Novelty Assessment Report

Paper: Can Small Training Runs Reliably Guide Data Curation? Rethinking Proxy-Model Practice

PDF URL: https://openreview.net/pdf?id=2FZC0c06jP

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Data teams at frontier AI companies routinely train small proxy models to make critical decisions about pretraining data recipes for full-scale training. However, the community has a limited understanding of whether and when conclusions drawn from small-scale experiments reliably transfer to large-scale production training. In this work, we uncover a critical issue in the standard practice of training small proxy models on each data recipe with a single set of hyperparameters. We demonstrate that each dataset requires its own optimal training configuration, and that dataset rankings can completely reverse with even minor adjustments to proxy training hyperparameters. Furthermore, this creates a disconnect from the actual model development pipeline, where hyperparameter optimization is a standard step. Consequently, we propose that the objective of data selection should be to identify the dataset that yields the best performance after its own hyperparameter optimization. We introduce a simple yet effective patch to the current proxy-model-based method: training proxy models with sufficiently small learning rates produces dataset rankings that strongly correlate with those obtained when large-scale models are properly tuned for each dataset. Theoretically, we prove that, for random-feature models, this approach preserves the ordering of datasets according to their optimal achievable losses. Empirically, we validate this approach through comprehensive experiments across 23 data recipes covering four critical dimensions of data curation decisions faced in production settings, demonstrating dramatic improvements in proxy model reliability.

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: **data curation using small proxy models for large-scale training**

A total of **43 papers** were analyzed and organized into a taxonomy with **29 categories**.

Taxonomy Overview

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

- **Proxy Model Design and Training Strategies**
- **Data Selection Objectives and Optimization**
- **Task-Specific Data Selection Applications**
- **Efficiency and Scalability Techniques**
- **Alternative Proxy Model Applications**
- **Domain-Specific Proxy Applications**
- **Auxiliary Topics and Methodological Foundations**

Complete Taxonomy Tree

- data curation using small proxy models for large-scale training Survey Taxonomy
- Proxy Model Design and Training Strategies
  - Hyperparameter Optimization for Proxy Models ★ (1 papers)
  - [0] Can Small Training Runs Reliably Guide Data Curation? Rethinking Proxy-Model Practice (Anon et al., 2026) View paper
  - Proxy Model Architecture and Scaling (3 papers)
  - [8] Selection via proxy: Efficient data selection for deep learning (Coleman, 2019) View paper
  - [20] Small-to-Large Generalization: Data Influences Models Consistently Across Scale (Khaddaj, 2025) View paper
  - [36] Small-to-Large Generalization: Training Data Influences Models Consistently Across Scale (A Khaddaj, n.d.) View paper
  - Training Trajectory and Learning Dynamics (3 papers)
  - [1] Smaller language models are capable of selecting instruction-tuning training data for larger language models (Mekala, 2024) View paper
  - [5] Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models (Yang Yu, 2024) View paper
  - [6] Bad students make great teachers: Active learning accelerates large-scale visual understanding (Evans, 2024) View paper
- Data Selection Objectives and Optimization
  - Regression-Based Data Mixture Optimization (1 papers)
  - [3] Regmix: Data mixture as regression for language model pre-training (Liu Qian, 2024) View paper
  - Clustering-Based Data Mixture Discovery (1 papers)
  - [4] Nemotron-CLIMB: CLustering-based Iterative Data Mixture Bootstrapping for Language Model Pre-training (Shizhe Diao, 2025) View paper
  - Reinforcement Learning for Data Selection (2 papers)
  - [9] RL-Guided Data Selection for Language Model Finetuning (Jha Animesh, 2025) View paper
  - [24] Proxy-RLHF: Decoupling Generation and Alignment in Large Language Model with Proxy (Zhu Yu, 2024) View paper
  - Compute-Constrained and Cost-Aware Selection (1 papers)
  - [21] Compute-Constrained Data Selection (Rush, 2024) View paper

- Task-Specific Data Selection Applications
  - Pretraining Data Selection (4 papers)
  - [2] Smaller Can Be Better: Efficient Data Selection for Pre-training Models (Guang Fang, 2024) View paper
  - [19] BLISS: A Lightweight Bilevel Influence Scoring Method for Data Selection in Language Model Pretraining (Hao Jie, 2025) View paper
  - [29] Foundation Models from a Data-Distributional View (Xie, 2024) View paper
  - [42] Spaced Scheduling for Large Language Model Training (N Chapados, n.d.) View paper
  - Instruction-Tuning and Supervised Fine-Tuning (3 papers)
  - [13] Small Language Model as Data Prospector for Large Language Model (Ni, 2024) View paper
  - [26] Tuning Language Models by Proxy (Liu, 2024) View paper
  - [34] Improving Large-Scale Conversational Assistants using Model Interpretation based Training Sample Selection (Stefan Schroedl, 2022) View paper
  - Vision-Language Model Data Selection (1 papers)
  - [7] Concept-skill Transferability-based Data Selection for Large Vision-Language Models (Lee Jae-Woo, 2024) View paper
  - Multimodal Model Distillation and Curation (2 papers)
  - [22] Active Data Curation Effectively Distills Large-Scale Multimodal Models (Vishaal Udandarao, 2024) View paper
  - [37] FAMIE: A Fast Active Learning Framework for Multilingual Information Extraction (Minh Van Nguyen, 2022) View paper
  - Domain-Specific and Medical Data Curation (2 papers)
  - [23] RefineNet: Elevating Medical Foundation Models Through Quality-Centric Data Curation by MLLM-Annotated Proxy Distillation (Ningyi Zhang, 2025) View paper
  - [38] SEED: Domain-Specific Data Curation With Large Language Models (Chen, 2023) View paper
- Efficiency and Scalability Techniques
  - Pruning-Based Efficiency Methods (2 papers)
  - [11] Pruning-based Data Selection and Network Fusion for Efficient Deep Learning (Bhatti, 2025) View paper
  - [41] PruneFuse: Efficient Data Selection via Weight Pruning and Network Fusion (H Kousar, n.d.) View paper
  - On-Device and Parameter-Efficient Training (1 papers)
  - [10] PieBridge: Fast and Parameter-Efficient On-Device Training via Proxy Networks (Wangsong Yin, 2024) View paper
  - Batch Selection and Online Mixing (2 papers)
  - [16] Research on Fine-Tuning Optimization of Large Language Models Based on Online Data Mixing (Y Liu, 2025) View paper
  - [43] Soft Sampling for Efficient Training of Deep Neural Networks on Massive Data (X Cui, n.d.) View paper
- Alternative Proxy Model Applications
  - Transfer Learning and Neural Data Servers (1 papers)
  - [12] Neural data server: A large-scale search engine for transfer learning data (Xi Yan, 2020) View paper
  - Fairness and Bias Mitigation (1 papers)
  - [28] Navigating Towards Fairness with Data Selection (Yixuan Zhang, 2024) View paper
  - Federated Learning Client Selection (1 papers)
  - [14] Proxy Model-Guided Reinforcement Learning for Client Selection in Federated Recommendation (Qu Liang, 2025) View paper
  - Neural Architecture Search (1 papers)
  - [15] Less is more: Proxy datasets in nas approaches (Brian Moser, 2022) View paper
- Domain-Specific Proxy Applications
  - Process Systems and Physics-Aware Modeling (1 papers)
  - [18] PAPM: A Physics-aware Proxy Model for Process Systems (Liu Peng-wei, 2024) View paper
  - Geoscience and Reservoir Engineering (1 papers)
  - [30] Small-Scale or Large-Scale Machine Learning Proxy Models? a Real Field Case Study (R. Amiri Kolajoobi, 2024) View paper
  - Language-Specific Data Filtering (2 papers)
  - [17] Filtering noisy parallel corpus using transformers with proxy task learning (Haluk AÃ§arÃ§iÃ§ek, 2020) View paper
  - [35] Cynical Selection of Language Model Training Data (Axelrod, 2022) View paper
- Auxiliary Topics and Methodological Foundations
  - Surveys and Comparative Studies (1 papers)
  - [27] A Comparative Survey: Reusing Small Pre-Trained Models for Efficient Large Model Training (Dhroov Pandey, 2024) View paper
  - Data Synthesis and Augmentation (1 papers)
  - [32] JiuZhang3.0: Efficiently Improving Mathematical Reasoning by Training Small Data Synthesis Models (Zhou, 2024) View paper
  - Model Confusion and Robustness Analysis (1 papers)
  - [31] Confusing Large Models by Confusing Small Models (Vitor Albiero, 2023) View paper
  - Privacy-Preserving Data Selection (1 papers)
  - [33] SelectFormer: Private and Practical Data Selection for Transformers (Ouyang Xu, 2023) View paper
  - Self-Supervised Proxy Tasks (1 papers)
  - [40] Colorization as a Proxy Task for Visual Understanding (Gustav Larsson, 2017) View paper
  - Complex Task Decomposition (1 papers)
  - [25] Generating Complex Question Decompositions in the Face of Distribution Shifts (Kelvin Han, 2025) View paper
  - Data-Centric Training Paradigms (1 papers)
  - [39] Smollm2: When smol goes bigâ□□data-centric training of a fully open small language model (A Lozhkov, n.d.) View paper

Narrative

Core task: data curation using small proxy models for large-scale training. The field addresses the challenge of selecting high-quality training data efficiently by leveraging smaller, computationally cheaper models to guide decisions for much larger target models. The taxonomy reveals several main branches: Proxy Model Design and Training Strategies explores how to construct and optimize these small surrogates, including hyperparameter tuning and architectural choices; Data Selection Objectives and Optimization focuses on the scoring functions and optimization frameworks that determine which examples to retain; Task-Specific Data Selection Applications examines domain-tailored approaches for language modeling, instruction tuning, and specialized tasks; Efficiency and Scalability Techniques investigates methods to handle massive datasets through batching, pruning, and online mixing; while Alternative and Domain-Specific Proxy Applications consider broader uses of proxy models beyond standard data curation. Works such as Selection via

Proxy[8] and Smaller Select Data[1] illustrate foundational principles, whereas recent efforts like Nemotron CLIMB[4] and BLISS[19] demonstrate scalable implementations across diverse settings.

A particularly active line of inquiry concerns the transferability of proxy model judgments to larger targets: studies like Small to Large[5] and Concept Skill Transferability[7] investigate when and why small models reliably predict large-model performance, while Bad Students Great Teachers[6] highlights scenarios where weaker proxies can paradoxically yield strong curation outcomes. Another contrasting theme involves online versus offline selection strategies, with Online Data Mixing[16] and RL Guided Selection[9] exploring adaptive, feedback-driven approaches. Proxy Model Practice[0] sits within the Hyperparameter Optimization for Proxy Models cluster, emphasizing systematic tuning of proxy configurations to maximize downstream gains. Compared to nearby works such as Regmix[3], which blends multiple proxy signals, or Smaller Can Be Better[2], which examines minimal proxy architectures, Proxy Model Practice[0] focuses specifically on the hyperparameter landscape that governs proxy fidelity and computational trade-offs, offering practitioners guidance on calibrating these small surrogates for diverse large-scale training regimes.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses specifically on hyperparameter optimization techniques (tuning, learning rate selection) to ensure proxy model reliability in data curation contexts. The sibling subtopics address complementary aspects: one examines architectural and scaling decisions for proxy models, while the other leverages dynamic training signals like trajectories and learnability. Together, these subtopics span the design, optimization, and utilization phases of small proxy models for large-scale data curation.

Similarities: - All three subtopics operate within the paradigm of using small proxy models to inform data curation for large-scale training - Each addresses reliability and effectiveness of proxy models, whether through hyperparameter tuning, architecture selection, or trajectory-based signals - All exclude applications of proxy models outside data curation contexts

Differences: - The original leaf targets optimization of training parameters (hyperparameters, learning rates), while 'Proxy Model Architecture and Scaling' focuses on structural design choices (model size, architecture) - 'Training Trajectory and Learning Dynamics' emphasizes temporal signals during training (trajectories, learnability over time), whereas the original leaf concerns static configuration decisions made before or during training - The original leaf explicitly excludes general training dynamics without hyperparameter focus, which is the core domain of 'Training Trajectory and Learning Dynamics' - 'Proxy Model Architecture and Scaling' addresses pre-training decisions about model capacity, while the original leaf addresses tuning decisions that occur during the training process

Suggested Search Directions: - Investigate interactions between hyperparameter choices and architecture selection for proxy models - Explore whether training trajectory signals can inform adaptive hyperparameter optimization strategies - Examine joint optimization frameworks that simultaneously address architecture, hyperparameters, and trajectory-based curation

Sibling Subtopics

- **Proxy Model Architecture and Scaling** (leaves: 1, papers: 3)
- Scope: Studies on proxy model size selection, architecture choices, and scaling properties for data curation.
- Exclude: Excludes methods using proxy models for non-curation tasks; covered under alternative proxy applications.
- **Training Trajectory and Learning Dynamics** (leaves: 1, papers: 3)
- Scope: Methods leveraging training trajectories, learning percentages, or learnability signals from small models.
- Exclude: Excludes static data selection without trajectory analysis; covered under static scoring methods.

Contributions Analysis

Overall novelty summary. The paper identifies a critical flaw in standard proxy-model-based data selection: dataset rankings can reverse when proxy training hyperparameters change, and proposes using sufficiently small learning rates to stabilize these rankings. Within the taxonomy, it occupies the 'Hyperparameter Optimization for Proxy Models' leaf under 'Proxy Model Design and Training Strategies'. Notably, this leaf contains only the original paper itself—no sibling papers—indicating this is a relatively sparse research direction. The broader parent category includes three leaves addressing proxy architecture, scaling, and training dynamics, suggesting the field has explored proxy model design from multiple angles but has not deeply investigated hyperparameter sensitivity until now.

The taxonomy reveals neighboring work in 'Proxy Model Architecture and Scaling' (three papers on model size selection) and 'Training Trajectory and Learning Dynamics' (three papers on learnability signals). These adjacent leaves focus on what proxy models to build and how to interpret their training signals, whereas the original paper addresses how to configure proxy training itself. The 'Data Selection Objectives and Optimization' branch (four leaves, seven papers total) explores scoring functions and mixture optimization but does not examine the hyperparameter configurations that produce those scores. This positioning suggests the paper bridges a gap between proxy model construction and optimization frameworks by questioning the reliability of the intermediate training step.

Among seventeen candidates examined, none clearly refute any of the three contributions. The first contribution (hyperparameter sensitivity identification) examined one candidate with no refutation. The second (tiny learning rate strategy) and third (theoretical proof with empirical validation) each examined eight candidates, again with no refutations found. This limited search scope—seventeen papers from semantic search and citation expansion—means the analysis captures nearby work but cannot claim exhaustive coverage. The absence of refutations among these candidates suggests the specific focus on hyperparameter-induced ranking reversals and the tiny learning rate remedy may be novel within the examined literature, though a broader search could reveal additional relevant prior work.

Given the sparse occupancy of the hyperparameter optimization leaf and the lack of refutations among seventeen examined candidates, the work appears to address an underexplored aspect of proxy-based data curation. However, the limited search scope and the paper's position in a relatively new subfield mean this assessment reflects current visibility rather than definitive novelty. The taxonomy structure indicates active research in related proxy model design areas, suggesting the community may soon expand attention to hyperparameter robustness as proxy methods mature.

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

Contribution 1: Identification of hyperparameter sensitivity in proxy-model-based dataset selection

Description: The authors reveal that standard proxy-model practices are fragile because dataset rankings can flip when training hyperparameters (especially learning rate) are slightly adjusted. This exposes a critical disconnect between fixed-hyperparameter evaluation and real-world workflows where hyperparameters are tuned per dataset.

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. Scalable gaussian process-based transfer surrogates for hyperparameter optimization

URL: View paper

Brief Assessment

Gaussian Process Surrogates[52] focuses on hyperparameter optimization for machine learning algorithms (e.g., SVMs, neural networks) rather than dataset selection for pretraining. The paper addresses transferring hyperparameter performance across different datasets, not the problem of ranking datasets based on their quality for model pretraining.

Contribution 2: Tiny learning rate strategy for reliable proxy model training

Description: The authors propose training proxy models with very small learning rates (e.g., 10^-5 to 10^-6) to improve transferability. This approach yields dataset rankings that remain consistent when models are scaled up and hyperparameters are optimized, addressing the fragility identified in current practices.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Layer-Wise Learning Rate Optimization for Task-Dependent Fine-Tuning of Pre-Trained Models: An Evolutionary Approach

URL: View paper

Brief Assessment

Layer Wise Learning Rate[48] focuses on evolutionary optimization of layer-wise learning rates for fine-tuning pre-trained models on specific NLP tasks, not on proxy model training for data curation or dataset selection with tiny learning rates.

2. Meta-learning innovates chemical kinetics: An efficient approach for surrogate model construction

URL: View paper

Brief Assessment

Kinetics Surrogate Construction[45] focuses on meta-learning for chemical kinetics surrogate models with fixed inner/meta step sizes (0.02), not on proxy model training for data curation with tiny learning rates (10^-5 to 10^-6) to improve cross-scale transferability.

3. pFedES: Generalized Proxy Feature Extractor Sharing for Model Heterogeneous Personalized Federated Learning

URL: View paper

Brief Assessment

pFedES[46] focuses on federated learning with heterogeneous models using proxy feature extractors for knowledge sharing across clients, not on proxy model training strategies for data curation or learning rate optimization for transferability.

4. A Study of Generalization of Stochastic Mirror Descent Algorithms on Overparameterized Nonlinear Models

URL: View paper

Brief Assessment

Stochastic Mirror Descent[51] focuses on convergence and generalization properties of SMD algorithms in overparameterized nonlinear models, not on proxy model training strategies for data curation or learning rate selection for dataset ranking transferability.

5. Quantifying the uncertainty of structural parameters using machine learningâ□□based surrogate models

URL: View paper

Brief Assessment

Structural Parameters Uncertainty[47] focuses on surrogate models for structural parameter uncertainty quantification in engineering contexts, not on data curation or proxy model transferability for language model pretraining. The candidate's mention of learning rate is in a different technical context.

6. Classical feature map surrogates and metrics for quantum control landscapes

URL: View paper

Brief Assessment

Classical Feature Surrogates[50] focuses on quantum control landscapes and feature map representations for quantum systems, not on proxy model training strategies for data curation in language models. The domains are fundamentally different.

7. Step-by-step enhancement of a graph neural network-based surrogate model for Lagrangian fluid simulations with flexible time step sizes

URL: View paper

Brief Assessment

GNN Surrogate Lagrangian[44] focuses on surrogate models for Lagrangian fluid simulations with flexible time steps, not on proxy model training strategies for data curation or learning rate optimization for dataset selection transferability.

8. An Empirical Study of P Learning Rate Transfer

URL: View paper

Brief Assessment

Learning Rate Transfer[49] focuses on μ-parameterization (μP) for scaling initializations and learning rates across model widths in transformers, not on using tiny learning rates to improve proxy model transferability for data curation decisions as proposed in the original paper.

Contribution 3: Theoretical proof for random-feature models and empirical validation across 23 data recipes

Description: The authors provide formal theoretical justification showing that tiny learning rates preserve dataset orderings relative to infinite-width optimal losses in random-feature models. They also conduct extensive experiments spanning multiple architectures, scales, and data curation scenarios to demonstrate dramatic improvements in proxy model reliability.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Bootstrap aggregating and random forest

URL: View paper

Brief Assessment

Bootstrap Aggregating[55] focuses on ensemble methods (bagging, random forests) for improving prediction robustness through model averaging, not on theoretical analysis of random-feature models or dataset ordering preservation under tiny learning rates.

2. Not-So-Random Features

URL: View paper

Brief Assessment

Not So Random Features[60] focuses on kernel learning via Fourier analysis for SVM classification, not on dataset ordering preservation or data curation for language model pretraining.

3. Data Sourcing Random Access using Semantic Queries for Massive IoT Scenarios

URL: View paper

Brief Assessment

Semantic Query Access[53] focuses on data retrieval protocols for IoT networks using semantic queries and random access channels, not on random-feature models or dataset ordering preservation in machine learning training.

4. Optimal Sequence Memory in Driven Random Networks

URL: View paper

Brief Assessment

Optimal Sequence Memory[58] focuses on sequence memory in driven random networks, which is a different domain from data curation and proxy model training for language models. The candidate paper does not address dataset ordering or proxy model reliability.

5. A hierarchical Vovk-Azoury-Warmuth forecaster with discounting for online regression in RKHS

URL: View paper

Brief Assessment

Hierarchical Vovk Azoury[59] focuses on online regression in RKHS using random features for dynamic regret bounds, not on dataset ordering preservation in language model pretraining with tiny learning rates.

6. Expected pinball loss for quantile regression and inverse cdf estimation

URL: View paper

Brief Assessment

Expected Pinball Loss[54] focuses on quantile regression and inverse CDF estimation using random feature models with asymptotic convergence analysis. The original paper addresses dataset ordering preservation in neural network training with tiny learning rates for data curation decisions, which is a fundamentally different problem domain.

7. Optimization and evaluation of ensemble learning models for intelligent lithology identification based on seismic data

URL: View paper

Brief Assessment

Ensemble Lithology Identification[56] focuses on lithology identification using seismic data with ensemble learning models. The candidate's context mentions random feature selection in decision trees and gradient optimization, but provides no evidence of theoretical analysis of random-feature models preserving dataset ordering or data curation for language models.

8. A randomized optimal k-mer indexing approach for efficient parallel genome sequence compression.

URL: View paper

Brief Assessment

Randomized Kmer Indexing[57] focuses on genome sequence compression using k-mer indexing techniques, which is entirely unrelated to random-feature models, dataset ordering, or machine learning training dynamics discussed in the original paper.

Appendix: Text Similarity Detection

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

References

- [0] Can Small Training Runs Reliably Guide Data Curation? Rethinking Proxy-Model Practice View paper
- [1] Smaller language models are capable of selecting instruction-tuning training data for larger language models View paper
- [2] Smaller Can Be Better: Efficient Data Selection for Pre-training Models View paper
- [3] Regmix: Data mixture as regression for language model pre-training View paper
- [4] Nemotron-CLIMB: CLustering-based Iterative Data Mixture Bootstrapping for Language Model Pre-training View paper
- [5] Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models View paper
- [6] Bad students make great teachers: Active learning accelerates large-scale visual understanding View paper
- [7] Concept-skill Transferability-based Data Selection for Large Vision-Language Models View paper
- [8] Selection via proxy: Efficient data selection for deep learning View paper
- [9] RL-Guided Data Selection for Language Model Finetuning View paper
- [10] PieBridge: Fast and Parameter-Efficient On-Device Training via Proxy Networks View paper
- [11] Pruning-based Data Selection and Network Fusion for Efficient Deep Learning View paper
- [12] Neural data server: A large-scale search engine for transfer learning data View paper
- [13] Small Language Model as Data Prospector for Large Language Model View paper
- [14] Proxy Model-Guided Reinforcement Learning for Client Selection in Federated Recommendation View paper
- [15] Less is more: Proxy datasets in nas approaches View paper
- [16] Research on Fine-Tuning Optimization of Large Language Models Based on Online Data Mixing View paper
- [17] Filtering noisy parallel corpus using transformers with proxy task learning View paper
- [18] PAPM: A Physics-aware Proxy Model for Process Systems View paper
- [19] BLISS: A Lightweight Bilevel Influence Scoring Method for Data Selection in Language Model Pretraining View paper
- [20] Small-to-Large Generalization: Data Influences Models Consistently Across Scale View paper
- [21] Compute-Constrained Data Selection View paper
- [22] Active Data Curation Effectively Distills Large-Scale Multimodal Models View paper

- [23] RefineNet: Elevating Medical Foundation Models Through Quality-Centric Data Curation by MLLM-Annotated Proxy Distillation View paper
- [24] Proxy-RLHF: Decoupling Generation and Alignment in Large Language Model with Proxy View paper
- [25] Generating Complex Question Decompositions in the Face of Distribution Shifts View paper
- [26] Tuning Language Models by Proxy View paper
- [27] A Comparative Survey: Reusing Small Pre-Trained Models for Efficient Large Model Training View paper
- [28] Navigating Towards Fairness with Data Selection View paper
- [29] Foundation Models from a Data-Distributional View View paper
- [30] Small-Scale or Large-Scale Machine Learning Proxy Models? a Real Field Case Study View paper
- [31] Confusing Large Models by Confusing Small Models View paper
- [32] JiuZhang3.0: Efficiently Improving Mathematical Reasoning by Training Small Data Synthesis Models View paper
- [33] SelectFormer: Private and Practical Data Selection for Transformers View paper
- [34] Improving Large-Scale Conversational Assistants using Model Interpretation based Training Sample Selection View paper
- [35] Cynical Selection of Language Model Training Data View paper
- [36] Small-to-Large Generalization: Training Data Influences Models Consistently Across Scale View paper
- [37] FAMIE: A Fast Active Learning Framework for Multilingual Information Extraction View paper
- [38] SEED: Domain-Specific Data Curation With Large Language Models View paper
- [39] Smollm2: When smol goes bigâ□□data-centric training of a fully open small language model View paper
- [40] Colorization as a Proxy Task for Visual Understanding View paper
- [41] PruneFuse: Efficient Data Selection via Weight Pruning and Network Fusion View paper
- [42] Spaced Scheduling for Large Language Model Training View paper
- [43] Soft Sampling for Efficient Training of Deep Neural Networks on Massive Data View paper
- [44] Step-by-step enhancement of a graph neural network-based surrogate model for Lagrangian fluid simulations with flexible time step sizes View paper
- [45] Meta-learning innovates chemical kinetics: An efficient approach for surrogate model construction View paper
- [46] pFedES: Generalized Proxy Feature Extractor Sharing for Model Heterogeneous Personalized Federated Learning View paper
- [47] Quantifying the uncertainty of structural parameters using machine learningâ□□based surrogate models View paper
- [48] Layer-Wise Learning Rate Optimization for Task-Dependent Fine-Tuning of Pre-Trained Models: An Evolutionary Approach View paper
- [49] An Empirical Study of P Learning Rate Transfer View paper
- [50] Classical feature map surrogates and metrics for quantum control landscapes View paper
- [51] A Study of Generalization of Stochastic Mirror Descent Algorithms on Overparameterized Nonlinear Models View paper
- [52] Scalable gaussian process-based transfer surrogates for hyperparameter optimization View paper
- [53] Data Sourcing Random Access using Semantic Queries for Massive IoT Scenarios View paper
- [54] Expected pinball loss for quantile regression and inverse cdf estimation View paper
- [55] Bootstrap aggregating and random forest View paper
- [56] Optimization and evaluation of ensemble learning models for intelligent lithology identification based on seismic data View paper
- [57] A randomized optimal k-mer indexing approach for efficient parallel genome sequence compression. View paper
- [58] Optimal Sequence Memory in Driven Random Networks View paper
- [59] A hierarchical Vovk-Azoury-Warmuth forecaster with discounting for online regression in RKHS View paper
- [60] Not-So-Random Features View paper