Novelty Assessment Report

Paper: How Text Quality Interventions Reshape Neural Scaling Laws for LLMs: Empirical Study

PDF URL: https://openreview.net/pdf?id=ZC5QBfdOw7

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Neural scaling laws are widely used for performance projection and resource planning, yet their sensitivity to data quality interventions remains poorly understood. We present an empirical study of how interventions—deduplication, heuristic filtering, and LLM-guided rewriting—reshape scaling behavior in large language model training. Using QualityPajama, a suite of 23 systematically filtered and synthetic datasets, we train over 2,000 models (100M–8B parameters, 100M–200B tokens) to measure how data quality affects scaling-law parameters and compute-optimal design decisions. Our results show that data interventions reshape scaling dynamics in non-trivial ways not captured by current theory, simultaneously moving exponents, coefficients, and constants in conflicting directions that exert opposing forces on loss. For example, an intervention may improve constants but hurt the exponents. Strategies that appear optimal at small scale can reverse at larger scale, and compute-optimal token–parameter ratios can vary by orders of magnitude depending on the intervention. These findings demonstrate that data curation and scaling strategy are deeply intertwined, and that evaluating interventions only at fixed scales can lead to misleading conclusions. We recommend evaluating interventions through their full scaling trajectories using scaling law projections.

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: **impact of text quality interventions on neural scaling laws**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

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

- **Data Quality Characterization and Filtering Methods**
- **Scaling Law Theory and Formulation**
- **Compute-Optimal Resource Allocation**
- **Data Augmentation and Synthesis**
- **Domain-Specific Applications and Datasets**
- **Training Infrastructure and Methodology**
- **Specialized Theoretical and Methodological Studies**

Complete Taxonomy Tree

- impact of text quality interventions on neural scaling laws Survey Taxonomy
- Data Quality Characterization and Filtering Methods
  - Quality Metrics and Scoring Systems (3 papers)
  - [3] Scalingfilter: Assessing data quality through inverse utilization of scaling laws (Ruihang Li, 2024) View paper
  - [4] Scaling laws revisited: modeling the role of data quality in language model pretraining (Subramanyam, 2025) View paper
  - [32] Scaling Parameter-Constrained Language Models with Quality Data (Ernie Chang, 2024) View paper
  - Heuristic and Rule-Based Filtering (2 papers)
  - [1] Optimizing dataset creation: A general purpose data filtering system for training large language models (Yanbing Wang, 2024) View paper
  - [44] Blu-WERP (Web Extraction and Refinement Pipeline): A Scalable Pipeline for Preprocessing Large Language Model Datasets (Gowtham, 2025) View paper
  - Model-Guided Data Curation (3 papers)
  - [22] Aleph-Alpha-GermanWeb: Improving German-language LLM pre-training with model-based data curation and synthetic data generation (Burns, 2025) View paper
  - [24] Data curation via joint example selection further accelerates multimodal learning (Talfan Evans, 2024) View paper
  - [25] Language Models Improve When Pretraining Data Matches Target Tasks (Mizrahi David, 2025) View paper
- Scaling Law Theory and Formulation
  - Quality-Aware Scaling Law Extensions ★ (3 papers)
  - [0] How Text Quality Interventions Reshape Neural Scaling Laws for LLMs: Empirical Study (Anon et al., 2026) View paper
  - [2] Farseer: A Refined Scaling Law in Large Language Models (H Li, 2025) View paper
  - [15] Revisiting scaling laws for language models: The role of data quality and training strategies (Zhengyu Chen, 2025) View paper
  - Specialized Scaling Frameworks (3 papers)
  - [5] The science of data filtering: Data curation cannot be compute agnostic (S Goyal, 2024) View paper
  - [8] Domain-aware scaling laws uncover data synergy (K Hamidieh, 2025) View paper
  - [47] Scaling Laws for Data Filteringâ□□Data Curation Cannot be Compute Agnostic (Sachin Goyal, 2024) View paper
  - Theoretical Limits and Power Law Alternatives (2 papers)
  - [6] Beyond neural scaling laws: beating power law scaling via data pruning (Sorscher, 2022) View paper

    - [10] Predictability and surprise in large generative models (Deep Ganguli, 2022) View paper
    - Scaling Law Analysis and Methodology (2 papers)
    - [19] (mis) fitting scaling laws: A survey of scaling law fitting techniques in deep learning (M Li, 2025) View paper
    - [23] Observational scaling laws and the predictability of langauge model performance (Tatsunori Hashimoto, 2024) View paper
- Compute-Optimal Resource Allocation
    - Dynamic Data Composition Optimization (2 papers)
    - [12] Adaptive data optimization: Dynamic sample selection with scaling laws (Jiang, 2024) View paper
    - [29] Autoscale: Scale-aware data mixing for pre-training llms (Kang, 2024) View paper
    - Static Data Source Selection (2 papers)
    - [26] AutoScale: Automatic Prediction of Compute-optimal Data Compositions for Training LLMs (F Kang, 2024) View paper
    - [46] Using Scaling Laws for Data Source Utility Estimation in Domain-Specific Pre-Training (Ostapenko, 2025) View paper
    - Data-Constrained Scaling Strategies (2 papers)
    - [34] Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization (Divya Shanmugam, 2022) View paper
    - [41] Scaling Data-Constrained Language Models (Muennighoff, 2023) View paper
- Data Augmentation and Synthesis
    - Synthetic Data Generation for Training (2 papers)
    - [13] ConvKGYarn: Spinning Configurable and Scalable Conversational Knowledge Graph QA Datasets with Large Language Models (Ronak Pradeep, 2024) View paper
    - [30] Scaling Laws of Synthetic Images for Model Training â¦ for Now (Lijie Fan, 2023) View paper
    - Data Transformation and Refinement (2 papers)
    - [21] Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy (Zeng Liang, 2025) View paper
    - [42] Data Curation Through the Lens of Spectral Dynamics: Static Limits, Dynamic Acceleration, and Practical Oracles (Yizhou Zhang, 2025) View paper
- Domain-Specific Applications and Datasets
    - Multimodal Vision-Language Systems (5 papers)
    - [7] Datacomp: In search of the next generation of multimodal datasets (Gadre, 2023) View paper
    - [9] No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance (Samuel Albanie, 2024) View paper
    - [43] Wan: Open and Advanced Large-Scale Video Generative Models (Wang Ang, 2025) View paper
    - [45] Frontier AI From the Outside In: Advances in Data Curation, Data Distillation and Model Evaluation (Feuer, 2025) View paper
    - [48] Scalable Vision Language Model Training via High Quality Data Curation (Hongyuan Dong, 2025) View paper
    - Specialized Language Modeling Domains (3 papers)
    - [17] Tele-LLMs: A Series of Specialized Large Language Models for Telecommunications (Maatouk, 2024) View paper
    - [39] Skywork-SWE: Unveiling Data Scaling Laws for Software Engineering in LLMs (Zeng Liang, 2025) View paper
    - [49] Empowering Large Language Models in Wireless Communication: A Novel Dataset and Fine-Tuning Framework (Yushen Lin, 2025) View paper
    - Specialized Sensor and Behavioral Data (2 papers)
    - [27] SensorLM: Learning the Language of Wearable Sensors (Zhang Yu-wei, 2025) View paper
    - [50] Scaling behavior of large language models in emotional safety classification across sizes and tasks (TÃ¼scher Oliver, 2025) View paper
- Training Infrastructure and Methodology
    - Distributed Training Systems (2 papers)
    - [20] Scaling Up Models and Data with t5x and seqio (Roberts, 2022) View paper
    - [37] Large-scale model training: Dataset construction, reliable scaling, and task-specific adaptation (Gadre, 2025) View paper
    - Post-Training Optimization Techniques (3 papers)
    - [14] OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling (Wang Zeng-zhi, 2025) View paper
    - [16] Self-Improving Vision-Language-Action Models with Data Generation via Residual RL (Xiao Wenli, 2025) View paper
    - [38] Training Data Curation for Language Models with Weak Supervision (Mekala, 2025) View paper
    - Model Architecture and Scaling Behavior (4 papers)
    - [18] Not-just-scaling laws: Towards a better understanding of the downstream impact of language model design decisions (Emmy Liu, 2025) View paper
    - [28] Limiting factors for the continued scaling of Large Langauge Models: Data Sets, efficient systems for training, model architecture and novel hardware (Hesslow, 2024) View paper
    - [35] The Cost of Down-Scaling Language Models: Fact Recall Deteriorates before In-Context Learning (Jin Tian, 2023) View paper
    - [36] Data scaling laws in NMT: The effect of noise and architecture (Bansal, 2022) View paper
- Specialized Theoretical and Methodological Studies
    - Knowledge Distillation Theory (1 papers)
    - [11] High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws (Ildiz, 2024) View paper
    - Non-NLP Neural Network Applications (3 papers)
    - [31] Leukemia Detection Using Invariant Structural Cascade Segmentation Based on Deep Vectorized Scaling Neural Network (A. Arthi, 2023) View paper
    - [33] Random Coupled Neural Network with Sand Cat Swarm Optimization for Automatic Object Detection in Aerial Images (Seeniappan Kaliappan, 2024) View paper
    - [40] Scale selection in convolutional neural networks with dimensional min-pooling and scaling filters (Vienken, 2016) View paper

Narrative

Core task: understanding how interventions on text quality affect neural scaling laws. The field has organized itself into several major branches that collectively address data quality, theoretical formulations, resource allocation, augmentation strategies, domain-specific applications, training infrastructure, and specialized methodological studies. Data Quality Characterization and Filtering Methods explore how to measure and improve dataset quality through filtering techniques, with works like General Purpose Filtering[1] and Scalingfilter[3] developing principled approaches to curate training corpora. Scaling Law Theory and Formulation investigates the mathematical relationships between model performance, data size, and compute, extending classical power-law formulations to account for quality dimensions; representative studies include Farseer[2] and Data Quality Scaling[4], which incorporate quality metrics into predictive frameworks. Compute-Optimal Resource Allocation examines trade-offs in distributing computational budgets across model size and training tokens, while Data Augmentation and Synthesis considers synthetic data generation as a lever for scaling. Domain-Specific Applications and Datasets apply these principles to specialized contexts such as vision-language tasks or scientific domains, and Training Infrastructure and Methodology addresses practical implementation challenges at scale.

A particularly active line of inquiry centers on quality-aware extensions to classical scaling laws, where researchers seek to move beyond simple data-volume metrics to incorporate notions of data cleanliness, diversity, and relevance. Text Quality Scaling[0] sits squarely within this branch, examining how targeted quality interventions shift the scaling behavior predicted by traditional formulations. This work shares thematic overlap with Farseer[2], which also models quality effects on downstream performance, and with Data Filtering Science[5], which systematically studies filtering strategies. Compared to Revisiting Scaling Laws[15], which re-examines foundational assumptions in scaling theory, Text Quality Scaling[0] emphasizes empirical interventions and their measurable impact on the scaling exponent. A central open question across these studies is whether quality improvements can substitute for raw data volume or whether they interact multiplicatively, and how to predict the returns from quality-focused curation at different scales.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Farseer: A Refined Scaling Law in Large Language Models

Authors: H Li, W Zheng, Q Wang, Z Ding, H Wang | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ 42], while other studies apply scaling principles to improve data curation, eg, by strategically â¦ By first defining a family and then empirically determining its scaling behavior, we can â¦

Relationship Analysis

Both papers belong to the Quality-Aware Scaling Law Extensions category, focusing on how data quality factors influence neural scaling law parameters beyond traditional model size and data volume considerations. The original paper empirically investigates how specific text quality interventions (deduplication, filtering, synthetic rewriting) reshape all components of scaling laws (coefficients, exponents, and asymptotic loss), while Farseer proposes a refined mathematical formulation of scaling laws with improved predictive accuracy and extrapolation capabilities. The key difference is that the original paper focuses on understanding the impact of diverse data quality interventions on scaling behavior, whereas Farseer develops a more accurate functional form for the scaling law itself without explicitly analyzing quality interventions.

2. Revisiting scaling laws for language models: The role of data quality and training strategies

Authors: Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ scaling law that better predicts performance in sub-scaling regimes, highlighting the importance of data quality â¦ Sub-Optimal Scaling Law: Our introduction of a sub-optimal scaling law â¦

Relationship Analysis

Both papers belong to the Quality-Aware Scaling Law Extensions category, investigating how data quality factors reshape traditional neural scaling laws for language models. They overlap in examining how data interventions (filtering, deduplication, quality metrics) affect scaling law parameters (exponents and coefficients) and compute-optimal training strategies. The original paper focuses on systematic text quality interventions (heuristic filtering, LLM-guided rewriting, synthetic data mixing) across 23 curated datasets with 2000+ models, while the candidate paper emphasizes data density as a primary driver of sub-scaling phenomena and proposes a sub-optimal scaling law formulation with decay factors to model diminishing returns in high-density datasets.

Contributions Analysis

Overall novelty summary. The paper investigates how data quality interventions—deduplication, heuristic filtering, and LLM-guided rewriting—reshape neural scaling laws through extensive empirical study. It resides in the 'Quality-Aware Scaling Law Extensions' leaf, which contains only three papers total, including this work and two siblings (Farseer and Data Quality Scaling). This represents a relatively sparse research direction within the broader taxonomy of 50 papers, suggesting the specific focus on decomposing scaling law parameters under quality interventions remains underexplored compared to adjacent areas like heuristic filtering or compute optimization.

The taxonomy reveals neighboring research in 'Specialized Scaling Frameworks' (multi-domain and constrained-resource contexts) and 'Scaling Law Analysis and Methodology' (fitting and validation techniques). The paper's emphasis on how interventions simultaneously affect exponents, coefficients, and constants distinguishes it from sibling work: Farseer models quality effects on downstream performance, while Data Quality Scaling examines quality metrics more abstractly. The scope note for this leaf explicitly includes 'incorporating data quality as an explicit parameter,' which aligns with the paper's decomposition approach, though the exclude note clarifies it differs from domain-specific or multi-source frameworks.

Among 30 candidates examined across three contributions, none were found to clearly refute any claim. The QualityPajama Benchmark examined 10 candidates with zero refutable overlaps; Full Scaling Law Decomposition and Data-Aware Scaling Strategies each examined 10 candidates with similar results. This suggests that within the limited search scope, the specific combination of systematic quality interventions, large-scale model training (2,000+ models), and full parameter decomposition appears distinctive. However, the analysis explicitly notes this is based on top-K semantic search plus citation expansion, not exhaustive coverage.

Given the sparse population of the quality-aware scaling law leaf and the absence of refuting prior work among 30 examined candidates, the contributions appear to occupy relatively novel ground within the analyzed scope. The scale of empirical validation (23 datasets, 2,000 models) and the focus on conflicting directional effects across scaling parameters distinguish this work from its immediate siblings, though the limited search scope means potentially relevant work outside the top-30 semantic matches may exist.

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

Contribution 1: QualityPajama Benchmark

Description: The authors present QualityPajama, a benchmark consisting of 23 systematically curated datasets derived from Common Crawl. Each dataset represents a different text quality intervention (including filtering, deduplication, and synthetic curation) to enable controlled study of how data quality affects scaling laws in large language models.

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. Data curation via joint example selection further accelerates multimodal learning

URL: View paper

Brief Assessment

Joint Example Selection[24] focuses on multimodal contrastive learning and batch-level data selection for vision-language models, not on creating benchmarks for evaluating text quality interventions on language model scaling laws.

2. Densing law of llms

URL: View paper

Brief Assessment

Densing Law[63] focuses on capability density metrics for evaluating LLM efficiency over time, not on benchmarks for text quality interventions or neural scaling laws.

3. Exponential scaling of factual inconsistency in data-to-text generation with fine-tuned LLMs

URL: View paper

Brief Assessment

Factual Inconsistency Scaling[66] focuses on scaling laws for factual inconsistency in data-to-text generation tasks, not on benchmarking text quality interventions for neural scaling in language models.

4. Scalingfilter: Assessing data quality through inverse utilization of scaling laws

URL: View paper

Brief Assessment

Scalingfilter[3] does not present a benchmark of systematically curated datasets. Instead, it proposes a filtering method that evaluates text quality based on perplexity differences between models of different sizes, without creating multiple intervention-based datasets for controlled study.

5. Scaling Laws for Downstream Task Performance in Machine Translation

URL: View paper

Brief Assessment

Downstream Task Scaling[70] focuses on machine translation downstream performance and does not present a benchmark for evaluating text quality interventions on neural scaling in language models. The candidate studies transfer learning from pretraining to finetuning, not systematic text quality interventions like filtering, deduplication, and synthetic curation.

6. Llm-generated natural language meets scaling laws: New explorations and data augmentation methods

URL: View paper

Brief Assessment

Natural Language Scaling[69] focuses on evaluating LLM-generated text quality through scaling laws (Zipf's, Heaps', etc.) rather than creating a benchmark for text quality interventions on neural scaling in language models. The candidate does not present a systematically curated dataset suite for studying how data quality affects scaling laws.

7. OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization

URL: View paper

Brief Assessment

OPT IML[67] focuses on instruction-tuning benchmarks for task generalization in NLP, not on text quality interventions or neural scaling laws for pretraining data curation.

8. CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation

URL: View paper

Brief Assessment

CopyBench[64] focuses on measuring literal and non-literal reproduction of copyright-protected text in language models, not on evaluating text quality interventions or neural scaling laws. The benchmarks serve entirely different purposes and research domains.

9. Exploring training and inference scaling laws in generative retrieval

URL: View paper

Brief Assessment

Generative Retrieval Scaling[65] focuses on scaling laws for generative retrieval systems (document generation from queries), not on benchmarking text quality interventions for language model pretraining. The domains are fundamentally different.

10. Towards trustable language models: Investigating information quality of large language models

URL: View paper

Brief Assessment

Trustable Language Models[68] focuses on mathematical formulation of information quality evaluation (consistency, relevance, accuracy) and general data quality issues in LLMs, not on systematic benchmarks for text quality interventions affecting scaling laws.

Contribution 2: Full Scaling Law Decomposition

Description: The authors conduct the first comprehensive empirical analysis showing that text quality interventions reshape all components of neural scaling laws (exponents, coefficients, and asymptotic loss terms), not just the exponents as prior work assumed. They demonstrate that stronger filtering produces conflicting shifts across different parameters rather than uniformly favorable changes.

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. (mis) fitting scaling laws: A survey of scaling law fitting techniques in deep learning

URL: View paper

Brief Assessment

Fitting Scaling Laws[19] is a survey paper focused on the technical details of fitting scaling law equations (forms, optimization methods, initialization). It does not study how text quality interventions affect scaling law components, which is the core novelty claim of the original paper.

2. Not-just-scaling laws: Towards a better understanding of the downstream impact of language model design decisions

URL: View paper

Brief Assessment

Not Just Scaling[18] focuses on predicting downstream task performance by incorporating architectural and data composition features beyond scale, rather than analyzing how text quality interventions reshape all components of neural scaling laws (exponents, coefficients, and asymptotic loss terms) as the original paper does.

3. Data curation via joint example selection further accelerates multimodal learning

URL: View paper

Brief Assessment

Joint Example Selection[24] does not analyze how text quality interventions affect all components of neural scaling laws (exponents, coefficients, asymptotic terms). It focuses on batch selection strategies for multimodal learning, not decomposing scaling law parameters.

4. Scalingfilter: Assessing data quality through inverse utilization of scaling laws

URL: View paper

Brief Assessment

Scalingfilter[3] focuses on using scaling laws inversely to assess data quality through perplexity differences, rather than analyzing how text quality interventions reshape all components (exponents, coefficients, asymptotic loss terms) of neural scaling laws.

5. Datacomp: In search of the next generation of multimodal datasets

URL: View paper

Brief Assessment

Datacomp[7] focuses on multimodal dataset curation and filtering strategies for image-text pairs, not on how text quality interventions reshape neural scaling law components (exponents, coefficients, asymptotic terms) for LLM pretraining.

6. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

URL: View paper

Brief Assessment

Noisy Text Supervision[61] focuses on visual and vision-language representation learning using noisy image-text pairs, not on neural scaling laws for LLMs or how text quality interventions affect scaling law components (exponents, coefficients, asymptotic terms).

7. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance

URL: View paper

Brief Assessment

Zero Shot Exponential[9] focuses on concept frequency in multimodal pretraining data and its log-linear relationship with downstream performance. This work does not address text quality interventions or neural scaling law components (exponents, coefficients, asymptotic terms) in LLM pretraining, which is the focus of the original contribution.

8. Sub-scaling laws: on the role of data density and training strategies in llms

URL: View paper

Brief Assessment

Sub Scaling Laws[62] focuses on sub-scaling phenomena (performance deceleration) in specific regimes (high data density, over-training), not on comprehensive decomposition of all scaling law components across diverse text quality interventions as claimed by the original paper.

9. Scaling Laws for Data Filteringâ□□Data Curation Cannot be Compute Agnostic

URL: View paper

Brief Assessment

Data Filtering Laws[47] focuses on vision-language models and data repetition effects, not text quality interventions in LLM pretraining. The candidate examines how data utility diminishes with repetition across compute budgets, while the original analyzes how text filtering/deduplication/rewriting reshape all scaling law components (exponents, coefficients, asymptotic terms) in language model training.

10. Scaling laws for reward model overoptimization in direct alignment algorithms

URL: View paper

Brief Assessment

Reward Overoptimization Scaling[60] focuses on reward model overoptimization in direct alignment algorithms for LLMs, not on how text quality interventions reshape neural scaling law components (exponents, coefficients, asymptotic terms) during pretraining.

Contribution 3: Data-Aware Scaling Strategies

Description: The authors demonstrate that data quality fundamentally affects compute-optimal design decisions in LLM training. They show that different quality interventions can shift the optimal number of parameters, training tokens, and their ratio by orders of magnitude, revealing that scaling strategies must explicitly account for data quality rather than treating it as a secondary concern.

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. Physics of language models: Part 3.3, knowledge capacity scaling laws

URL: View paper

Brief Assessment

Knowledge Capacity Scaling[56] focuses on knowledge storage capacity (bits per parameter) in language models, not on compute-optimal scaling strategies or data quality interventions affecting training decisions. The candidate examines how model size relates to knowledge storage, while the original contribution addresses how data quality shifts optimal parameter/token ratios during training.

2. Is training data quality or quantity more impactful to small language model performance

URL: View paper

Brief Assessment

Quality Versus Quantity[58] focuses on small language models (≤100M parameters) using the TinyStories dataset, examining duplication effects on character-level prediction. The original paper studies large-scale LLM training (100M-8B parameters, 100M-200B tokens) with diverse quality interventions (filtering, deduplication, synthetic rewriting) and their impact on compute-optimal scaling laws. These are fundamentally different experimental scopes and research questions.

3. Datacomp-lm: In search of the next generation of training sets for language models

URL: View paper

Brief Assessment

Datacomp LM[51] focuses on data curation methods (filtering, deduplication, mixing) to improve model performance, but does not systematically analyze how data quality affects all scaling law components (α, β, a, b, e) or compute-optimal design decisions across different scales and resource constraints.

4. Will we run out of data? an analysis of the limits of scaling datasets in machine learning

URL: View paper

Brief Assessment

Data Limits Analysis[54] focuses on exhaustion timelines for available training data stocks (language/vision data running out by 2030-2060), not on how data quality affects compute-optimal design decisions or scaling law parameters (α, β, a, b, e) as the original paper demonstrates.

5. Position: Will we run out of data? Limits of LLM scaling based on human-generated data

URL: View paper

Brief Assessment

Human Data Limits[59] focuses on the availability and exhaustion of human-generated text data as a constraint on LLM scaling, not on how data quality interventions affect compute-optimal design decisions or scaling law parameters.

6. The falcon series of open language models

URL: View paper

Brief Assessment

Falcon Series[52] focuses on web data quality and filtering strategies for LLM pretraining, not on how data quality interventions reshape compute-optimal scaling laws or affect the relationship between optimal parameters, tokens, and their ratios across different quality levels.

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

URL: View paper

Brief Assessment

Data Quality Scaling[4] focuses on introducing a dimensionless quality parameter q into scaling laws to model data corruption and deficiency effects. The original paper examines how diverse text quality interventions (filtering, deduplication, synthetic data) reshape all scaling law components and compute-optimal design decisions. These are complementary perspectives on data quality's role in scaling.

8. Kanana: Compute-efficient Bilingual Language Models

URL: View paper

Brief Assessment

Kanana[53] focuses on practical techniques for building compute-efficient bilingual models (filtering, staged pre-training, pruning) rather than studying how data quality fundamentally affects scaling law parameters and compute-optimal design decisions across different scales and resource constraints.

9. Data Engineering for Scaling Language Models to 128K Context

URL: View paper

Brief Assessment

Context Engineering[55] focuses on continual pretraining for extending context length (4k to 128k tokens) through data engineering, not on compute-optimal scaling strategies or how data quality affects the fundamental scaling law parameters (α, β, a, b, e) across different model sizes and training token budgets.

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

URL: View paper

Brief Assessment

Compute Optimal Training[57] focuses on optimal model size and training token allocation for a fixed compute budget using a single dataset (MassiveText). The original paper examines how different data quality interventions systematically affect scaling law components across 23 curated datasets, which is a fundamentally different research question.

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 2 similarity segment(s) across 2 paper(s).

The following **2 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. Exploring training and inference scaling laws in generative retrieval

Detected in: Contribution: contribution_1

⚠ **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.

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

Detected in: Contribution: contribution_3

⚠ **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] How Text Quality Interventions Reshape Neural Scaling Laws for LLMs: Empirical Study View paper
- [1] Optimizing dataset creation: A general purpose data filtering system for training large language models View paper
- [2] Farseer: A Refined Scaling Law in Large Language Models View paper
- [3] Scalingfilter: Assessing data quality through inverse utilization of scaling laws View paper
- [4] Scaling laws revisited: modeling the role of data quality in language model pretraining View paper
- [5] The science of data filtering: Data curation cannot be compute agnostic View paper
- [6] Beyond neural scaling laws: beating power law scaling via data pruning View paper
- [7] Datacomp: In search of the next generation of multimodal datasets View paper
- [8] Domain-aware scaling laws uncover data synergy View paper
- [9] No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance View paper
- [10] Predictability and surprise in large generative models View paper
- [11] High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws View paper
- [12] Adaptive data optimization: Dynamic sample selection with scaling laws View paper
- [13] ConvKGYarn: Spinning Configurable and Scalable Conversational Knowledge Graph QA Datasets with Large Language Models View paper
- [14] OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling View paper
- [15] Revisiting scaling laws for language models: The role of data quality and training strategies View paper
- [16] Self-Improving Vision-Language-Action Models with Data Generation via Residual RL View paper
- [17] Tele-LLMs: A Series of Specialized Large Language Models for Telecommunications View paper
- [18] Not-just-scaling laws: Towards a better understanding of the downstream impact of language model design decisions View paper
- [19] (mis) fitting scaling laws: A survey of scaling law fitting techniques in deep learning View paper
- [20] Scaling Up Models and Data with t5x and seqio View paper
- [21] Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy View paper
- [22] Aleph-Alpha-GermanWeb: Improving German-language LLM pre-training with model-based data curation and synthetic data generation View paper
- [23] Observational scaling laws and the predictability of langauge model performance View paper
- [24] Data curation via joint example selection further accelerates multimodal learning View paper
- [25] Language Models Improve When Pretraining Data Matches Target Tasks View paper
- [26] AutoScale: Automatic Prediction of Compute-optimal Data Compositions for Training LLMs View paper
- [27] SensorLM: Learning the Language of Wearable Sensors View paper
- [28] Limiting factors for the continued scaling of Large Langauge Models: Data Sets, efficient systems for training, model architecture and novel hardware View paper
- [29] Autoscale: Scale-aware data mixing for pre-training llms View paper
- [30] Scaling Laws of Synthetic Images for Model Training â□¦ for Now View paper
- [31] Leukemia Detection Using Invariant Structural Cascade Segmentation Based on Deep Vectorized Scaling Neural Network View paper
- [32] Scaling Parameter-Constrained Language Models with Quality Data View paper
- [33] Random Coupled Neural Network with Sand Cat Swarm Optimization for Automatic Object Detection in Aerial Images View paper
- [34] Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization View paper
- [35] The Cost of Down-Scaling Language Models: Fact Recall Deteriorates before In-Context Learning View paper
- [36] Data scaling laws in NMT: The effect of noise and architecture View paper
- [37] Large-scale model training: Dataset construction, reliable scaling, and task-specific adaptation View paper
- [38] Training Data Curation for Language Models with Weak Supervision View paper
- [39] Skywork-SWE: Unveiling Data Scaling Laws for Software Engineering in LLMs View paper
- [40] Scale selection in convolutional neural networks with dimensional min-pooling and scaling filters View paper
- [41] Scaling Data-Constrained Language Models View paper
- [42] Data Curation Through the Lens of Spectral Dynamics: Static Limits, Dynamic Acceleration, and Practical Oracles View paper
- [43] Wan: Open and Advanced Large-Scale Video Generative Models View paper
- [44] Blu-WERP (Web Extraction and Refinement Pipeline): A Scalable Pipeline for Preprocessing Large Language Model Datasets View paper
- [45] Frontier AI From the Outside In: Advances in Data Curation, Data Distillation and Model Evaluation View paper
- [46] Using Scaling Laws for Data Source Utility Estimation in Domain-Specific Pre-Training View paper
- [47] Scaling Laws for Data Filteringâ□□Data Curation Cannot be Compute Agnostic View paper
- [48] Scalable Vision Language Model Training via High Quality Data Curation View paper
- [49] Empowering Large Language Models in Wireless Communication: A Novel Dataset and Fine-Tuning Framework View paper
- [50] Scaling behavior of large language models in emotional safety classification across sizes and tasks View paper
- [51] Datacomp-lm: In search of the next generation of training sets for language models View paper
- [52] The falcon series of open language models View paper
- [53] Kanana: Compute-efficient Bilingual Language Models View paper
- [54] Will we run out of data? an analysis of the limits of scaling datasets in machine learning View paper
- [55] Data Engineering for Scaling Language Models to 128K Context View paper
- [56] Physics of language models: Part 3.3, knowledge capacity scaling laws View paper
- [57] An empirical analysis of compute-optimal large language model training View paper
- [58] Is training data quality or quantity more impactful to small language model performance View paper
- [59] Position: Will we run out of data? Limits of LLM scaling based on human-generated data View paper
- [60] Scaling laws for reward model overoptimization in direct alignment algorithms View paper
- [61] Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision View paper
- [62] Sub-scaling laws: on the role of data density and training strategies in llms View paper
- [63] Densing law of llms View paper

- [64] CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation View paper
- [65] Exploring training and inference scaling laws in generative retrieval View paper
- [66] Exponential scaling of factual inconsistency in data-to-text generation with fine-tuned LLMs View paper
- [67] OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization View paper
- [68] Towards trustable language models: Investigating information quality of large language models View paper
- [69] Llm-generated natural language meets scaling laws: New explorations and data augmentation methods View paper
- [70] Scaling Laws for Downstream Task Performance in Machine Translation View paper