Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3170

A THREE-STAGE GENERALIZED SCALING LAW FOR LARGE LANGUAGE MODEL PRE-TRAINING

GengYin Li[1], Xue Yuan[2*], XueTong Zang[2]
[1]*School of Energy and Materials, Shanghai Polytechnic University, Shanghai 201209, China.*
[2]*School of Resources and Environmental Engineering, Shanghai Polytechnic University, Shanghai 201209, China.*
[*]*Corresponding Author: Xue Yuan*

Abstract: Large language models (LLMs) acquire their capabilities primarily through scaling, and the rapidly rising cost of training has turned compute-budget allocation into a central practical concern. Classical scaling laws express loss as a function of model size and data volume only, and therefore cannot quantify data quality or domain mixture, two factors that are increasingly decisive in modern pre-training. This paper develops a three-stage framework for a generalized scaling law. First, the classical power law is fitted to 1,176 standardized Pythia training points by multi-start nonlinear least squares, with row- and block-bootstrap confidence intervals. Second, four quality-augmented candidate forms are compared on a held-out N-D-Q experiment grid; the additive form, in which the quality penalty enters linearly with the quality deficit, is selected, and a large supplementary matrix is identified as contaminated by its reversed quality monotonicity, its violations of the entropy floor, and the sign reversal of its quality coefficient when pooled. Third, the law is unified with centered-log-ratio domain-mixture coordinates, from which equal-loss substitution rates, elasticities, and compute-normalized marginal gains are derived analytically. The classical law attains an RMSE of 0.00011 nat with stable exponents across five data sources; the quality law reaches a hold-out correlation of 0.99; and the predictions remain accurate for models up to ten trillion parameters, where quality improvement proves up to 30,000 times more cost-effective than parameter scaling. The framework thus provides a quantitative basis for compute-constrained resource allocation in LLM pre-training.
Keywords: Large language model; Scaling law; Data quality; Domain mixture; Marginal substitution

1 INTRODUCTION

Large language models (LLMs), exemplified by LLaMA and similar foundation models [1], have become the dominant approach to natural language processing and to a growing range of downstream tasks. Their progress is driven primarily by scaling: for a fixed compute budget, enlarging the number of model parameters N and the volume of training tokens D reliably reduces cross-entropy loss [2]. State-of-the-art training runs, however, demand ever-larger amounts of compute, so using a given compute budget efficiently has become a problem of both theoretical and practical importance.

The empirical study of scaling was pioneered by Kaplan et al. [3], who reported that test loss follows approximate power laws in model size and data size. Hoffmann et al. refined these laws and showed that earlier models were substantially over-sized relative to their data [2], yielding the compute-optimal Chinchilla allocation. Once unique tokens become scarce, Muennighoff et al. quantified the loss penalty of token reuse [4]. A parallel line of work treats data as a heterogeneous mixture rather than a homogeneous stream: Xie et al. proposed DoReMi [5], which reweights domains using a small reference run, and Ye et al. formulated data mixing as a regression problem [6]. A third line of work raises the information content of the training data itself through quality-based pruning [7], deduplication [8], curated synthetic textbooks [9], domain-specific scientific corpora [10], and web-scale cleaning [11].

Despite this progress, two limitations remain. First, most scaling laws are still two-dimensional functions of N and D; data quality Q and domain composition p have been studied separately and have not been integrated into a single loss surface, so the marginal trade-offs among the four dimensions remain unmeasured. Second, evidence is frequently drawn from a single model family, and mis-calibrated supplementary data are seldom screened before estimation. To close these gaps, this paper builds on the standardized logs of the Pythia suite and constructs a three-stage generalized scaling law $L(N, D, Q, p)$ [12]. The classical law is estimated and cross-validated over five data sources; a quality term is selected by held-out competition while a contaminated matrix is identified; and domain proportions are integrated through the centered log-ratio transform [13]. The unified law is then applied to quantify equal-loss substitution and compute-normalized marginal gains, and to test extrapolation toward trillion-parameter scales.

2 MODEL ESTABLISHMENT

This chapter develops the three nested models in turn. The stages are cumulative: each model extends its predecessor by one additional data dimension, and every candidate in Stages 2 and 3 is required to reduce exactly to the classical law of Stage 1 when Q=1, which keeps the framework internally consistent.

2.1 Classical Power-Law Scaling Model

© By the Author(s) 2026, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

Let N denote the model size in billions of parameters and D the number of training tokens in billions. The classical law attributes the cross-entropy loss L, measured in nat, to an irreducible term and two reducible power-law terms:

$$L(N, D) = E + AN^{-\alpha} + BD^{-\beta} \quad (1)$$

Here, $L(N,D)$ is the expected validation loss as a function of model size N and data size D. The constant E is the irreducible loss of natural language—the floor that even infinite parameters and data cannot beat. A and B are the amplitudes (coefficients) of the parameter term and the data term, respectively, while α and β are the corresponding decay exponents that control how fast each reducible term shrinks as N or D grows. Under the standard dense-decoder approximation, the training compute C satisfies

$$C \approx \kappa ND \quad (2)$$

where C is the total training compute in FLOPs and $\kappa \approx 6\times10^{18}$ FLOPs is the compute-per-parameter-token constant under the dense-decoder approximation; the product ND thus maps directly to a compute budget. The five parameters (E,A,B,α,β) are estimated by multi-start nonlinear least squares (NLS):

$$\arg\min_{\theta} \sum_{i=1}^{n} [L_i - E - AN_i^{-\alpha} - BD_i^{-\beta}]^2 \quad (3)$$

Here, $\theta=(E,A,B,\alpha,\beta)$ is the parameter vector being estimated; the index i runs over all observed training points (N_i, D_i, L_i), with n the total sample size; the minimization searches for the θ that minimizes the squared residual between the observed loss L_i and the model prediction. The search is constrained by $E, A, B \geq 0$ and $\alpha < 0, \beta < 3$, and multiple starting points are used to avoid poor local minima. Ninety-five-percent confidence intervals are obtained by two bootstrap schemes: resampling individual rows and resampling whole size groups (blocks).

2.2 Quality-Augmented Generalized Scaling Model

Let $Q \in (0, 1]$ be a normalized quality score of the training data, with $Q = 1$ representing pristine data and lower values indicating noisier or less useful data. Because the functional form of the quality penalty is not known a priori, four competing candidate forms are considered: an additive penalty M_add, a global multiplicative power law M_mult_pow, a power factor on the data term M_data_pow, and an exponential factor M_data_exp; all four reduce exactly to Eq. (1) at $Q = 1$. The additive candidate is

$$L(N, D) = E + AN^{-\alpha} + BD^{-\beta} + G(1 - Q) \quad (4)$$

Here, G is the loss penalty per unit quality deficit, measured in nat: when data quality drops from $Q = 1$ to $Q < 1$, the loss increases by $G(1-Q)$. The additive form was ultimately selected because it gives the best hold-out fit while satisfying the exact nesting condition. The candidates are compared on a hold-out split that is independent of the estimation set, with the Akaike information criterion used as an auxiliary index:

$$AIC = n\ln\frac{SSR}{n} + 2k \quad (5)$$

Here, n is the sample size, SSR is the residual sum of squares, and k is the number of free parameters in the model; AIC balances goodness of fit (low SSR) against model complexity (low k), penalizing over-parameterized models. Before pooling any supplementary source, three contamination diagnostics are imposed: within fixed (N, D) groups the Spearman correlation $\rho(L, Q)$ must be negative; losses must remain above the entropy floor E; and the estimated sign of the quality coefficient must be stable across sources. A source that fails these diagnostics is excluded from the main estimation.

2.3 Unified Generalized Scaling Model with Domain Mixture

Let $p = (p_1, ..., p_K)^T$ be the vector of domain proportions, where each p_k is the fraction of training data coming from domain k, and the proportions are constrained to sum to 1 ($\Sigma\, p_k = 1$). Direct regression on p is ill-defined because the coordinates are mutually dependent, so the centered log-ratio (CLR) transform is applied with smoothing constant ε:

$$z_k(\boldsymbol{p}) = \ln\frac{p_k + \varepsilon}{g(\boldsymbol{p})}, \quad g(\boldsymbol{p}) = \prod_{k=1}^{K} (p_k + \varepsilon)^{\frac{1}{K}} \quad (6)$$

Here, $z_k(p)$ is the CLR coordinate of domain k, $g(p)$ is the geometric mean of the smoothed coordinates ($p_k + \varepsilon$), and K is the total number of domains; the coordinate of a preselected reference domain is removed to avoid singularity (since the CLR coordinates sum to zero). Let b denote the vector of CLR slopes and p_{ref} a reference composition. The unified generalized law is then

$$L(N, D, Q, \boldsymbol{p}) = E + AN^{-\alpha} + BD^{-\beta} + G(1 - Q) + \boldsymbol{b}^T[\boldsymbol{z}(\boldsymbol{p}) - \boldsymbol{z}_{ref}] \quad (7)$$

Here, the term $b^T[z(p) - z_{ref}]$ captures the domain-mixture effect: each component of b measures how much shifting one unit of CLR coordinate for a given domain changes the loss, and the reference term z_{ref} anchors the mixture perturbation at zero on average across the training recipes. The slope vector b is estimated with the quality term kept explicit, which prevents the mixture effect and the quality effect from being confounded. The unified law also yields, in closed form, the substitution between quality and parameters at equal loss: holding D fixed and equating the losses before and after a quality improvement ΔQ gives

$$N_2 = [N_1^{-\alpha} - \frac{G\Delta Q}{A}]^{-\frac{1}{\alpha}} \quad (8)$$

Here, N_1 is the original parameter count and N_2 is the equivalent parameter count that achieves the same loss after a quality improvement ΔQ; the formula tells us how many extra parameters are needed to match the loss reduction obtained by cleaning the data. This substitution is feasible only when $N_1^{-\alpha} > G\Delta Q/A$. The infeasibility boundary is

$$N_{crit} = \left(\frac{G\Delta Q}{A}\right)^{-\frac{1}{\alpha}} \tag{9}$$

where N_{crit} is the largest model size for which a quality gain ΔQ can still be offset by increasing parameters; beyond it the required quality gain would exceed the entire parameter term, so the marginal substitution rate diverges. The information flow among the three stages and their data interfaces is summarized in Figure 1.

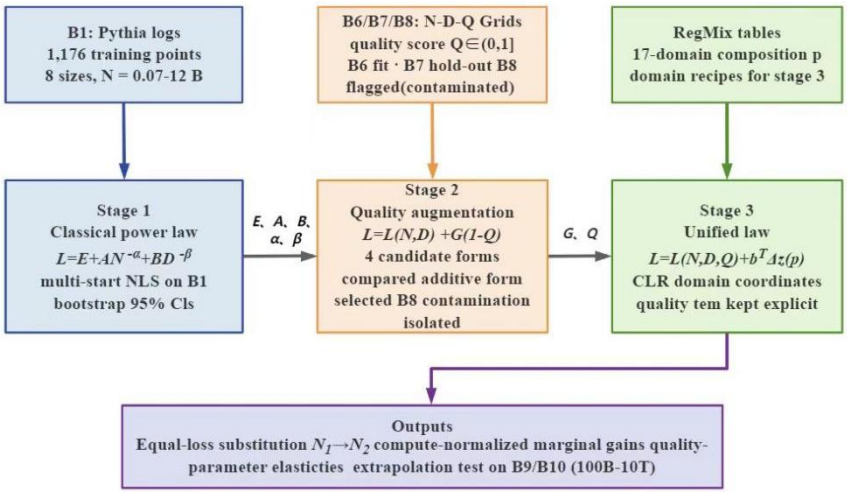

Figure 1 The Three-Stage Generalized-Scaling-Law Framework and Its Data Interfaces

3 RESULTS AND ANALYSIS

The experimental data come from three public sources: the Pythia training logs of the EleutherAI Pythia Scaling Suite (https://github.com/EleutherAI/pythia), the Epoch AI model database (https://epoch.ai/data), and the RegMix mixture tables (https://github.com/HazyResearch/regmix). Throughout this section the data sources are labeled B1-B10: B1 denotes the 1,176 Pythia training points, B2 the Cerebras logs, B3 independent training trajectories, B4 a twelve-family scaling baseline, B5 previously published scaling data, B6-B8 the semi-synthetic N-D-Q experiment grids, and B9-B10 the large-model reference points used in extrapolation. Sources B6-B8 are calibration-based experiment grids rather than direct observations, and this distinction is preserved in every conclusion below.

3.1 Classical Law Estimation and Cross-Source Validation

Fitting the classical model of Section 2.1 to the 1,176 B1 points yields the estimates in Table 1.

Table 1 Classical-Law Parameter Estimates and 95% Bootstrap Confidence Intervals

Parameter	Estimate	95% CI, row bootstrap	95% CI, block bootstrap
E	1.6898	[1.6897, 1.6899]	[1.6896, 1.6900]
A	0.3540	[0.3539, 0.3541]	[0.3538, 0.3542]
α	0.3400	[0.3399, 0.3400]	[0.3399, 0.3402]
B	1.2403	[1.2402, 1.2404]	[1.2401, 1.2405]
β	0.2799	[0.2798, 0.2799]	[0.2798, 0.2800]

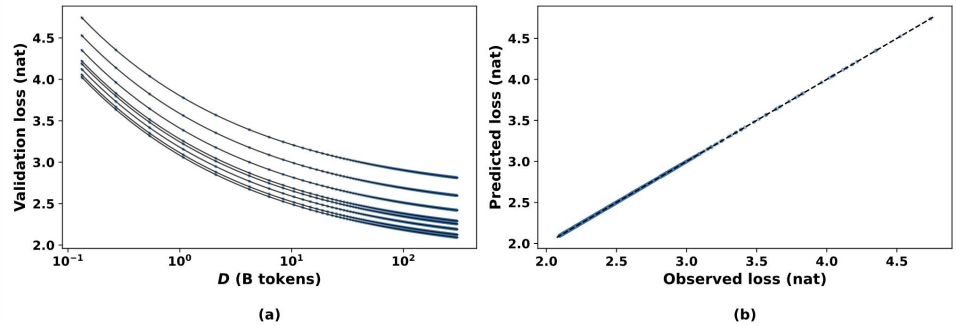

Figure 2 Classical-Law Fit on B1

The classical model demonstrates an exceptional fit with an R^2 of 1.000 and an RMSE of 1.1×10^{-4} nat, while residual analysis confirms a zero mean and no systematic deviation across any size group. Specifically, the residuals exhibit negligible rank-order dependence on both logarithmic model size (lg N) and logarithmic training data size (lg D), as evidenced by minimal Spearman's rank correlation coefficients (ρ) of +0.01 and +0.06, indicating that the error terms are randomly distributed. Furthermore, both bootstrap schemes yield narrow confidence intervals, and fixing parameter E at

1.69 leaves the remaining four parameters virtually unchanged, confirming a well-identified and stable surface. Figure 2 visually corroborates these statistical findings by displaying the tightly fitted curves alongside the observed-versus-predicted scatter plot.

The parameter estimation scheme demonstrates high robustness, as fixing parameter E at its reference value of 1.69 leaves the remaining four parameters virtually unchanged, indicating a well-identified surface. Figure 2(a) visually confirms that the power-law form accurately captures the training dynamics across eight Pythia model sizes, showing validation loss decreasing as training tokens scale from 0.1B to 200B. Furthermore, Figure 2(b) validates the model's predictive accuracy, with all 1,176 observed losses aligning almost perfectly along the 45-degree diagonal with their predicted values, demonstrating the absence of systematic bias. Finally, cross-source validation (summarized in Table 2 and Figure 3) is conducted to determine whether this estimated shape can successfully generalize beyond the B1 dataset.

Table 2 Cross-Source Validation Summary

Source	α	β	RMSE (nat)	Shape r
B1 Pythia (main)	0.340	0.280	0.0001	1.00
B3 trajectories	—	0.280	0.004	1.00
B2 Cerebras (free)	0.336	0.277	0.011	—
B4 twelve families	—	—	0.193	0.98
B5 published points	—	—	0.200	0.91

The D exponent estimated from the B3 trajectories (0.2797) differs from the B1 value by only 0.0002, and the B1 parameters predict the B3 points with RMSE 0.004. The B2 family, fitted freely, yields exponents (0.336, 0.277) that match the B1 exponents despite a different absolute level. After a single global level offset, the B4 and B5 points give shape correlations of 0.98 and 0.91. The decay exponents therefore generalize across model families and published studies, while the level offsets capture family-specific training settings.

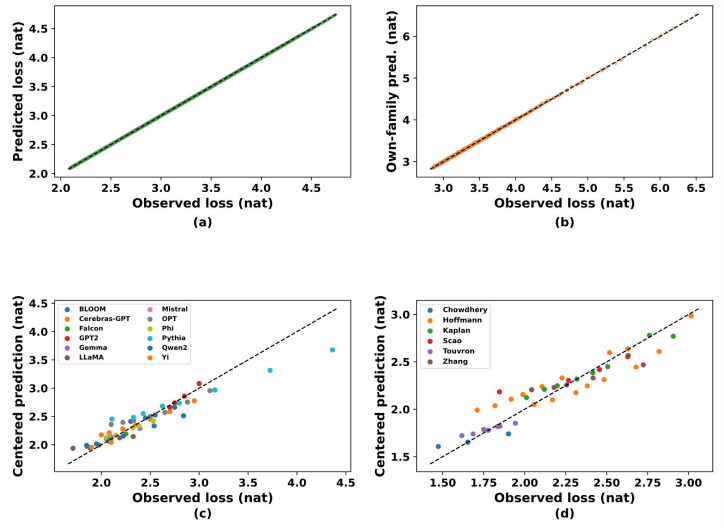

Figure 3 Cross-Source Validation

Figure 3 provides the visual counterpart to Table 2. Panel (a) shows the B3 independent training trajectories: each point is an observed loss, and the solid curve is the B1-fitted prediction; the near-perfect overlap confirms that the D-exponent estimated on Pythia transfers to an unrelated training run. Panel (b) displays the B2 Cerebras model family fitted freely with its own level offset; the scatter along the diagonal confirms that the same (α, β) shape holds even though the absolute loss level differs. Panels (c) and (d) extend the comparison to the B4 twelve-family baseline and the B5 published points, respectively: after removing a single global level offset per source, the centered predictions track the observations closely (shape correlations 0.98 and 0.91), demonstrating that the power-law exponents are universal while the level offset absorbs source-specific training details.

3.2 Quality Augmentation and Contaminated-Data Identification

The four quality candidates were estimated on the 360-point B6 grid ($Q \in \{0.05, 0.3, 0.7, 1\}$) and tested on the 90 new B7 grid points at $Q = 0.5$ and 0.7; the results are listed in Table 3.

Table 3 Quality-Model Competition on B6 (Estimation) and B7 New Points (Hold-Out)

Candidate model	B6 RMSE (nat)	B7 RMSE (nat)	B7 r
M_add (free)	0.059	0.042	0.992
M_add (fixed classical)	0.076	0.060	0.990
M_mult_pow (free)	0.065	0.053	0.990

Candidate model	B6 RMSE (nat)	B7 RMSE (nat)	B7 r
M_data_pow (free)	0.073	0.058	0.987
M_data_exp (free)	0.062	0.049	0.988

M_add achieves both the smallest hold-out RMSE (0.042 nat) and the highest correlation (0.992), and its Q = 1 face coincides with the B1 classical surface to a median offset of 2×10^{-4} nat, satisfying the nesting requirement exactly. Figure 4 shows the fitted quality curves, the hold-out scatter, and the contrast with source B8.

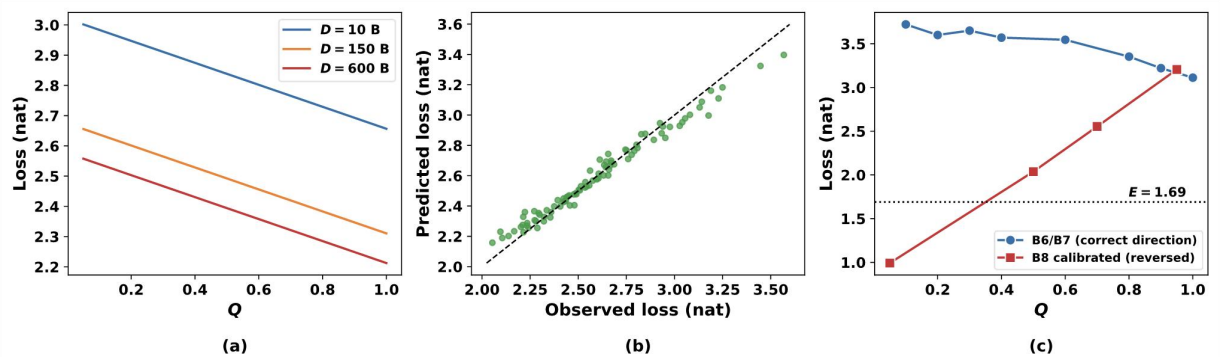

Figure 4 Quality Augmentation

Figure 4(a) plots the fitted loss-quality curves for three representative model sizes (N = 0.10, 0.50, and 1.0 B) at fixed D: as Q rises from 0.05 toward 1, the loss decreases monotonically, and the additive penalty $G(1 - Q)$ produces straight, parallel lines whose common slope is G. Figure 4(b) is the hold-out scatter on the 90 B7 points: predicted losses are plotted against observed losses, and the tight alignment along the diagonal (correlation 0.992) shows that the selected model generalizes to unseen quality levels. Figure 4(c) exposes the contamination diagnostic on source B8: unlike B6/B7, the B8 calibrated segment rises as Q approaches 1 (the opposite of the expected monotone decrease), and more than half of its points fall below the entropy floor E = 1.69 marked by the dashed line; this reversal is why B8 is excluded from the main estimation.

3.3 Marginal Substitution and Large-Scale Extrapolation

On a reference grid covering $N \in [0.125, 700]$ B, the loss elasticities with respect to N, D, and Q at N = 1.4 B, D = 200 B are approximately (−0.13, −0.10, −0.18) for Q = 0.4 and (−0.18, −0.13, −0.61) for Q = 1; the quality lever thus becomes the most elastic one as the data depart from pristine quality. After converting every lever to the same compute unit (10^{21} FLOPs), the parameter and quality routes are compared in Figure 5.

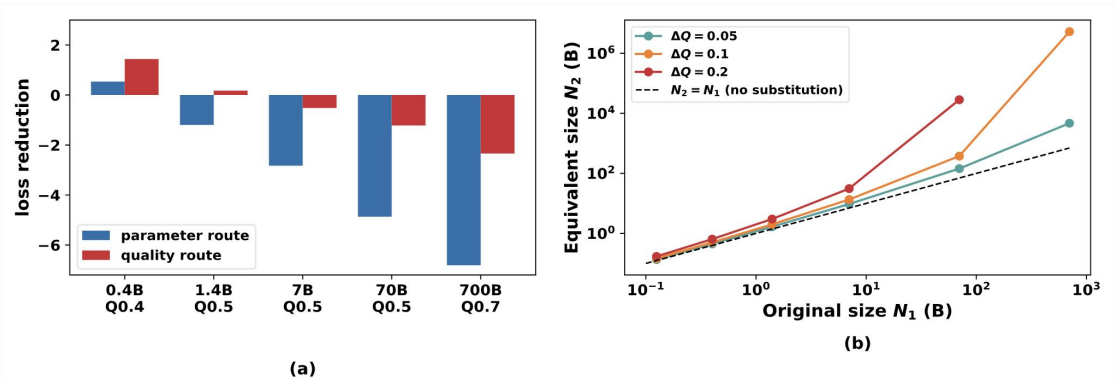

Figure 5 Marginal Trade-Offs

Figure 5(a) compares the two improvement routes on a common compute budget: each pair of bars shows the loss reduction achieved by adding parameters (blue, "parameter route") versus improving data quality (red, "quality route") for the same 10^{21} FLOPs of additional compute, at five model scales from 0.4 B to 700 B. At the 0.4 B scale the quality route is already about 8 times more cost-effective, and the advantage grows to 2.9×10^4 times at the 700 B scale: as N grows the parameter term $AN^{-\alpha}$ saturates while the quality penalty $G(1 - Q)$ remains linearly available, so data cleaning increasingly dominates. Figure 5(b) translates the same comparison into equal-loss parameter equivalents: each curve gives the model size N_2 that achieves the same loss as the original size N_1 after a quality improvement ΔQ; the dashed diagonal marks no substitution ($N_2 = N_1$). The curves bend sharply upward with scale, meaning that a fixed quality gain replaces ever-larger parameter increases at large models. The equal-loss equivalents are listed in Table 4.

Table 4 Equal-Loss Parameter Equivalents for Quality Improvements at Fixed D

N1 (B)	N2, ΔQ = 0.05 (B)	N2, ΔQ = 0.1 (B)	N2, ΔQ = 0.2 (B)
0.125	0.135	0.146	0.171
0.4	0.448	0.503	0.646
1.4	1.667	2.006	3.021
7	9.525	13.438	31.106
70	144.0	375.4	28344
700	4681	5.2×10^{6}	infeasible

Table 4 explicitly quantifies the equal-loss parameter equivalents for various quality improvements at a fixed data size D. A quality improvement ΔQ = 0.1 substitutes for parameter increases ranging from 0.10 B (at $N_1 = 0.4$ B) to 5.2×10^6 B (at $N_1 = 700$ B); the effect grows sharply with scale, and for ΔQ = 0.2 the substitution becomes infeasible beyond the boundary $N_{crit} = 105$ B. Extrapolation validity was tested on the 128 B9/B10 points covering models from 100 B to 10 T.

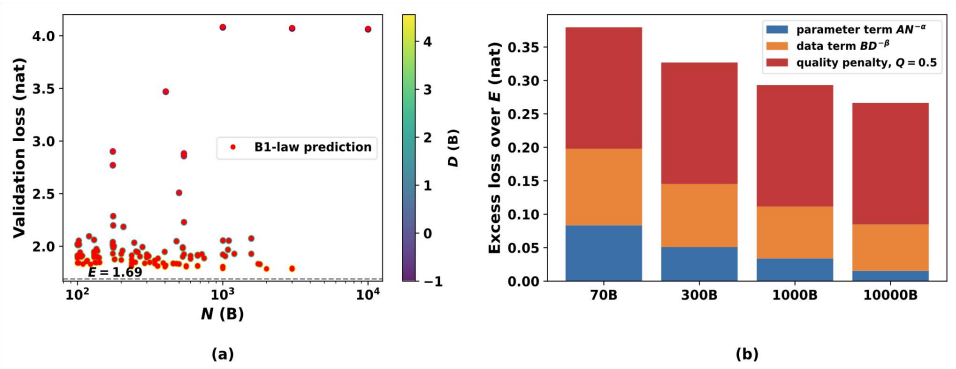

Figure 6 Large-Scale Extrapolation

Figure 6(a) tests the unified law on the 128 B9/B10 large-model reference points spanning 100 B to 10 T parameters. Each point is colored by its training-token count D; the B1-fitted prediction curve (black, labeled "B1-law prediction") passes through the cloud of observed losses, and the points cluster tightly around it with RMSE = 0.0011 nat and correlation 1.00. Even under-trained models such as DLRM-2021 (N = 1,000 B, D = 0.1 B) sit on the same curve, confirming that the law generalizes across the entire parameter range. Figure 6(b) decomposes the excess loss (loss minus E) into the parameter term $AN^{-\alpha}$ (blue), the data term $BD^{-\beta}$ (orange), and the quality penalty for $Q = 0.5$ (red), stacked across model sizes from 70 B to 10,000 B. The combined share of the data and quality terms rises from 48% at 70 B to 68% at 10,000 B, while the parameter term shrinks proportionally; this shows that future loss reductions will come increasingly from data-side improvements (more tokens or higher quality) rather than from simply scaling the model size, see Table 5.

Table 5 Selected CLR Domain Coefficients of the Unified Law

Domain	CLR coefficient (nat)	Direction
ubuntu_irc	−0.031	favorable
dm_mathematics	−0.029	favorable
stackexchange	−0.023	favorable
nih_exporter	+0.015	unfavorable
philpapers	+0.017	unfavorable
enron_emails	+0.046	unfavorable

The domain coefficients are also economically interpretable: adding ubuntu_irc, dm_mathematics, or stackexchange shifts excess loss downward by 0.023 to 0.031 nat per unit CLR coordinate, whereas enron_emails shifts it upward by 0.046 nat. The mean perturbation over the training recipes remains zero, so these effects describe reallocation rather than a global shift in the loss surface.

4 CONCLUSION AND OUTLOOK

This paper addressed the compute-constrained allocation problem of LLM pre-training through a three-stage generalized scaling law. The classical power law was estimated on 1,176 Pythia points with multi-start NLS and dual bootstrap intervals, and its exponents were validated over five independent sources. Four quality-augmented candidates were then compared on a held-out grid; the additive model $L = E + AN^{-\alpha} + BD^{-\beta} + G(1 - Q)$ was selected, and a contaminated source was identified and isolated. A unified law further incorporated CLR domain-mixture coordinates and provided analytic equal-loss substitution rates and compute-normalized marginal gains. Extrapolation tests up to ten-trillion-parameter models retained an RMSE of 0.001 nat, and quality investment proved up to 3×10^4 times more cost-effective than parameter scaling at large scale. The framework thus supplies a practical quantitative tool for allocating parameters, data volume, data quality, and domain mixture under a fixed compute budget.

The additive quality term is a first-order approximation whose validity beyond the observed quality range remains to be checked, and the CLR slopes currently come from a single mixture study. Future work will introduce higher-order mixture interactions, estimate the quality-cost curve from direct cleaning experiments, and extend the law to inference-time and test-time compute.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

ACKNOWLEDGMENTS

GengYin Li, Xue Yuan and XueTong Zang contributed equally to this work by perfomming an equivalent amount of experimental tasks and data analysis.

REFERENCES

[1] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv, 2023. DOI: 10.48550/arXiv.2302.13971.
[2] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022: 30016-30030.
[3] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv, 2020. DOI: 10.48550/arXiv.2001.08361.
[4] Muennighoff N, Rush A, Barak B, et al. Scaling data-constrained language models. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 50358-50376.
[5] Xie S M, Pham H, Dong X, et al. DoReMi: reweighting data for faster language model pre-training. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 69798-69818.
[6] Liu Q, Zheng X, Muennighoff N, et al. RegMix: data mixture as regression for language model pre-training. arXiv, 2024. DOI: 10.48550/arXiv.2407.01492.
[7] Sorscher B, Geirhos R, Shekhar S, et al. Beyond neural scaling laws: beating power law scaling via data pruning. NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022: 19523-19536.
[8] Lee K, Ippolito D, Nystrom A, et al. Deduplicating training data makes language models better. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 8424-8445. DOI: 10.18653/v1/2022.acl-long.577.
[9] Gunasekar S, Zhang Y, Aneja J, et al. Textbooks are all you need. arXiv, 2023. DOI: 10.48550/arXiv.2306.11644.
[10] Taylor R, Kardas M, Cucurull G, et al. Galactica: a large language model for science. arXiv, 2022. DOI: 10.48550/arXiv.2211.09085.
[11] Penedo G, Kydlíček H, Ben Allal L, et al. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. arXiv, 2024. DOI: 10.48550/arXiv.2406.17557.
[12] Biderman S, Schoelkopf H, Anthony Q, et al. Pythia: a suite for analyzing large language models across training and scale. Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023, 202: 2397-2430.
[13] Kudugunta S. Fixing the foundation: principled approaches to data curation and scaling. Seattle: University of Washington, 2026.