training. The key difference is that the original paper (Scaling with Collapse) focuses on practical LLM training at scale with the Celerity family, emphasizing the role of AdamW timescale τ and TPP ratio in achieving collapse, while the candidate paper (Scaling Collapse Reveals Universal Dynamics) provides deeper theoretical analysis of collapse mechanisms through power-law scaling laws and SGD noise dynamics, introducing the concept of "supercollapse" where deviations fall below noise floors.

Contributions Analysis

Overall novelty summary. The paper demonstrates that training loss curves collapse onto a universal trajectory when optimization hyperparameters are scaled optimally with model size and data budget. It resides in the 'Compute-Optimal Collapse Phenomena' leaf, which contains only two papers total, indicating a relatively sparse research direction within the broader 'Loss Curve Collapse and Universality' branch. The work extends recent findings on loss curve collapse by showing the phenomenon holds under practical joint scaling of width, depth, learning rate, batch size, and weight decay—a setting closer to real-world LLM training than prior studies.

The taxonomy reveals that this work sits at the intersection of multiple research threads. Its parent branch 'Loss Curve Collapse and Universality' is adjacent to 'Scaling Laws and Loss Prediction', which contains foundational power-law studies across seven papers in four sub-categories. Neighboring branches include 'Training Dynamics and Phase Transitions' (examining temporal behavior) and 'Optimization and Hyperparameter Scaling' (studying how hyperparameters adapt with scale). The paper bridges these areas by linking collapse phenomena to compute-optimal hyperparameter choices, connecting geometric curve structure to optimization efficiency in ways that prior scaling law formulations did not emphasize.

Among 29 candidates examined, the core collapse demonstration (Contribution 1) shows one refutable candidate from 9 examined, suggesting some prior work on collapse exists but coverage is limited. The Celerity LLM family (Contribution 2) examined 10 candidates with none refutable, indicating the specific model instantiation appears novel within this search scope. The early stopping method (Contribution 3) found 2 refutable candidates among 10 examined, suggesting related hyperparameter tuning approaches exist. The limited search scale means these statistics reflect top-semantic-match coverage rather than exhaustive field surveys, and the sparse taxonomy leaf suggests this research direction remains relatively unexplored.

Given the small sibling set and limited candidate pool examined, the work appears to occupy a genuinely sparse area where loss curve collapse meets compute-optimal training. The analysis covers top-30 semantic matches and does not claim exhaustive coverage of all hyperparameter tuning or scaling law literature. The taxonomy structure suggests the field is actively fragmenting into specialized sub-problems, with this paper carving out a niche at the intersection of collapse phenomena and practical scaling recipes.

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

Contribution 1: Demonstration that training loss curves collapse under optimal hyperparameter scaling

Description: The authors show that normalized training loss curves (TLCs) collapse onto a universal trajectory across different model sizes when the AdamW timescale τ, tokens-per-parameter ratio (TPP), and learning rate schedule are properly aligned. This collapse emerges as a signature of compute-efficient training.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Scaling laws for neural language models

URL: View paper

Brief Assessment

Scaling Laws[1] focuses on power-law relationships between loss and compute/model/data size, not on normalized training loss curve collapse across scales under optimal hyperparameter alignment (τ, TPP, learning rate schedule).

---

2. nanoLM: an Affordable LLM Pre-training Benchmark via Accurate Loss Prediction across Scales

URL: View paper

Brief Assessment

nanoLM[35] focuses on predicting final training loss values across model scales using μP-based hyperparameter transfer, not on demonstrating that normalized training loss curves collapse onto a universal trajectory when hyperparameters are optimally scaled.

---

3. Critical Batch Size Revisited: A Simple Empirical Approach to Large-Batch Language Model Training

URL: View paper

Brief Assessment

Critical Batch Size[51] focuses on measuring critical batch size through branched training experiments and does not address training loss curve collapse across model scales with optimal hyperparameters.

---

4. Scaling Collapse Reveals Universal Dynamics in Compute-Optimally Trained Neural Networks

URL: View paper

Prior Art Analysis

Scaling Collapse Universal[27] demonstrates that compute-optimally trained models exhibit training loss curve collapse when normalized appropriately. The paper shows that 'loss curves from models of varying sizes collapse onto a single universal curve when training compute and loss are normalized to unity at the end of training' and introduces the concept of 'supercollapse' where differences fall below noise floors. This work was published prior to the original paper and explicitly demonstrates the same phenomenon of normalized training loss curves collapsing across different model sizes under compute-optimal conditions, directly challenging the novelty claim.

Evidence

Evidence 1 - **Rationale**: The original paper explicitly cites 'qiu et al. (2025)' as prior work demonstrating TLC collapse, and the candidate paper (Scaling Collapse Universal[27]) shows this phenomenon across multiple architectures and conditions, establishing prior demonstration of the collapse phenomenon. - **Original**: qiu et al. (2025) only recently demonstrated this striking regularity in tlcs, showing collapse when training with μp on small-scale autoregressive tasks. - **Candidate**: we observe supercollapse across learning rate schedules, datasets, and architectures, including transformers trained on next-token prediction, and find it breaks down when hyperparameters are scaled suboptimally, providing a precise and practical indicator of good scaling.

Evidence 2 - **Rationale**: Both papers establish that collapse occurs under optimal hyperparameter settings aligned with compute-optimal training, demonstrating the same fundamental insight about the conditions for collapse. - **Original**: we show that it does: loss curves collapse across scales precisely when optimization hyperparameters are set optimally for the given data budget, in accordance with recent empirical scaling laws. - **Candidate**: we first show that for loss curves following typical neural scaling laws, collapse occurs precisely when models are trained for constant multiples of their compute-optimal horizons

---

5. Hyperparameter Transfer Enables Consistent Gains of Matrix-Preconditioned Optimizers Across Scales

URL: View paper

Brief Assessment

Hyperparameter Transfer Preconditioned[52] focuses on hyperparameter transfer for matrix-preconditioned optimizers (Shampoo, SOAP, Muon) across model scales, not on training loss curve collapse phenomena. The paper studies learning rate and weight decay scaling rules to achieve consistent optimizer performance, which is a different technical contribution from demonstrating universal training loss curve trajectories.

---

6. Warmstarting for scaling language models

URL: View paper

Brief Assessment

Warmstarting Scaling[50] focuses on warmstarting techniques for transferring weights and hyperparameters across model scales using μTransfer, not on training loss curve collapse phenomena or optimal hyperparameter scaling laws.

---

7. Resolving discrepancies in compute-optimal scaling of language models

URL: View paper

Brief Assessment

Compute-optimal Discrepancies[47] focuses on resolving discrepancies between different compute-optimal scaling laws (Kaplan et al. vs Hoffmann et al.) by analyzing factors like flop counting, warmup duration, and optimizer tuning. It does not address training loss curve collapse across model scales as a phenomenon or signature of compute-efficient training.

---

8. Exploring molecular pretraining model at scale

URL: View paper

Brief Assessment

Molecular Pretraining Scale[49] focuses on molecular pretraining models and scaling laws in chemistry/biology domains, not on training loss curve collapse across language model scales with hyperparameter optimization.

---

9. Simplifying DINO via Coding Rate Regularization

URL: View paper

Brief Assessment

Simplifying DINO[48] focuses on simplifying self-supervised learning pipelines through coding rate regularization, not on training loss curve collapse across model scales with optimal hyperparameters.

Contribution 2: Celerity LLM family trained with collapse-inducing hyperparameter scaling

Description: The authors introduce Celerity, the first large-scale LLM family (300M–3.9B parameters) explicitly trained in fixed-TPP bands with optimal τ scaling to achieve training loss curve collapse. This family demonstrates compute-efficiency and provides practical validation of collapse principles at scale.

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. Jet-Nemotron: Efficient Language Model with Post Neural Architecture Search

URL: View paper

Brief Assessment

Jet-Nemotron[42] focuses on neural architecture search for hybrid-architecture models with linear attention blocks, not on training loss curve collapse or hyperparameter scaling principles for compute-efficient training.

---

2. Minicpm: Unveiling the potential of small language models with scalable training strategies

URL: View paper

Brief Assessment

MiniCPM[40] focuses on small language models (1.2B-2.4B parameters) using a WSD (warmup-stable-decay) learning rate scheduler for continuous training and domain adaptation. While both papers explore compute-efficient training strategies, MiniCPM[40] does not demonstrate training loss curve collapse across model scales as a signature of compute-efficient training, nor does it explicitly train model families in fixed tokens-per-parameter bands with optimal τ scaling.

---

3. Tuning large neural networks via zero-shot hyperparameter transfer

URL: View paper

Brief Assessment

Zero-shot Hyperparameter Transfer[39] focuses on transferring hyperparameters across model widths using maximal update parametrization (μP), not on training LLM families with collapse-inducing hyperparameter scaling or demonstrating training loss curve collapse as a signature of compute-efficient training.

---

4. Scaling laws for generative mixed-modal language models

URL: View paper

Brief Assessment

Mixed-modal Scaling[45] focuses on scaling laws for generative mixed-modal language models across different modalities (text, speech, images, code), not on training loss curve collapse or hyperparameter scaling for compute-efficient training within a single modality family.

---

5. Communication-Efficient Language Model Training Scales Reliably and Robustly: Scaling Laws for DiLoCo

URL: View paper

Brief Assessment

DiLoCo Scaling[44] focuses on distributed training with relaxed synchronization and scaling laws for communication-efficient training, not on training loss curve collapse or hyperparameter scaling to achieve collapse across model sizes.

---

6. Scaling laws for differentially private language models

URL: View paper

Brief Assessment

Differentially Private Scaling[46] focuses on privacy-preserving language model training with differential privacy constraints, not on training loss curve collapse or compute-efficient hyperparameter scaling without privacy considerations.

---

7. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer

URL: View paper

Brief Assessment

Tensor Programs Five[38] focuses on zero-shot hyperparameter transfer across model widths using maximal update parametrization (μP), not on training loss curve collapse or compute-efficient LLM families with optimal τ scaling.

---

8. Cerebras-GPT: Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster

URL: View paper

Brief Assessment

Cerebras-GPT[41] focuses on compute-optimal training following Chinchilla scaling rules (20 tokens per parameter) without exploring training loss curve collapse or hyperparameter scaling for collapse. The candidate does not address the collapse phenomenon or τ-based hyperparameter optimization that defines Celerity's novelty.

---

9. Scaling data-constrained language models

URL: View paper

Brief Assessment

Data-constrained Scaling[37] focuses on training language models under data constraints using repeated data and scaling laws for compute allocation, not on training loss curve collapse or hyperparameter scaling to achieve collapse across model sizes.

---

10. A system for massively parallel hyperparameter tuning

URL: View paper

Brief Assessment

Massively Parallel Tuning[43] focuses on hyperparameter optimization infrastructure and the ASHA algorithm for distributed tuning, not on training LLM families with specific hyperparameter scaling recipes to achieve training loss curve collapse.

Contribution 3: Early stopping method for hyperparameter tuning using collapse predictions

Description: The authors propose a functional form for normalized TLCs that can be fit on small-scale runs and used to extrapolate final loss from partial trajectories. This enables selecting optimal hyperparameters after only 10–30% of training, significantly reducing tuning compute costs.

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. Improving Hyperparameter Optimization with Checkpointed Model Weights

URL: View paper

Brief Assessment

Checkpointed Model Weights[54] focuses on using logged network weights in a Gaussian process surrogate model for hyperparameter optimization, not on predicting loss curves from partial training trajectories for early stopping.

---

2. Optimizing coronary artery disease diagnosis: a heuristic approach using robust data preprocessing and automated hyperparameter tuning of eXtreme gradient â¦

URL: View paper

Brief Assessment

Coronary Artery Optimization[57] uses early stopping as a built-in XGBoost feature for model training, not as a method for hyperparameter tuning via loss curve prediction. The contexts are fundamentally different: medical diagnosis optimization versus LLM training efficiency.

---

3. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves.

URL: View paper

Prior Art Analysis

Extrapolation Learning Curves[62] demonstrates prior work on early stopping for hyperparameter tuning using learning curve prediction. The candidate paper presents a probabilistic model that extrapolates performance from partial learning curves to enable early termination of poor hyperparameter configurations. Both papers address the same core problem: predicting final performance from

partial training trajectories to enable early stopping in hyperparameter optimization. The candidate explicitly states their method 'mimic[s] the early termination of bad runs using a probabilistic model that extrapolates the performance from the first part of a learning curve' and demonstrates 'predictive termination speeds up current hyperparameter optimization methods for dnns by roughly a factor of two.' This directly challenges the novelty claim that the original authors were first to propose using loss curve extrapolation for early stopping in hyperparameter tuning.

Evidence

Evidence 1 - **Rationale**: Both papers propose using learning curve extrapolation to enable early termination. The candidate explicitly describes a probabilistic model for extrapolating from partial learning curves, which is the same fundamental approach as the original paper's predictive model for normalized TLCs. - **Original**: we show collapse enables principled early stopping in tuning, and introduce a predictive model-fit at small scales, and re-used to extrapolate large-scale tlcs. - **Candidate**: in this work, we mimic this early termination of bad runs with the help of a probabilistic model that extrapolates performance from the first part of a learning curve to its remainder, enabling us to automatically identify and terminate bad runs to save time.

Evidence 2 - **Rationale**: Both papers demonstrate that early stopping based on learning curve prediction significantly reduces hyperparameter tuning time. The candidate shows 2x speedup, while the original shows reliable selection at 10-30% of training—both achieving the same goal of computational savings through early prediction. - **Original**: predicted best achieves negligible loss gaps when stopping after just 30% and 10% of training, respectively... key takeaway 3: collapse makes early stops reliable: align each tlc to a small-scale predictor, infer l(t), and choose the best hyperparameters by 10-30% of training-saving tuning compute. - **Candidate**: experiments with different neural network architectures on the prominent object recognition benchmarks cifar-10, cifar-100 and mnist show that predictive termination speeds up current hyperparameter optimization methods for dnns by roughly a factor of two, enabling them to find dnn settings that yiel...

---

4. Learning curve prediction with Bayesian neural networks

URL: View paper

Brief Assessment

Bayesian Learning Curves[61] focuses on predicting learning curves for individual hyperparameter configurations to enable early termination of poorly-performing runs. The original paper's contribution involves using collapse phenomena across model scales to enable early stopping in hyperparameter tuning by extrapolating final loss from partial trajectories at 10-30% of training. These are technically distinct approaches: one predicts individual curve behavior, the other exploits scale-invariant collapse patterns.

---

5. On the difficulty of DNN hyperparameter optimization using learning curve prediction

URL: View paper

Brief Assessment

Learning Curve Difficulty[59] focuses on the challenges and limitations of using learning curve prediction for early termination in hyperparameter optimization, demonstrating that effectiveness varies drastically with task and hyperparameter choices. The original paper proposes a specific functional form based on collapse theory for LLM training curves, which is a distinct technical approach not addressed in the candidate.

---

6. Neural Velocity for hyperparameter tuning

URL: View paper

Brief Assessment

Neural Velocity[60] focuses on early stopping using neural velocity (rate of change of neuron transfer functions) rather than loss curve collapse predictions. The candidate does not demonstrate prior work on predicting final loss from normalized training loss curves or using collapse phenomena for hyperparameter selection.

---

7. Scaling laws for hyperparameter optimization

URL: View paper

Prior Art Analysis

Hyperparameter Optimization Scaling[53] demonstrates that learning curves can be predicted using power law functions, enabling early stopping in hyperparameter tuning after observing only partial training trajectories. The paper shows that by fitting power law models on small-scale runs, they can extrapolate final performance and select optimal hyperparameters after 10-30% of training, which directly addresses the same problem as the original paper's contribution of using collapse predictions for early stopping.

Evidence

Evidence 1 - **Rationale**: Both papers propose methods for early stopping in hyperparameter tuning by predicting learning curves. The candidate uses power law models while the original uses normalized training loss curves (TLCs), but both enable stopping training early based on predictions. - **Original**: we propose a simple functional form for normalized tlcs, and showing that fitting this form on small-scale training runs enables early stopping in large-scale hyperparameter tuning - **Candidate**: in this work, we propose deep power laws (dpl), an ensemble of neural network models conditioned to yield predictions that follow a power-law scaling pattern. our method dynamically decides which configurations to pause and train incrementally by making use of gray-box evaluations.

Evidence 2 - **Rationale**: Both papers demonstrate the ability to predict final performance from partial learning curves and use this for early stopping in hyperparameter tuning, achieving similar goals of reducing computational costs. - **Original**: collapse makes early stops reliable: align each tlc to a small-scale predictor, infer l(t), and choose the best hyperparameters by 10-30% of training-saving tuning compute. - **Candidate**: in this experiment, we evaluate the predictive performance of forecasting models that given a fraction of the observed learning curve, estimate the remaining unobserved segment of the curve, on the lcbench benchmark.

---

8. Surpassing early stopping: A novel correlation-based stopping criterion for neural networks

URL: View paper

Brief Assessment

Correlation-based Stopping[55] focuses on preventing overfitting by monitoring train-validation correlation divergence, not on hyperparameter selection via loss curve extrapolation from partial trajectories as in the original paper.

---

9. Keeping deep learning models in check: A history-based approach to mitigate overfitting

URL: View paper

Brief Assessment

History-based Overfitting[56] focuses on detecting and preventing overfitting in software engineering deep learning models using validation loss histories, not on hyperparameter tuning via loss curve extrapolation for LLM training.

---

10. Early stopping on CNN-LSTM development to improve classification performance

URL: View paper

Brief Assessment

CNN-LSTM Early Stopping[58] focuses on preventing overfitting in CNN-LSTM models during training by monitoring validation loss, not on hyperparameter tuning or loss curve prediction for selecting optimal hyperparameters across model scales.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Scaling with Collapse: Efficient and Predictable Training of LLM Families View paper
- [1] Scaling laws for neural language models View paper
- [2] An empirical analysis of compute-optimal large language model training View paper
- [3] Same pre-training loss, better downstream: Implicit bias matters for language models View paper
- [4] Scaling Laws for Precision View paper
- [5] Inverse Scaling: When Bigger Isn't Better View paper
- [6] Small-scale proxies for large-scale transformer training instabilities View paper
- [7] Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models View paper
- [8] Feature-Learning Networks Are Consistent Across Widths At Realistic Scales View paper
- [9] Loss-to-loss prediction: Scaling laws for all datasets View paper
- [10] Critical data size of language models from a grokking perspective View paper
- [11] Unraveling the Mystery of Scaling Laws: Part I View paper
- [12] Infinite limits of multi-head transformer dynamics View paper
- [13] A loss curvature perspective on training instabilities of deep learning models View paper
- [14] Are protein language models compute optimal? View paper
- [15] Efficient training of self-supervised speech foundation models on a compute budget View paper
- [16] AlphaZero Neural Scaling and Zipf's Law: a Tale of Board Games and Power Laws View paper
- [17] An empirical model of large-batch training View paper
- [18] Learning rate collapse prevents training recurrent neural networks at scale View paper
- [19] Training Trajectories of Language Models Across Scales View paper
- [20] Functional Scaling Laws in Kernel Regression: Loss Dynamics and Learning Rate Schedules View paper
- [21] Training Dynamics Underlying Language Model Scaling Laws: Loss Deceleration and Zero-Sum Learning View paper
- [22] Exploiting Vocabulary Frequency Imbalance in Language Model Pre-training View paper
- [23] Dynamic Loss-Based Sample Reweighting for Improved Large Language Model Pretraining View paper
- [24] Characterizing Model Behavior Under Synthetic Data Training: An Empirical Study Across Scales and Mixing Ratios View paper
- [25] The Geometry of Forgetting: Analyzing Machine Unlearning through Local Learning Coefficients View paper
- [26] Towards Real-Time Monitoring of High-Voltage Insulators: Progressive Flashover Classification Using Quantized Deep Learning View paper
- [27] Scaling Collapse Reveals Universal Dynamics in Compute-Optimally Trained Neural Networks View paper
- [28] AdaLRS: Loss-Guided Adaptive Learning Rate Search for Efficient Foundation Model Pretraining View paper
- [29] The Epochal Sawtooth Phenomenon: Unveiling Training Loss Oscillations in Adam and Other Optimizers: Q. Liu, W. Ma View paper
- [30] Learning in Large Neural Networks View paper
- [31] A Dynamical Model of Neural Scaling Laws View paper
- [32] Integrating Independent Layer-Wise Rank Selection with Low-Rank SVD Training for Model Compression: A Theory-Driven Approach View paper
- [33] Exploration of replacing Dense Layers with Higher Efficiency Structures View paper
- [34] Generalization and Optimization in the Interpolation Regime: From Linear Models to Neural Networks View paper
- [35] nanoLM: an Affordable LLM Pre-training Benchmark via Accurate Loss Prediction across Scales View paper
- [36] A Universal Trade-off Between the Model Size, Test Loss, and Training Loss of Linear Predictors View paper
- [37] Scaling data-constrained language models View paper
- [38] Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer View paper
- [39] Tuning large neural networks via zero-shot hyperparameter transfer View paper
- [40] Minicpm: Unveiling the potential of small language models with scalable training strategies View paper
- [41] Cerebras-GPT: Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster View paper
- [42] Jet-Nemotron: Efficient Language Model with Post Neural Architecture Search View paper
- [43] A system for massively parallel hyperparameter tuning View paper
- [44] Communication-Efficient Language Model Training Scales Reliably and Robustly: Scaling Laws for DiLoCo View paper
- [45] Scaling laws for generative mixed-modal language models View paper
- [46] Scaling laws for differentially private language models View paper
- [47] Resolving discrepancies in compute-optimal scaling of language models View paper
- [48] Simplifying DINO via Coding Rate Regularization View paper
- [49] Exploring molecular pretraining model at scale View paper
- [50] Warmstarting for scaling language models View paper
- [51] Critical Batch Size Revisited: A Simple Empirical Approach to Large-Batch Language Model Training View paper
- [52] Hyperparameter Transfer Enables Consistent Gains of Matrix-Preconditioned Optimizers Across Scales View paper
- [53] Scaling laws for hyperparameter optimization View paper

- [54] Improving Hyperparameter Optimization with Checkpointed Model Weights View paper
- [55] Surpassing early stopping: A novel correlation-based stopping criterion for neural networks View paper
- [56] Keeping deep learning models in check: A history-based approach to mitigate overfitting View paper
- [57] Optimizing coronary artery disease diagnosis: a heuristic approach using robust data preprocessing and automated hyperparameter tuning of eXtreme gradient â¦ View paper
- [58] Early stopping on CNN-LSTM development to improve classification performance View paper
- [59] On the difficulty of DNN hyperparameter optimization using learning curve prediction View paper
- [60] Neural Velocity for hyperparameter tuning View paper
- [61] Learning curve prediction with Bayesian neural networks View paper
- [62] Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. View paper