Novelty Assessment Report

Paper: Pre-training under infinite compute
PDF URL: https://openreview.net/pdf?id=ck0aZTAnwK
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Since compute grows much faster than web text available for language model pre-training, we ask how one should approach pre-training under fixed data and no compute constraints. We first show that existing data-constrained approaches of increasing epoch count and parameter count overfit, and we improve upon such recipes by tuning regularization, finding that the optimal weight decay is $30\times$ larger than standard practice. Since our regularized recipe monotonically decreases loss following a power law in parameter count, we estimate its best possible performance via the \textbf{asymptote} of its scaling law rather than the performance at a fixed compute budget. We then identify that ensembling independently trained models achieves a significantly lower loss asymptote than the regularized recipe. Our best intervention combining epoching, regularization, parameter scaling, and ensemble scaling achieves an asymptote at 200M tokens using $5.17\times$ less data than our baseline, and our data scaling laws predict that this improvement persists at higher token budgets. We find that our data efficiency gains can be realized at smaller parameter counts as we can distill an ensemble into a student model that is 8$\times$ smaller and retains 83% of the ensembling benefit. Finally, our interventions designed for validation loss generalize to downstream benchmarks, achieving a 9% improvement for pre-training evals. Our results show that simple algorithmic improvements can enable significantly more data-efficient pre-training in a compute-rich future.

Core Task Landscape

This paper addresses: **data-efficient language model pre-training under compute abundance**

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:

- **Compute-Optimal Scaling and Resource Allocation**
- **Data Curation and Selection Methods**
- **Adaptive and Continual Pre-Training**
- **Sample-Efficient Training Objectives and Architectures**
- **Multimodal and Cross-Modal Pre-Training**
- **System Infrastructure and Distributed Training**
- **Privacy-Preserving and Federated Training**
- **Benchmarking and Evaluation Frameworks**
- **Transfer Learning and Model Reuse**
- **Security and Robustness**
- ... and 1 more categories

Complete Taxonomy Tree

- data-efficient language model pre-training under compute abundance Survey Taxonomy
- Compute-Optimal Scaling and Resource Allocation
  - Scaling Law Formulation and Analysis ★ (6 papers)
  - [0] Pre-training under infinite compute (Anon et al., 2026) View paper
  - [1] Training compute-optimal large language models (Hoffmann, 2022) View paper
  - [18] Scaling laws revisited: modeling the role of data quality in language model pretraining (Subramanyam, 2025) View paper
  - [28] An empirical analysis of compute-optimal large language model training (J Hoffmann, 2022) View paper
  - [30] Scaling Laws for Predicting Downstream Performance in LLMs (Chen Yangyi, 2024) View paper
  - [49] Compute-Optimal LLMs Provably Generalize Better With Scale (Finzi, 2025) View paper
  - Empirical Training Strategies and Recipes (3 papers)
  - [4] Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster (Dey, 2023) View paper
  - [15] Cramming: Training a Language Model on a single GPU in one day. (Geiping, 2023) View paper
  - [47] A Minimalist Optimizer Design for LLM Pretraining (Li Jia-Xiang, 2025) View paper
- Data Curation and Selection Methods
  - Quality-Based Data Filtering (7 papers)
  - [7] Improving pretraining data using perplexity correlations (Thrush, 2024) View paper
  - [8] Language models improve when pretraining data matches target tasks (Mizrahi David, 2025) View paper
  - [10] How to train data-efficient llms (Sachdeva, 2024) View paper
  - [26] Dataman: Data manager for pre-training large language models (Peng Ru, 2025) View paper
  - [38] Enhancing Multilingual LLM Pretraining with Model-Based Data Selection (Messmer, 2025) View paper
  - [41] Judging Quality Across Languages: A Multilingual Approach to Pretraining Data Filtering with Language Models (Ali Mehdi, 2025) View paper

    - [44] AttentionInfluence: Adopting Attention Head Influence for Weak-to-Strong Pretraining Data Selection (Hua Kai, 2025) View paper
    - Diversity and Coverage Optimization (6 papers)
    - [12] Doremi: Optimizing data mixtures speeds up language model pretraining (Xie, 2023) View paper
    - [21] Entropy law: The story behind data compression and llm performance (Mingjia Yin, 2024) View paper
    - [22] Combatting dimensional collapse in llm pre-training data via diversified file selection (Fan Ziqing, 2025) View paper
    - [27] Regmix: Data mixture as regression for language model pre-training (Liu Qian, 2024) View paper
    - [34] Multi-agent collaborative data selection for efficient llm pretraining (Tianyi Bai, 2024) View paper
    - [43] DataDecide: How to Predict Best Pretraining Data with Small Experiments (Magnusson, 2025) View paper
    - Deduplication Techniques (1 papers)
    - [29] Softdedup: an efficient data reweighting method for speeding up language model pre-training (HE Nan, 2024) View paper
    - Synthetic Data Generation (3 papers)
    - [14] Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls (Kang, 2025) View paper
    - [25] Rephrasing the web: A recipe for compute and data-efficient language modeling (Maini, 2024) View paper
    - [35] Seed-Free Synthetic Data Generation Framework for Instruction-Tuning LLMs: A Case Study in Thai (Parinthapat Pengpun, 2024) View paper
- Adaptive and Continual Pre-Training
    - Domain and Task Adaptation (4 papers)
    - [3] Unsupervised corpus aware language model pre-training for dense passage retrieval (Luyu Gao, 2022) View paper
    - [6] Don't stop pretraining: Adapt language models to domains and tasks (Gururangan, 2020) View paper
    - [31] Nuner: Entity recognition encoder pre-training via llm-annotated data (Bogdanov Sergei, 2024) View paper
    - [40] Cost-Efficient Domain-Adaptive Pretraining of Language Models for Optoelectronics Applications. (Dingyun Huang, 2025) View paper
    - Temporal and Continual Learning (2 papers)
    - [36] Continual Pre-training of MoEs: How robust is your router? (ThÃ©rien, 2025) View paper
    - [48] TiC-LM: A Web-Scale Benchmark for Time-Continual LLM Pretraining (Li, 2025) View paper
    - Context Length Scaling (1 papers)
    - [9] Data engineering for scaling language models to 128k context (Fu Yao, 2024) View paper
- Sample-Efficient Training Objectives and Architectures
    - Alternative Pre-Training Objectives (1 papers)
    - [11] Electra: Pre-training text encoders as discriminators rather than generators (Kevin Clark, 2020) View paper
    - Model Architecture Variants (2 papers)
    - [32] Stacking your transformers: A closer look at model growth for efficient llm pre-training (Du, 2024) View paper
    - [37] Cpm-2: Large-scale cost-effective pre-trained language models (Zhengyan Zhang, 2021) View paper
- Multimodal and Cross-Modal Pre-Training
    - Vision-Language Pre-Training (2 papers)
    - [2] Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm (Li, 2021) View paper
    - [39] NeuCLIP: Efficient Large-Scale CLIP Training with Neural Normalizer Optimization (Xiyuan Wei, 2025) View paper
    - Speech-Language Pre-Training (1 papers)
    - [45] Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages (Matassoni, 2025) View paper
- System Infrastructure and Distributed Training
    - Parallelism and Distributed Training Systems (1 papers)
    - [5] Efficient large-scale language model training on gpu clusters using megatron-lm (Narayanan, 2021) View paper
    - Hardware Architectures and Memory Optimization (1 papers)
    - [16] 3D-METRO: Deploy Large-Scale Transformer Model on A Chip Using Transistor-Less 3D-Metal-ROM-Based Compute-in-Memory Macro (Yi-Ming Chen, 2025) View paper
    - Workload Analysis and Cost Modeling (2 papers)
    - [17] Characterization of large language model development in the datacenter (Hu Qinghao, 2024) View paper
    - [20] System-performance and cost modeling of Large Language Model training and inference (Guo WenZhe, 2025) View paper
- Privacy-Preserving and Federated Training (3 papers)
    - [23] Pre-text: Training language models on private federated data in the age of llms (Hou, 2024) View paper
    - [24] The future of large language model pre-training is federated (Sani Lorenzo, 2024) View paper
    - [42] MLLM-FL: Multimodal Large Language Model Assisted Federated Learning on Heterogeneous and Long-tailed Data (Zhang Jianyi, 2024) View paper
- Benchmarking and Evaluation Frameworks (2 papers)
    - [13] Datacomp-lm: In search of the next generation of training sets for language models (Li, 2024) View paper
    - [33] LLM-datasets: An open framework for pretraining datasets of large language models (M Ostendorff, 2024) View paper
- Transfer Learning and Model Reuse (1 papers)
    - [50] Meta-learning and synthetic data for automated pretraining and finetuning (Ferreira, 2025) View paper
- Security and Robustness (1 papers)
    - [46] Backdoor samples detection based on perturbation discrepancy consistency in pre-trained language models. (Zuquan Peng, 2025) View paper
- Domain-Specific Applications (1 papers)
    - [19] Large-scale pretraining improves sample efficiency of active learning-based virtual screening (Zhonglin Cao, 2024) View paper

Narrative

Core task: data-efficient language model pre-training under compute abundance. When computational resources are plentiful, the central challenge shifts from simply scaling up to making the most effective use of available data and compute together. The taxonomy reflects this dual focus through several major branches. Compute-Optimal Scaling and Resource Allocation examines how to balance model size, data volume, and training duration—building on foundational work like Chinchilla[1] and extending into newer analyses of generalization trade-offs (Compute Optimal Generalization[49]). Data Curation and Selection Methods addresses the quality and composition of training corpora, spanning retrieval-based approaches (Corpus Aware Retrieval[3]), domain-specific filtering (DoReMi[12], DataComp LM[13]),

and synthetic data generation (Synthetic Data Scaling[14]). Adaptive and Continual Pre-Training explores how models can be updated or specialized over time (Domain Adaptation Pretraining[6]), while Sample-Efficient Training Objectives and Architectures investigates alternative learning signals (ELECTRA[11]) and architectural innovations. Additional branches cover multimodal extensions, distributed infrastructure (Megatron[5], Datacenter LLM Development[17]), privacy-preserving methods (Federated Pre-text[23]), and domain-specific applications, forming a comprehensive landscape of strategies for efficient pre-training.

Several active lines of work reveal key trade-offs and open questions. One thread focuses on scaling law formulation: understanding how loss, model capacity, and data interact under different resource constraints (Chinchilla[1], Compute Optimal Analysis[28], Downstream Scaling Laws[30]). Another emphasizes data quality over sheer volume, with methods that curate, deduplicate (SoftDedup[29]), or synthesize high-value examples (Rephrasing the Web[25], Seed Free Synthetic[35]). A third direction tackles the practical realities of large-scale training, from cost modeling (LLM Cost Modeling[20]) to system-level optimizations (Cerebras GPT[4]). Infinite Compute Pretraining[0] sits squarely within the Compute-Optimal Scaling branch, specifically addressing scaling law formulation and analysis. Its emphasis on scenarios where compute is abundant but data remains finite contrasts with earlier work like Chinchilla[1], which derived optimal ratios under joint resource constraints, and complements recent studies on generalization bounds (Compute Optimal Generalization[49]) by exploring how to allocate unlimited compute when data quality or diversity becomes the bottleneck.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Training compute-optimal large language models

Authors: Hoffmann, Jordan, Borgeaud, Sebastian, Jordan Hoffmann, et al. (66 authors total) | **Year/Venue**: 2022 | **URL**: View paper

Abstract

We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are significantly undertrained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant. By training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens, we find that for compute-optimal training, the model size and...

Relationship Analysis

Both papers belong to the Scaling Law Formulation and Analysis category, investigating optimal relationships between compute, model size, data volume, and performance. The original paper focuses on data-constrained pre-training under infinite compute, deriving scaling laws for parameter count, epoch count, and ensembling while emphasizing asymptotic performance and regularization. The candidate paper (Chinchilla) derives compute-optimal scaling laws showing that model size and training tokens should scale equally, contradicting prior work that suggested scaling model size faster than data, and demonstrates this through training a 70B parameter model on significantly more tokens than standard practice.

---

2. Scaling laws revisited: modeling the role of data quality in language model pretraining

Authors: Subramanyam, Anirudh, Chen, Yuxin, Anirudh Subramanyam, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Scaling laws for language model training traditionally characterize how performance scales with model size and dataset volume. Prior work has explored architecture variants and data treatments such as dataset filtering and noise injection in language model pretraining; however, these studies have not formalized data quality within a principled scaling law. We introduce a dimensionless data-quality parameter Q, and propose a quality-aware scaling law extending the Chinchilla framework to predict ...

Relationship Analysis

Both papers belong to the Scaling Law Formulation and Analysis category, focusing on deriving scaling laws that relate compute, model size, data volume, and performance under data-constrained conditions. They overlap in addressing how to optimize language model pre-training when data is limited relative to compute, with both proposing extensions to standard Chinchilla-style scaling laws. The key difference is that the original paper focuses on algorithmic interventions (regularization, ensembling, distillation) to achieve data efficiency under infinite compute, while the candidate paper introduces an explicit data quality parameter Q into the scaling law formulation to model how data corruption and deficiency affect loss.

---

3. An empirical analysis of compute-optimal large language model training

Authors: J Hoffmann, S Borgeaud, A Mensch | **Year/Venue**: 2022 | **URL**: View paper

Abstract

â¦ While pre-training a large language model has a considerable compute cost, downstream â¦ of the pre-training data alone can lead to substantial improvements on this benchmark. â¦

Relationship Analysis

Both papers belong to the Scaling Law Formulation and Analysis category, investigating the relationship between compute, model size, data volume, and performance in language model pre-training. While the original paper focuses on data-efficient pre-training under infinite compute by exploring regularization, ensembling, and asymptotic scaling laws when data is fixed, the candidate paper (Chinchilla) empirically derives compute-optimal scaling laws by training over 400 models to determine the optimal balance between model size and training tokens for a given compute budget. The key difference is that the original paper assumes unlimited compute with fixed data constraints and evaluates recipes by their loss asymptotes, whereas the candidate paper operates under fixed compute budgets and proposes equal scaling of parameters and tokens, challenging prior work that recommended faster parameter scaling.

---

4. Scaling Laws for Predicting Downstream Performance in LLMs

Authors: Chen Yangyi, Huang, Binxuan, Yangyi Chen, Gao Yifan, et al. (14 authors total) | **Year/Venue**: 2024 • Trans. Mach. Learn. Res. | **URL**: View paper

Abstract

Precise estimation of downstream performance in large language models (LLMs) prior to training is essential for guiding their development process. Scaling laws analysis utilizes the statistics of a series of significantly smaller sampling language models (LMs) to predict the performance of the target LLM. For downstream performance prediction, the critical challenge lies in the emergent abilities in LLMs that occur beyond task-specific computational thresholds. In this work, we focus on the pre-...

Relationship Analysis

Both papers belong to the Scaling Law Formulation and Analysis category, focusing on deriving and extending scaling laws for language model pre-training. They overlap in their use of power law relationships to predict model performance based on computational resources and model size. However, the original paper focuses on data-constrained pre-training under infinite compute, exploring regularization, ensembling, and asymptotic performance limits, while the candidate paper focuses on predicting downstream task

performance from pre-training loss across different model sizes and data mixtures, particularly addressing emergent abilities and multi-source datasets.

---

5. Compute-Optimal LLMs Provably Generalize Better With Scale

Authors: Finzi, Marc, Kapoor, Sanyam, Marc Finzi, et al. (18 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Why do larger language models generalize better? To investigate this question, we develop generalization bounds on the pretraining objective of large language models (LLMs) in the compute-optimal regime, as described by the Chinchilla scaling laws. We introduce a novel, fully empirical Freedman-type martingale concentration inequality that tightens existing bounds by accounting for the variance of the loss function. This generalization bound can be decomposed into three interpretable components:...

Relationship Analysis

Both papers belong to the Scaling Law Formulation and Analysis category, investigating how compute, model size, data volume, and performance relate in language model pre-training. While the original paper focuses on data-constrained pre-training under infinite compute, deriving empirical scaling laws for parameter count, epoch count, and ensembling with asymptotic analysis, the candidate paper provides theoretical generalization bounds for compute-optimal (Chinchilla) scaling, proving that larger models generalize better through decreased loss variance and improved quantizability. The key difference is that the original paper addresses data scarcity with algorithmic interventions (regularization, ensembling, distillation), whereas the candidate paper explains why compute-optimal scaling works from a generalization theory perspective.

Contributions Analysis

Overall novelty summary. The paper proposes a data-efficient pre-training framework combining regularization, parameter scaling, and ensemble methods to optimize performance under fixed data budgets. It resides in the 'Scaling Law Formulation and Analysis' leaf, which contains six papers including foundational work like Chinchilla and recent extensions examining generalization trade-offs. This leaf sits within the broader 'Compute-Optimal Scaling and Resource Allocation' branch, indicating a moderately populated research direction focused on theoretical scaling principles rather than empirical recipes or system implementations.

The taxonomy reveals neighboring work in 'Empirical Training Strategies and Recipes' (three papers on practical protocols) and 'Data Curation and Selection Methods' (multiple leaves addressing quality filtering, diversity optimization, and synthetic generation). The paper's focus on asymptotic scaling laws under compute abundance distinguishes it from sibling papers examining joint compute-data constraints (Chinchilla) or generalization bounds. The scope_note clarifies this leaf excludes empirical recipes without theoretical analysis, positioning the work as extending scaling law theory rather than proposing purely practical training configurations.

Among 22 candidates examined across three contributions, no clearly refuting prior work was identified. The regularized parameter scaling contribution examined 10 candidates with zero refutations, suggesting novelty in the specific combination of tuned weight decay and asymptote-based evaluation. The ensemble scaling framework similarly examined 10 candidates without refutation, indicating the asymptote-focused methodology may be distinctive. The joint scaling recipe examined only 2 candidates, reflecting a more limited search scope for this compositional contribution. These statistics indicate the analysis covered a focused set of semantically related papers rather than an exhaustive field survey.

The limited search scope (22 candidates from semantic search) means the analysis captures closely related scaling law research but may not cover all relevant empirical training studies or data curation methods. The absence of refuting papers among examined candidates suggests the specific combination of techniques—particularly the asymptote-based evaluation framework and ensemble scaling under data constraints—appears novel within the sampled literature, though broader coverage might reveal additional overlaps in adjacent research directions.

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

Contribution 1: Regularized parameter scaling recipe with tuned weight decay

Description: The authors introduce a regularized pre-training recipe that jointly tunes weight decay, learning rate, and epoch count at each parameter count. This approach achieves monotonic loss decrease following a power law in parameter count, with optimal weight decay being 30 times larger than the standard 0.1 value used in practice.

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. How to set AdamW's weight decay as you scale model and dataset size

URL: View paper

Brief Assessment

AdamW Weight Decay[64] focuses on scaling weight decay with model/dataset size through an EMA timescale framework, while the original paper addresses data-constrained pre-training with joint tuning of weight decay, learning rate, and epochs to prevent overfitting. These are distinct optimization contexts with different objectives.

---

2. SGD with weight decay secretly minimizes the ranks of your neural networks

URL: View paper

Brief Assessment

Weight Decay Rank[67] focuses on how weight decay affects the rank of weight matrices during training, not on parameter scaling recipes or data-constrained pre-training optimization. The candidate examines implicit bias toward low-rank matrices, while the original develops a regularized recipe for data-efficient pre-training with specific epoch and parameter count tuning.

---

3. Rethinking weight decay for robust fine-tuning of foundation models

URL: View paper

Brief Assessment

Robust Weight Decay[65] focuses on selective weight decay for fine-tuning foundation models, not pre-training recipes with parameter scaling laws. The contexts are fundamentally different.

---

4. Weight decay induces low-rank attention layers

URL: View paper

Brief Assessment

Low Rank Attention[66] focuses on the theoretical and empirical effects of weight decay on inducing low-rank structures in attention layers of transformers, not on parameter scaling recipes for pre-training or optimal weight decay values for data-constrained scenarios.

---

5. Rotational equilibrium: How weight decay balances learning across neural networks

URL: View paper

Brief Assessment

Rotational Equilibrium[69] focuses on how weight decay affects individual neuron update dynamics and rotational equilibrium in neural networks, not on parameter scaling recipes for data-constrained language model pre-training or achieving monotonic loss decrease through joint tuning of weight decay, learning rate, and epoch count at different parameter scales.

---

6. Hallmarks of Optimization Trajectories in Neural Networks: Directional Exploration and Redundancy

URL: View paper

Brief Assessment

Optimization Trajectories[70] focuses on analyzing directional structure and redundancy in optimization trajectories through trajectory complexity measures, not on weight decay tuning for parameter scaling recipes in data-constrained pre-training.

---

7. Rank minimization, alignment and weight decay in neural networks

URL: View paper

Brief Assessment

Rank Minimization Alignment[63] focuses on the spectral dynamics of weight matrices (singular values/vectors) and how weight decay affects rank minimization across architectures. It does not address parameter scaling recipes, epoch tuning, or power law scaling relationships that are central to the original paper's contribution.

---

8. Weight Decay With Tailored Adam on Scale-Invariant Weights for Better Generalization

URL: View paper

Brief Assessment

Tailored Adam Decay[72] focuses on modifying the Adam optimizer to work better with weight decay for image classification tasks, not on scaling recipes for language model pre-training with jointly tuned hyperparameters across parameter counts.

---

9. Understanding decoupled and early weight decay

URL: View paper

Brief Assessment

Decoupled Weight Decay[71] focuses on understanding why decoupled weight decay works better than L2 regularization with adaptive optimizers, not on parameter scaling recipes or optimal weight decay values for data-constrained pre-training.

---

10. Explicit regularisation, sharpness and calibration

URL: View paper

Brief Assessment

Explicit Regularisation Sharpness[68] focuses on the relationship between weight decay, loss landscape sharpness, and calibration in image classification tasks, not on parameter scaling recipes for language model pre-training or power law scaling behavior.

Contribution 2: Ensemble scaling recipe and asymptote-based evaluation framework

Description: The authors propose an ensembling recipe that trains multiple independent models and averages their logits, achieving a lower loss asymptote than parameter scaling alone. They introduce evaluating scaling recipes by the asymptote of their scaling law rather than performance at fixed compute budgets.

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. When Ensembling Smaller Models is More Efficient than Single Large Models

URL: View paper

Brief Assessment

Smaller Model Ensembles[60] focuses on computational efficiency trade-offs (accuracy vs. FLOPs) for image classification, not on data-constrained pre-training or asymptote-based scaling law evaluation under infinite compute.

---

2. Using Bayesian model averaging to calibrate forecast ensembles

URL: View paper

Brief Assessment

Bayesian Model Averaging[54] focuses on postprocessing weather forecast ensembles to calibrate probabilistic predictions, not on training multiple independent models to achieve lower loss asymptotes in language model pre-training.

---

3. Ensemble learning using decorrelated neural networks

URL: View paper

Brief Assessment

Decorrelated Ensembles[53] focuses on decorrelation training methods for small neural network ensembles in regression tasks, not on scaling laws or asymptote-based evaluation frameworks for language model pre-training under data constraints.

---

4. Ensemble-learning approaches for network security and anomaly detection

URL: View paper

Brief Assessment

Network Security Ensemble[57] focuses on ensemble learning for network security and anomaly detection tasks, not language model pre-training or scaling laws. The domains and objectives are fundamentally different.

---

5. Counting the Cost: Quantifying the Rising Impacts of Heat-Related Productivity Losses in the United States (2001â□□2023)

URL: View paper

Brief Assessment

Heat Productivity Losses[59] focuses on quantifying economic impacts of heat-related productivity losses in the United States, not on machine learning ensemble methods or scaling laws for language models.

---

6. Snapshot Ensemble One-Dimensional Convolutional Neural Networks for Ballistic Target Recognition

URL: View paper

Brief Assessment

Snapshot Ensemble CNN[56] focuses on ballistic target recognition using snapshot ensembles with cosine annealing, not on language model pre-training or asymptote-based scaling law evaluation.

---

7. Deep Learning Ensemble Method for Classifying Glaucoma Stages Using Fundus Photographs and Convolutional Neural Networks

URL: View paper

Brief Assessment

Glaucoma Ensemble[52] focuses on medical image classification using CNN ensembles for glaucoma diagnosis, not language model pre-training or scaling law asymptotes. The domains and methodologies are fundamentally different.

---

8. Boost Neural Networks by Checkpoints

URL: View paper

Brief Assessment

Checkpoint Boosting[58] focuses on ensembling checkpoints from a single training run to reduce computational cost, not on training multiple independent models to achieve lower loss asymptotes than parameter scaling.

---

9. Prune and tune ensembles: low-cost ensemble learning with sparse independent subnetworks

URL: View paper

Brief Assessment

Prune and Tune[51] focuses on creating ensembles through pruning and fine-tuning child networks from a single parent, not on scaling laws or asymptotic performance evaluation under infinite compute constraints.

---

10. Hybrid and Ensemble Methods of Two Days Ahead Forecasts of Electric Energy Production in a Small Wind Turbine

URL: View paper

Brief Assessment

Wind Turbine Ensemble[55] focuses on ensemble methods for wind turbine energy forecasting, not language model pre-training or scaling laws. The ensembling approach is domain-specific to renewable energy prediction rather than addressing loss asymptotes in neural language models.

Contribution 3: Joint scaling recipe composing parameter and ensemble scaling

Description: The authors develop a joint scaling recipe that composes both parameter scaling and ensemble scaling by taking the double limit as both parameter count and ensemble member count approach infinity. This combined approach achieves significantly improved data efficiency compared to standard pre-training recipes.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Simultaneous Learning the Dimension and Parameter of a Statistical Model with Big Data

URL: View paper

Brief Assessment

Dimension Parameter Learning[62] is not available for comparison. The candidate paper's full text context is marked as 'n/a', making it impossible to assess whether it addresses joint parameter and ensemble scaling for neural networks or any related methodology.

---

2. The surprising ineffectiveness of pre-trained visual representations for model-based reinforcement learning

URL: View paper

Brief Assessment

Visual Pretraining MBRL[61] focuses on pre-trained visual representations for model-based reinforcement learning in robotics/control tasks, not on joint parameter and ensemble scaling for language model pre-training.

Appendix: Text Similarity Detection

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

References

- [0] Pre-training under infinite compute View paper
- [1] Training compute-optimal large language models View paper
- [2] Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm View paper
- [3] Unsupervised corpus aware language model pre-training for dense passage retrieval View paper
- [4] Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster View paper
- [5] Efficient large-scale language model training on gpu clusters using megatron-lm View paper
- [6] Don't stop pretraining: Adapt language models to domains and tasks View paper
- [7] Improving pretraining data using perplexity correlations View paper
- [8] Language models improve when pretraining data matches target tasks View paper
- [9] Data engineering for scaling language models to 128k context View paper
- [10] How to train data-efficient llms View paper
- [11] Electra: Pre-training text encoders as discriminators rather than generators View paper
- [12] Doremi: Optimizing data mixtures speeds up language model pretraining View paper
- [13] Datacomp-lm: In search of the next generation of training sets for language models View paper
- [14] Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls View paper
- [15] Cramming: Training a Language Model on a single GPU in one day. View paper
- [16] 3D-METRO: Deploy Large-Scale Transformer Model on A Chip Using Transistor-Less 3D-Metal-ROM-Based Compute-in-Memory Macro View paper
- [17] Characterization of large language model development in the datacenter View paper

- [18] Scaling laws revisited: modeling the role of data quality in language model pretraining View paper
- [19] Large-scale pretraining improves sample efficiency of active learning-based virtual screening View paper
- [20] System-performance and cost modeling of Large Language Model training and inference View paper
- [21] Entropy law: The story behind data compression and llm performance View paper
- [22] Combatting dimensional collapse in llm pre-training data via diversified file selection View paper
- [23] Pre-text: Training language models on private federated data in the age of llms View paper
- [24] The future of large language model pre-training is federated View paper
- [25] Rephrasing the web: A recipe for compute and data-efficient language modeling View paper
- [26] Dataman: Data manager for pre-training large language models View paper
- [27] Regmix: Data mixture as regression for language model pre-training View paper
- [28] An empirical analysis of compute-optimal large language model training View paper
- [29] Softdedup: an efficient data reweighting method for speeding up language model pre-training View paper
- [30] Scaling Laws for Predicting Downstream Performance in LLMs View paper
- [31] Nuner: Entity recognition encoder pre-training via llm-annotated data View paper
- [32] Stacking your transformers: A closer look at model growth for efficient llm pre-training View paper
- [33] LLM-datasets: An open framework for pretraining datasets of large language models View paper
- [34] Multi-agent collaborative data selection for efficient llm pretraining View paper
- [35] Seed-Free Synthetic Data Generation Framework for Instruction-Tuning LLMs: A Case Study in Thai View paper
- [36] Continual Pre-training of MoEs: How robust is your router? View paper
- [37] Cpm-2: Large-scale cost-effective pre-trained language models View paper
- [38] Enhancing Multilingual LLM Pretraining with Model-Based Data Selection View paper
- [39] NeuCLIP: Efficient Large-Scale CLIP Training with Neural Normalizer Optimization View paper
- [40] Cost-Efficient Domain-Adaptive Pretraining of Language Models for Optoelectronics Applications. View paper
- [41] Judging Quality Across Languages: A Multilingual Approach to Pretraining Data Filtering with Language Models View paper
- [42] MLLM-FL: Multimodal Large Language Model Assisted Federated Learning on Heterogeneous and Long-tailed Data View paper
- [43] DataDecide: How to Predict Best Pretraining Data with Small Experiments View paper
- [44] AttentionInfluence: Adopting Attention Head Influence for Weak-to-Strong Pretraining Data Selection View paper
- [45] Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages View paper
- [46] Backdoor samples detection based on perturbation discrepancy consistency in pre-trained language models. View paper
- [47] A Minimalist Optimizer Design for LLM Pretraining View paper
- [48] TiC-LM: A Web-Scale Benchmark for Time-Continual LLM Pretraining View paper
- [49] Compute-Optimal LLMs Provably Generalize Better With Scale View paper
- [50] Meta-learning and synthetic data for automated pretraining and finetuning View paper
- [51] Prune and tune ensembles: low-cost ensemble learning with sparse independent subnetworks View paper
- [52] Deep Learning Ensemble Method for Classifying Glaucoma Stages Using Fundus Photographs and Convolutional Neural Networks View paper
- [53] Ensemble learning using decorrelated neural networks View paper
- [54] Using Bayesian model averaging to calibrate forecast ensembles View paper
- [55] Hybrid and Ensemble Methods of Two Days Ahead Forecasts of Electric Energy Production in a Small Wind Turbine View paper
- [56] Snapshot Ensemble One-Dimensional Convolutional Neural Networks for Ballistic Target Recognition View paper
- [57] Ensemble-learning approaches for network security and anomaly detection View paper
- [58] Boost Neural Networks by Checkpoints View paper
- [59] Counting the Cost: Quantifying the Rising Impacts of Heat-Related Productivity Losses in the United States (2001â□□2023) View paper
- [60] When Ensembling Smaller Models is More Efficient than Single Large Models View paper
- [61] The surprising ineffectiveness of pre-trained visual representations for model-based reinforcement learning View paper
- [62] Simultaneous Learning the Dimension and Parameter of a Statistical Model with Big Data View paper
- [63] Rank minimization, alignment and weight decay in neural networks View paper
- [64] How to set AdamW's weight decay as you scale model and dataset size View paper
- [65] Rethinking weight decay for robust fine-tuning of foundation models View paper
- [66] Weight decay induces low-rank attention layers View paper
- [67] SGD with weight decay secretly minimizes the ranks of your neural networks View paper
- [68] Explicit regularisation, sharpness and calibration View paper
- [69] Rotational equilibrium: How weight decay balances learning across neural networks View paper
- [70] Hallmarks of Optimization Trajectories in Neural Networks: Directional Exploration and Redundancy View paper
- [71] Understanding decoupled and early weight decay View paper
- [72] Weight Decay With Tailored Adam on Scale-Invariant Weights for Better Generalization View paper