InfoLaw: Information Scaling Laws for Large Language Models with Quality-Weighted Mixture Data and Repetition

Fengze Liu [*1] **Weidong Zhou** [*1] **Binbin Liu** [1] **Ping Guo** [1] **Zijun Wang** [1 2] **Bingni Zhang** [1] **Yifan Zhang** [1] **Yifeng Yu** [1] **Xiaohuan Zhou** [1] **Taifeng Wang** [1]

Abstract

Upweighting high-quality data in LLM pretraining often improves performance, but in data-limited regimes, especially under overtraining, stronger upweighting increases repetition and can degrade performance. However, standard scaling laws do not reliably extrapolate across mixture recipes or under repetitions, making the selection for optimal data recipes at scaling underdetermined. To solve this, we introduce **InfoLaw** (Information Scaling Laws), a data-aware scaling framework that predicts loss from consumed tokens, model size, data mixture weights, and repetition. The key idea is to model pretraining as information accumulation, where quality controls information density and repetition induces scale-dependent diminishing returns. We first collect the model performance after training on datasets that vary in scale, quality distribution, and repetition level. Then we build up the modeling for information so that information accurately predicts those model performance. InfoLaw predicts performance on unseen data recipes and larger-scale runs (up to 7B, 425B tokens) with 0.15% mean and 0.96% max absolute error in loss, and it extrapolates reliably across overtraining levels, enabling efficient data-recipe selection under varying compute budgets.

1. Introduction

Training large language models (LLMs) requires access to high-quality data (Brown et al., 2020a; Chowdhery et al., 2023). However, the availability of high-quality data is severely limited (Villalobos et al., 2024), and in the data-constrained settings, upweighting higher-quality data inevitably increases repetition, which has been shown to impair performance when excessive (Muennighoff et al., 2023). This issue is further exacerbated by the widespread adoption of overtraining (Touvron et al., 2023; Yang et al., 2025)—a strategy that reduces inference costs compared to the compute-optimal regime (Hoffmann et al., 2022).

To address the shortage of high-quality data as model scale increases, a common compromise is to incorporate lower-quality data, thereby reducing the repetition of high-quality samples. Intuitively, high-quality data provides greater performance gains than low-quality data upon first exposure, but as repetition increases, the marginal benefit decays—eventually approaching that of unseen low-quality data. However, the optimal balance between quality and repetition remains unclear. A standard approach for identifying optimal mixing strategies is to run smaller-scale experiments and extrapolate performance to larger compute budgets using scaling laws (OpenAI et al., 2024; Hoffmann et al., 2022; Chowdhery et al., 2023). Yet, as shown in Figure 1, under conditions of data repetition, standard scaling laws fail to reliably predict model performance at scale (Hernandez et al., 2022; Muennighoff et al., 2023). Moreover, they do not generalize across different mixing strategies , necessitating grid searches over data recipes—an approach that is costly even at small scales.

In this paper, we study the problem of scaling large language models in a data-aware regime, where training data consists of a heterogeneous mixture with varying quality levels, and each quality level is repeated to different extents. We introduce a theoretical framework, the InfoLaw, which accounts for both the scaling effects of mixture weights and the impact of repetition. Our formulation views training as a process of accumulating information from the dataset, with model performance determined by the total information gained by the end of training. At each step, the information gain is modeled as the sum of contributions from different quality ranges. Within each quality range, the gain depends on two factors: an information density function, parameterized by quality (with higher quality assigned higher density), and an exponential decay term that captures the interactions between model scale, data scale, and repetition level.

To fit the parameters of the InfoLaw, we construct a suite of

[1]ByteDance [2]UC Santa Cruz. Correspondence to: Fengze Liu <fengze.liu@bytedance.com>, Weidong Zhou <zhouweidong.66@bytedance.com>.

Preprint. May 5, 2026.

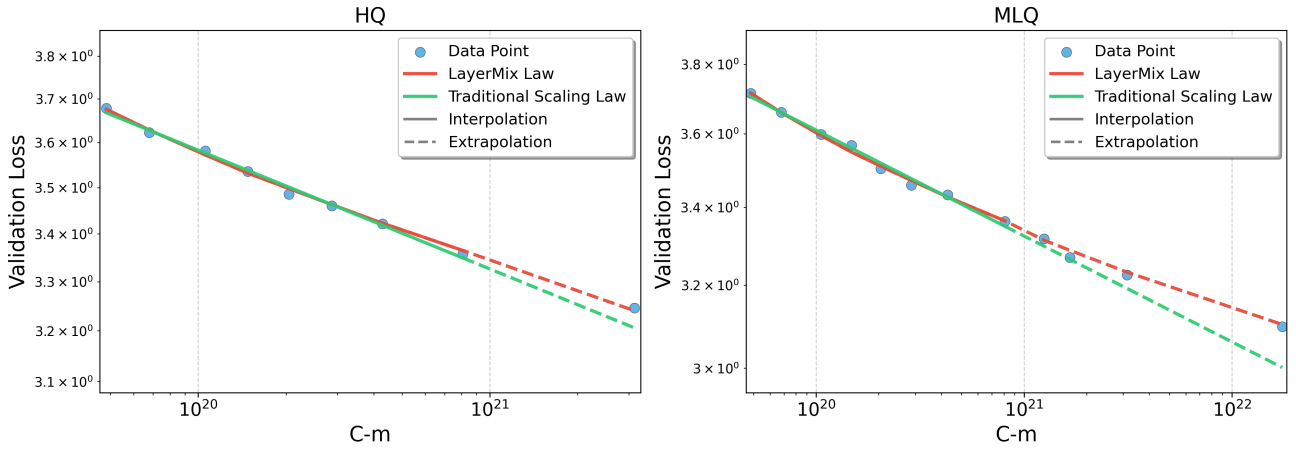


Figure 1. Validation loss versus compute C_m in the loss–C view under LayerMix data with repetition. Curves are fit on 252M–1.2B and extrapolated to larger models. The traditional scaling law mis-extrapolates under repetition, while InfoLaw tracks both interpolation and extrapolation across recipes (HQ and MLQ).

datasets that vary along three axes: scale, quality, and repetition level. Specifically, we partition the source dataset into buckets according to quality scores, and then sample from each bucket with different weights, a procedure we refer to as LayerMix sampling. Following the data-constrained setting, the source dataset is first downsampled to the target scale to ensure stable repetition effects. We then train 9 models ranging from 252M to 1.2B parameters from scratch, each under the same 3.6x over-trained ratio (Gadre et al., 2024). For each model, we construct three datasets with distinct LayerMix sampling configurations, resulting in 27 total training runs. Model performance is evaluated as the average perplexity across five downstream tasks. Finally, we fit the InfoLaw to these results, estimating the parameters that best capture the relationship between information gain and observed performance.

We evaluate the generalization of InfoLaw along three axes: (i) unseen mixture recipes (new LayerMix sampling weights), (ii) larger compute scales, and (iii) a higher overtraining ratio ($25\times$). Across these settings, InfoLaw accurately predicts loss on unseen recipes and scales up to a 7B model trained on 425B tokens, with 0.15% mean and 0.96% maximum absolute error. Moreover, using the fitted law we search over candidate mixtures and identify a data recipe for a 2.5B model that outperforms four randomly sampled baselines without additional training runs. The same parameters also extrapolate well to the $25\times$ overtraining regime.

2. Related Work

Scaling Laws Empirical studies have shown that transformer language models exhibit predictable power-law scaling with model size and training data (Hestness et al., 2017; Vaswani et al., 2017; Chowdhery et al., 2023; Radford et al., 2019), which has motivated the development of many large-scale systems, including dense models (Brown et al., 2020b; Rae et al., 2021; Grattafiori et al., 2024) and mixture-of-experts variants (DeepSeek-AI et al., 2025; Yang et al., 2025; Fedus et al., 2021). Compute-based scaling laws further formalize how to allocate model capacity and training tokens under a fixed compute budget: Hoffmann et al. (2022) characterized the compute-optimal regime, while subsequent work explored alternative allocations and the interaction between compute C and optimization choices such as batch size and learning rate (Kaplan et al., 2020; DeepSeek-AI et al., 2024).

In parallel, training smaller models on substantially more tokens than the compute-optimal point has become increasingly common for efficiency and deployment reasons (Touvron et al., 2023; Yang et al., 2025). Sardana et al. (2024) extended the Chinchilla framework by incorporating factors such as data quality and inference requirements, and Gadre et al. (2024) showed that scaling relations can remain reliable in overtrained regimes. For predicting downstream performance, Isik et al. (2025) studied how downstream metrics scale after fine-tuning, and Schaeffer et al. (2023) linked non-linear evaluation metrics to perplexity, supporting perplexity as a more stable proxy than earlier observations of emergent/unstable metrics (Wei et al., 2022).

Data-Aware Scaling Traditional scaling laws often assume effectively unlimited data, but in practice high-quality data is scarce and therefore frequently upsampled (Lin et al., 2022). Under repetition, prior work reports diminishing returns and, beyond some point, performance degradation when upsampling subsets or repeating datasets (Hernandez et al., 2022; Muennighoff et al., 2023). At the same time, Xue et al. (2023) suggests that, in certain regimes,

continuing to train on repeated data can still be preferable to stopping early, highlighting that the effect of repetition is non-trivial and not captured by classical laws. More recently, Chen et al. (2025) studied how scaling interacts with data density, providing a finer-grained view in limited-data regimes.

A separate line of work uses scaling laws to optimize data recipes. Ye et al. (2025) incorporated mixture weights into loss prediction, and Kang et al. (2025) argued that optimal mixing can be model-scale dependent. Liu et al. (2024) uses proxy models to search mixture ratios without training the full-scale model, while Gu et al. (2024); Que et al. (2024) leverage scaling insights in continued pre-training and domain-mixture design; Chang et al. (2024) further analyzes the interaction between scaling and data quality. In contrast, our goal is to predict loss under quality-weighted mixtures *with explicit repetition*, enabling extrapolation across both mixture recipes and repetition levels.

3. Limitations of Conventional Scaling Laws

In this section, we reveal and substantiate a critical limitation of conventional scaling laws in the context of data repetition and quality selection. First, we introduce the LayerMix sampling function in section 3.1, to imitate real scenario where the data is a mixture of different quality and repetition degrees. Next, we compare the relationship between the model's loss L and amount of compute C in cases with and without repetition in section 3.2, and the results show that the traditional scaling law performs well on data without repetition

3.1. LayerMix Sampling Function

Source Data We obtain our training corpora from Common Crawl (Common Crawl Foundation), following Penedo et al. (2023) and obtain 15T English tokens. We ran global fuzzy deduplication across all snapshots to ensure there is no repeat data in the corpora. The final dataset contains 3.7T token. Details are in Appendix A.

Training Data Sampling We assign each document a quality score following Liu et al. (2025): we apply two quality classifiers (Penedo et al., 2024; Li et al., 2025) and take the average of their normalized scores. We rank all documents by this score and partition the corpus into six buckets by percentile: 0–5%, 5–20%, 20–40%, 40–60%, 60–80%, and 80–100%.

We then define a LayerMix sampling function $H(w, K, S, B)$ to construct a packed training set. Here S is the number of tokens in the source corpus to sample from, K is the total number of tokens in the packed training set (we use one-epoch training to avoid additional epoch-induced repetition), $w = [w_0, \ldots, w_5]$ with $\sum_d w_d = 1$ specifies the target token proportions of the six buckets in the training set, and $B = [B_0, \ldots, B_5]$ specifies the bucket proportions in the source corpus (in our setting $B = [0.05, 0.15, 0.20, 0.20, 0.20, 0.20]$).

For bucket d, the training set contains $K_d = w_d K$ tokens sampled from $S_d = B_d S$ source tokens. Let $M_d = \min(K_d, S_d)$ denote the number of unique (non-repeated) tokens from bucket d that appear in the packed training set, and define the average repetition factor as $R_d = K_d/M_d = w_d K/M_d$, so $R_d = 1$ when $K_d \le S_d$ and $R_d > 1$ otherwise. The full packing procedure is given in Appendix C.

By varying (w, K, S), LayerMix produces datasets with different scale, quality mixture, and repetition. We enforce $w_d \ge w_{d+1}$ to keep higher-quality buckets more represented. We use five preset mixtures (HQ, MHQ, MQ, MLQ, LQ; Table 1), and set $w_5 = 0$ to drop the lowest 20% bucket. Unless stated otherwise, we set $K = S$ to isolate the repetition effects induced by w.

3.2. Traditional Scaling Law Between Loss and Amount of Compute

We compare the relationship between model loss L and total compute C under regimes with and without repetition in an overtrained setting. Specifically, under the compute-optimal scheme, $C_{opt} = N_{opt} K_{opt}$, where K is the consumed tokens, N is the non-embedding FLOPs per token as defined in DeepSeek-AI et al. (2024) and N_{opt}, K_{opt} is the Chinchilla-optimal pair. Then in the overtrained setting, following Gadre et al. (2024), we set $K_m = \sqrt{m} K_{opt}$, $N_m = \frac{1}{\sqrt{m}} N_{opt}$, $C_m = K_m N_m$ with $m = 3.6$. And Gadre et al. (2024) shows that the the Loss–Compute relation preserves the fitted exponent for models trained with the same overtraining factor m.

The model loss is collected by training on datasets sampled with LayerMix parameters HQ and MLQ, see details in Table 1. Dataset HQ has more high quality data but with more repitition, while MLQ has more diverse data with less repitition. We then visualize the relationship between compute C_m and model loss L in the loss–C_m view in Figure 1. Here L is the average perplexity over five downstream tasks—HellaSwag (Zellers et al., 2019), ARC-E/ARC-C (Clark et al., 2018), MMLU (Hendrycks et al., 2021), and TriviaQA (Joshi et al., 2017). Following Schaeffer et al. (2023), we convert downstream accuracies into perplexity to obtain a smoother scaling signal. As shown in Figure 1, although a conventional power-law scaling curve can interpolate within the fitting regime (252M–1.2B), it systematically mis-extrapolates as C_m increases under LayerMix data with repetition, yielding overly optimistic loss reductions. This failure appears consistently across representative mixture recipes, indicating that compute alone is

Table 1. Preset LayerMix sampling weights and Searched optimal sampling weights for 2.5B model.

Name	$w0$	$w1$	$w2$	$w3$	$w4$	$w5$
HQ (High Quality)	0.80	0.10	0.03	0.03	0.02	0.0
MHQ (Medium-High Quality)	0.66	0.22	0.05	0.03	0.02	0.0
MQ (Medium Quality)	0.48	0.23	0.13	0.07	0.07	0.0
MLQ (Medium-Low Quality)	0.38	0.21	0.20	0.11	0.08	0.0
LQ (Low Quality)	0.24	0.20	0.19	0.18	0.17	0.0
Optimal Recipe of 2.5B model with $m = 3.6$	0.50	0.49	0.01	0.0	0.0	0.0

insufficient to characterize scaling behavior in the presence of quality-weighted mixtures and repetition.

These observations suggest that traditional scaling laws are not reliably predictive under quality-weighted mixture data with repetition, especially for extrapolation. Therefore, we need a modified scaling law that explicitly incorporates both the data quality distribution and the degree of data repetition as core variables.

4. Information Scaling Laws

In this section, we introduce the design of InfoLaw. We treat the training process as gaining information from the dataset and propose to calculate Information as accumulation of information gain throughout the training process, which synthesizes the impacts of data quality, repetition level, model scales and total training tokens, and then build power-law relationship with the model's final validation loss.

4.1. Information Measurement

To build intuition for how repetition interacts with data quality, we compare two 850M runs trained with different LayerMix sampling weights. In the more repetition-heavy recipe (HQ), the top 5% quality bucket is repeated by roughly $16\times$, whereas in a less repetitive recipe (MQ) it is repeated by roughly $10\times$. Empirically, the two runs achieve similar evaluation loss early in training, but the more repetitive run improves substantially more slowly in the later stage and converges to a worse final loss, indicating diminishing returns from repeated exposures. See Appendix E Figure 5(b) for the training-time curves.

Based on this observation, we propose an exponential decay function to model the decreasing information gain of repeated data. Assuming the Information a document i contains is I_i, then the information a language model gets at t-th learning from the document i is:

$$I_{i\_\text{part}}(t, \lambda(N)) = I_i \cdot \lambda(N) e^{-\lambda(N)t} \tag{1}$$

where $\lambda(N)$ is a nonnegative rate parameter that depends on the model's non-embedding FLOPs/token N and is fitted from data.

When a language model learning the document for total T times, the Information learned from the document is:

$$I_{i\_\text{total}}(T, \lambda(N)) = \int_0^T I_{i\_\text{part}}(t, \lambda(N))dt = I_i \cdot (1 - e^{-\lambda(N)T}) \tag{2}$$

Equation 2 captures the principle of diminishing returns in learning: repeated exposure to a document yields progressively smaller gains, causing the total acquired information to saturate and asymptotically approach the document's full information content I_i.

To capture the empirically observed slowdown in marginal gains relative to the total training budget K, we incorporate a logarithmic normalization factor. This formulation is empirically grounded and essential for generalizing the scaling law across orders of magnitude in training volume, as validated in Appendix B.

$$I_{i\_\text{part}}(t, \lambda(N), K) = I_i \cdot \lambda(N) e^{-\lambda(N)t/\log(K)} \tag{3}$$

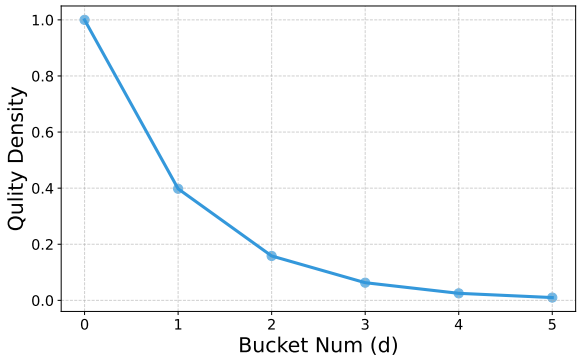
Then the Equation 2 becomes to:

$$\begin{aligned} I_{i,\text{total}}(t, \lambda(N), K) &= \int_0^T I_{i,\text{part}}(t, \lambda(N), K)dt \\ &= I_i \cdot \log(K) \left(1 - e^{-\lambda(N)T/\log(K)}\right) \end{aligned} \tag{4}$$

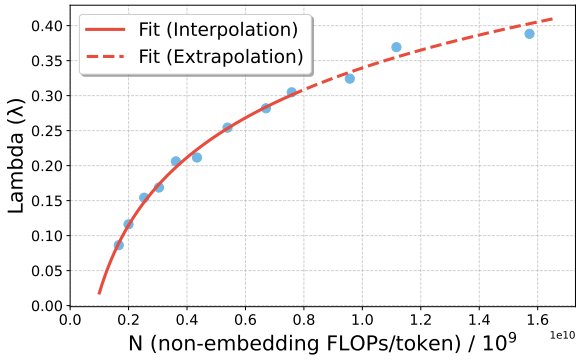
For all the training data, we sum them together as the final Information the language model learned from the training corpora, denoting as info:

$$\begin{aligned} &\text{info}(w, K, S, f, \lambda(N)) \\ &= \sum_d I_d \cdot \log(K) \left(1 - e^{-\lambda(N)R_d/\log(K)}\right) \\ &= \sum_d f_d M_d \log(K) \cdot \left(1 - e^{-\lambda(N)R_d/\log(K)}\right) \end{aligned} \tag{5}$$

where d is the quality bucket number from 0 to 5. I_d is the total Information in d-the bucket, which can be calculated



(a) The fitted quality density function f_d. The quality density is a monotonically decreasing function of the bucket index, meaning buckets with higher-quality data are assigned a higher density value.



(b) The relationship between λ and N with a fitted curve. The blue scattered points represent the observed data. The solid red line shows the fit within the data range, while the dashed line represents the extrapolation.

Figure 2. The fitted function of quality density function and relationship between $\lambda(N)$ and N

by the multiplication of number of unique tokens $M_d = min(w_d K, B_d S)$ and information density f_d, which is a parameterized quality density function. $R_d = \frac{w_d K}{M_d}$ is the average repeat times for the data from the d-th bucket and $\lambda(N)$ is related with N, which are to be fitted from the data.

Equation 5 can be divided into two parts: the first term is $I_d = f_d M_d \log(K)$, it represents the total Information contained in the packed data of the d-th bucket, and the second term is $1 - e^{-\lambda(N) R_d / \log(K)}$, it represents the language model's learning ability on this data when repeated an average of R_d times. And the total Information learned by the language model is the product of these two terms.

We propose Information, a metric computed from LayerMix sampling weights w, train token K and two fitted functions (f_d, $\lambda(N)$), to quantify the knowledge learned during training. Since it is designed to be monotonic with model performance, it enables loss prediction for various training configurations prior to any actual runs. The fitting of f_d and $\lambda(N)$ is described in Section 5.2.

4.2. Information-Loss Power Law

As illustrated in Figure 1, in the loss–C_m view conventional scaling laws are not reliably predictive under quality-weighted mixture data with repetition, with extrapolation errors that grow at larger compute. This motivates replacing compute with a repetition and quality aware effective data signal. In the next section, we show that our Information collapses results across mixture recipes and scales onto a unified power-law curve.

We use the Information proposed in Section 4.1 and plot the L-$info$ figure. As illustrated in Figure 3f, when we replace the traditional computation axis C with our novel metric: Information, the experimental points with different LayerMix sampling weights w, Model non-embedding FLOPs/token N and Train Token K now collapse perfectly onto a single, unified power-law curve, where they were previously scattered and separated.

Then the relationship between the loss L and $info$ can be measured using power-law formulation as:

$$L = \alpha \cdot info^{-\beta} \tag{6}$$

In our experiment, $\alpha = 3.7373$ and $\beta = 0.0441$. We show them in a log-log plot, so it appears as a straight line with a slope of $-\beta$ and an intercept of $\log(\alpha)$.

Like the traditional scaling law (Hoffmann et al., 2022), we can now conduct experiments on small models to compare the advantages and disadvantages of different experimental configurations, and then use our proposed information scaling law to extrapolate the performance of larger models under larger training tokens.

5. FITTING EXPERIMENTS

5.1. Training setup

We train 9 models ranging from 252M to 1.2B on 3 layermix sampling weights HQ, MQ, and LQ, with 3.6x over-trained ratio, resulting in 27 experiment runs in total to collect data for fitting the InfoLaw parameters. We use transformer architecture (Vaswani et al., 2017), SwiGLU (Shazeer, 2020) as the activation function and RoPE embeddings (Su et al., 2024). We use a tokenizer with 250k vocabulary. See Appendix C and Appendix D for details about LayerMix sampling weights, model structure, learning rate and optimizer.

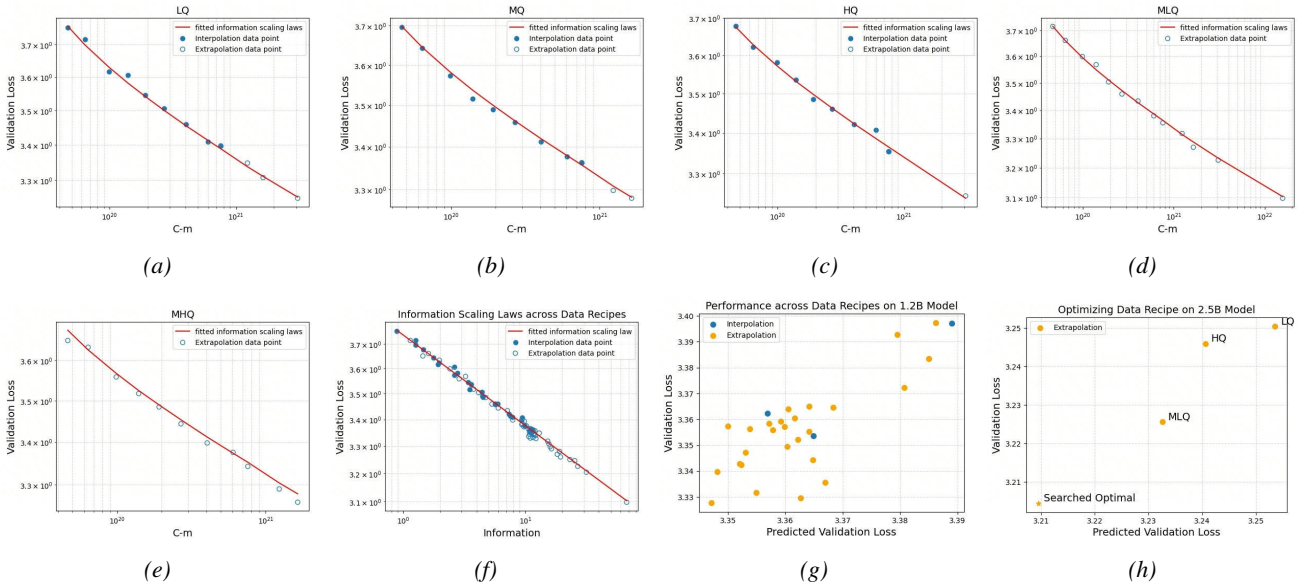


Figure 3. **Verification, Unification, and Application of Information Scaling Laws.** Panels **(a)-(e)** demonstrate that information scaling laws hold independently across varying data quality distributions (LQ to MHQ), consistently following power-law trajectories. **(f)** Illustrates the **Information Scaling Laws**, where diverse data recipes collapse onto a single curve when mapped to the information quantity metric, confirming the universality of the law. **(g)** Validates predictive capability on a 1.2B model, showing strong correlation between predicted and actual validation loss for both interpolation and extrapolation settings. **(h)** Demonstrates optimization on a 2.5B model, where the "Searched Optimal" recipe identified by our framework achieves lower validation loss compared to fixed baselines.

5.2. Fitting the curve

In this section, we introduce how to fit the parameters in InfoLaw to predict the model performance collected in Section 5.1. Since Information $info$ indicates the knowledge learned by the model, we expect larger $info$ to correspond to lower evaluation loss L. Considering that there may exist scale difference between $info$ and model loss L, we choose Spearman correlation ρ_s as the fitting metric, i.e., the object is to find the optimal quality density f and $\lambda(N)$ such that the Spearman correlation between evaluation loss L and $info$ is minimized for all the experiments over N, w :

$$(f^*, \lambda^*) = \underset{f,\lambda}{\operatorname{argmin}} \sum_{N,w} \rho_s\big(L_N, \text{info}(w, K_N, S_N, f, \lambda(N))\big) \tag{7}$$

To prevent from over-fitting, we make some assumption based on naive intuition. For f, as it indicates the quality density, the higher-quality bucket should have larger f. As smaller d corresponds to higher-quality buckets, we define f in the following form to ensure it is a decreasing function:

$$f_d(\theta) = e^{-\theta * d} \tag{8}$$

where θ is a hyperparameter and $\theta > 0$.

$\lambda(N)$ is related to the model's learning capacity, so $\lambda(N)$ should increase as N increases. But we need to find the formula for $\lambda(N)$ related with N so that it can scale to larger N. To do this we first sample 100,000 combinations of θ and $\lambda(N)$ from the parameter space, then select optimal θ^* and λ_N^* based on Equation 7. The fitted quality density $f(\theta^*)$ is shown in Figure 2a with fitted $\theta^* = 0.922$.

Having the $\lambda(N)^*$ values of different models, as is shown in Figure 2b, we try to fit the $\lambda(N)$-N curve. The relationship between $\lambda(N)$ and N is observed to be non-linear, exhibiting rapid growth for smaller N and gradually saturating as N increases. This trend is well-approximated by a logarithmic function. Therefore, we choose the $\lambda(N)$-N curve using following formula:

$$\lambda(N)(a, b) = a \cdot \ln(N) + b \tag{9}$$

Using existing $\lambda(N)^*$, we fit the $\lambda(N)$-N curve in Figure 2b with fitted $a^* = 0.140$, $b^* = 0.018$. To validate this fit, we compute $\lambda(N)^*$ for larger N under the fixed θ^*, and examine whether these values lie on the predicted $\lambda(N)$-N curve. As illustrated in Figure 2b, the results demonstrate strong extrapolation performance, supporting the correctness of our formulation. We compared with different formats of 9 in Appendix G and the log function best fit the trend and extrapolates.

Finally, with $f(\theta^*)$ and $\lambda(N)(a^*, b^*)$, we can calculate the Information for arbitrary layermix sampling weights w, train token K, source token S and model non-embedding

Table 2. The best data recipe for different models and train token

Model	Train Token	Source Token	$w0$	$w1$	$w2$	$w3$	$w4$	$w5$
7B	300B	500B	0.548	0.444	0.004	0.003	0.002	0.000
	500B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	800B	500B	0.439	0.430	0.130	0.001	0.000	0.000
	1000B	500B	0.395	0.387	0.214	0.003	0.001	0.000
1.8B	300B	500B	0.619	0.376	0.004	0.001	0.000	0.000
	500B	500B	0.548	0.444	0.004	0.003	0.002	0.000
	800B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	1000B	500B	0.491	0.487	0.017	0.005	0.000	0.000
1.2B	300B	500B	0.758	0.229	0.012	0.001	0.000	0.000
	500B	500B	0.619	0.376	0.004	0.001	0.000	0.000
	800B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	1000B	500B	0.496	0.492	0.007	0.003	0.002	0.000

FLOPs/token N.

6. Extrapolation

After fitting InfoLaw on the 252M–1.2B models, we evaluate its extrapolation along three axes: unseen mixture recipes, larger model scales, and a higher overtraining ratio. Finally, we use our InfoLaw to predict optimal data recipe under different training budgets and validate the optimal recipe by comparing with preset recipes.

Comparing with traditional scaling laws

Figure 1 contrasts our InfoLaw with traditional power scaling law in the loss–C plane. Both curves are fit using models in the 252M–1.2B range and then extrapolated to larger models. The Info curve tracks the MLQ data more closely within the fitting regime and remains accurate when extrapolating up to 7B models, avoiding the overly optimistic loss reductions predicted by the traditional law at high compute. Concretely, the traditional scaling law tends to under-estimate loss as C_m grows, whereas the Info curve better matches the realized validation losses of larger models.

Extrapolation to other LayerMix Sampling Weights

We first test the ability to generalize to an unseen LayerMix sampling weights. We test on unseen dataset generated with MLQ, MHQ on model scales ranging from $252M$ to $1.2B$, which are within the range of training data. Also we random sample 25 more sampling weights and run experiments on $1.2B$ model only.

The result is shown in Figure 3 . As can be seen, these points align remarkably well with the scaling law curve established by the initial HQ, MQ, LQ data, demonstrating the predictive power of our model on unseen LayerMix sampling weights. The traditional scaling laws requires additional experiments on different data recipes to fit new curves, while ours can directly predict loss on unseen recipes.

Extrapolation to Larger Models

To test the extrapolation ability on model scale, we use the same Layermix sampling weights MQ,LQ to train models ranging from 1.5B to 2.5B and HQ, LQ to train model with 2.5B parameters, which are out of the range of training data. The experimental results of larger models are shown in Figure 3(a-e), we can see InfoLaw predict the loss on larger scale accurately for all three sampling weights, proving the ability of scaling on model size.

Combination of Extrapolation

Furthermore, we combine the two extrapolation above and test the effectiveness on both unseen LayerMix sampling weights and unseen scales. We run experiments with MLQ,MHQ on models ranging from $1.5B$ to $7B$. As shown in Figure 3f, InfoLaw also generalise well on these combined extrapolation condition. On all the unseen data points, including unseen LayerMix sampling weights (MLQ, MHQ and other 25 sets random sampled weights) and unseen model scales, InfoLaw predict the validation loss with 0.15% average absolute error and maximum error is 0.96%. This proves that our proposed information scaling law has reliable extrapolation capability.

Extrapolation to Larger Overtrain Degree

To explore the model's reliability under varying suboptimality, we conducted a second series of experiments at a higher overtrain degree, $m' = 25$. This new regime was anchored by a 1.2B model trained on 640B tokens (the $C_{m'}$ experiment), contrasting with our initial C_m experiment anchored at 106B tokens.

For the $C_{m'}$-experiment, we calculated the Information using the same quality density $f(\theta^*)$ and $\lambda(N)(a^*, b^*)$ fitted previously on the C_m data. As shown in Figure 4, the new experimental points align with a new scaling law curve. The resulting curves for C_m and $C_{m'}$ appear nearly parallel, suggesting the overtrain degree m primarily shifts the curve's

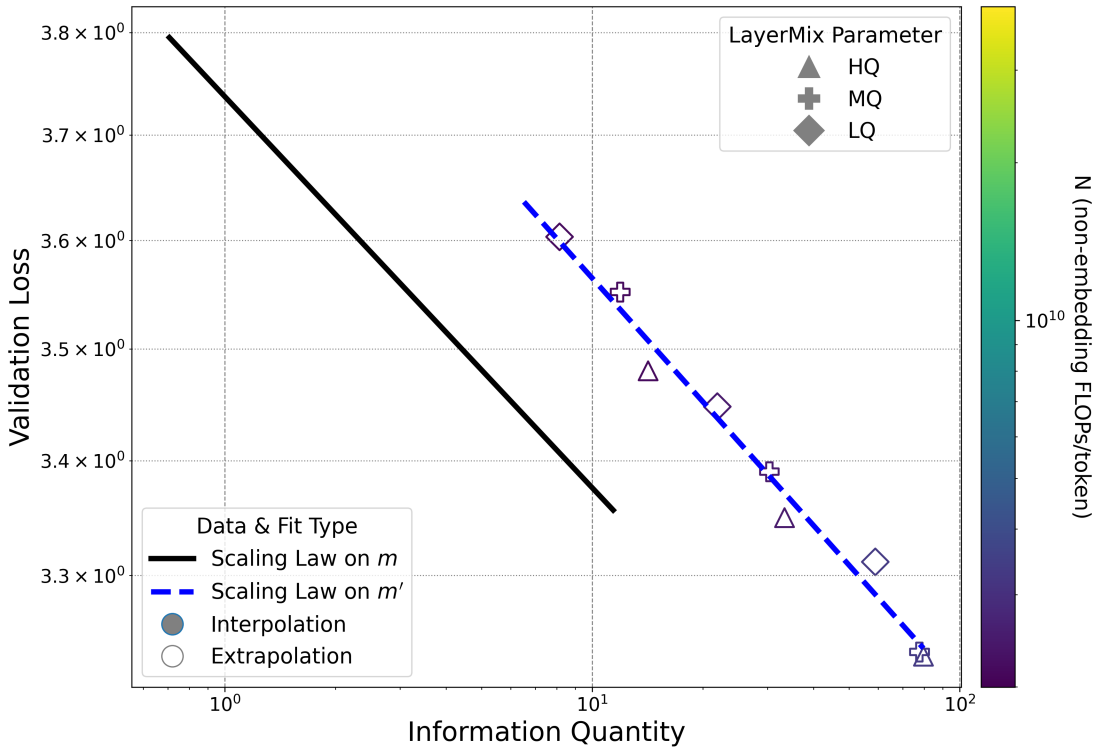


Figure 4. Cross-Regime Prediction of the Scaling Law. The blue line ($C_{m'}$) is a pure prediction, generated using parameters fitted only on the C_m data (black line). The fit for the $C_{m'}$ points demonstrates our InfoLaw's power to extrapolate across different overtrain degrees.

intercept. This confirms that our proposed Information Scaling Law is effective across different overtrain degrees.

Optimizing Data Recipe with InfoLaw

The ability of predicting loss on unseen data recipes and scales enables us to search for best data recipe without additional experiments. Similar to Liu et al. (2024). We randomly sample 100k LayerMix parameters from the parameter space, compute the information for each set of parameters, and convert it to loss via Equation 6. We then select the parameter that minimizes the predicted validation loss as the optimal LayerMix configuration for each training setting. To verify the optimal recipe, we conduct experiments on $2.5B$ model with optimal data recipe and 3 other Layermix sampling weights. The result optimal recipe is as in Table 1. As shown in Figure 3h, our optimal recipe achieves the best validation loss.

We additionally test generalization to unseen LayerMix parameters: on 25 held-out configurations for the 1.2B model, predicted and measured validation losses achieve a Pearson correlation of 0.76, suggesting InfoLaw can reliably rank recipes for efficient search.

In Table 2, we present the optimal LayerMix parameters for different model sizes and training-token counts under a fixed source-token budget of 500B tokens. The optimal LayerMix parameters exhibit two clear trends. First, at a fixed training-token count, smaller models favor a higher fraction of high-quality data, whereas larger models benefit more from diversity and thus allocate a smaller fraction to the high-quality data. Second, as the total training tokens increase, the optimal LayerMix parameters shift from a high-quality emphasis toward greater diversity. More results are shown in Appendix J. In short: Small models or small training budgets prioritize quality; large models or large training budgets prioritize diversity.

7. Conclusion

In this paper, we propose a refined scaling law modeling **InfoLaw**, which focus on predicting model performance on downstream tasks under data-constrained settings with weighted-quality mixing. The InfoLaw provides accurate predictions of model performance on unseen data recipes at larger computational scales, achieving an average absolute error of only 0.15% and a maximum error of 0.96%. This enables efficient discovery of optimal data recipes without the need for extensive additional experiments. Furthermore, the InfoLaw extrapolates reliably across varying degrees of over-training, offering an effective tool for selecting data recipes under different computational budgets.

8. Impact Statement

This paper aims to advance machine learning by improving our understanding of large language model performance under different data mixing and repetition strategies. Our InfoLaw can support more efficient pretraining by reducing expensive trial-and-error over data recipes. We do not anticipate direct negative societal consequences arising uniquely from this contribution. Broader ethical issues associated with LLMs, such as bias, misuse, and unsafe deployment, remain important but are not specifically introduced or materially amplified by our method beyond general improvements in training efficiency.

References

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Brown, T., Mann, B., Ryder, N., Subbiah, M., and et al. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., and et al. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chang, E., Paltenghi, M., Li, Y., Lin, P.-J., Zhao, C., Huber, P., Liu, Z., Rabatin, R., Shi, Y., and Chandra, V. Scaling parameter-constrained language models with quality data, 2024. URL https://arxiv.org/abs/2410.03083.

Chen, Z., Wang, S., Xiao, T., Wang, Y., Chen, S., Cai, X., He, J., and Wang, J. Sub-scaling laws: On the role of data density and training strategies in llms, 2025. URL https://arxiv.org/abs/2507.10613.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., and et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Common Crawl Foundation. Common Crawl. http://commoncrawl.org.

DeepSeek-AI, :, Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., Gao, H., Gao, K., Gao, W., and et al. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL https://arxiv.org/abs/2401.02954.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., and et al. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961, 2021. URL https://arxiv.org/abs/2101.03961.

Gadre, S. Y., Smyrnis, G., Shankar, V., Gururangan, S., Wortsman, M., Shao, R., Mercat, J., Fang, A., Li, J., Keh, S., et al. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., and et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gu, J., Yang, Z., Ding, C., Zhao, R., and Tan, F. Cmr scaling law: Predicting critical mixture ratios for continual pretraining of language models, 2024. URL https://arxiv.org/abs/2407.17467.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net, 2021. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2021.html#HendrycksBBZMSS21.

Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Henighan, T., Hume, T., Johnston, S., Mann, B., Olah, C., Olsson, C., Amodei, D., Joseph, N., Kaplan, J., and McCandlish, S. Scaling laws and interpretability of learning from repeated data, 2022. URL https://arxiv.org/abs/2205.10487.

Hestness, J., Narang, S., Ardalani, N., Diamos, G. F., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *CoRR*, abs/1712.00409, 2017. URL http://arxiv.org/abs/1712.00409.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J. W., and Sifre, L. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Isik, B., Ponomareva, N., Hazimeh, H., Paparas, D., Vassilvitskii, S., and Koyejo, S. Scaling laws for downstream task performance in machine translation, 2025. URL https://arxiv.org/abs/2402.04177.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Kang, F., Sun, Y., Wen, B., Chen, S., Song, D., Mahmood, R., and Jia, R. Autoscale: Scale-aware data mixing for pre-training llms, 2025. URL https://arxiv.org/abs/2407.20177.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., and et al. Datacomp-lm: In search of the next generation of training sets for language models, 2025. URL https://arxiv.org/abs/2406.11794.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., and et al. Few-shot learning with multilingual generative language models. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616/.

Liu, F., Zhou, W., Liu, B., Yu, Z., Zhang, Y., Lin, H., Yu, Y., Zhang, B., Zhou, X., Wang, T., et al. Quadmix: Quality-diversity balanced data selection for efficient llm pretraining. *arXiv preprint arXiv:2504.16511*, 2025.

Liu, Q., Zheng, X., Muennighoff, N., Zeng, G., Dou, L., Pang, T., Jiang, J., and Lin, M. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024.

Muennighoff, N., Rush, A. M., Barak, B., Le Scao, T., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., and Raffel, C. Scaling data-constrained language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., and et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. URL https://arxiv.org/abs/2306.01116.

Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., Werra, L. V., and Wolf, T. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Que, H., Liu, J., Zhang, G., Zhang, C., Qu, X., Ma, Y., Duan, F., Bai, Z., Wang, J., Zhang, Y., Tan, X., Fu, J., Su, W., Wang, J., Qu, L., and Zheng, B. D-cpt law: Domain-specific continual pre-training scaling law for large language models, 2024. URL https://arxiv.org/abs/2406.01375.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, H. F., Aslanides, J., Henderson, S., Ring, R., Young,

S., Rutherford, E., and et al. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. URL https://arxiv.org/abs/2112.11446.

Sardana, N., Portes, J., Doubov, S., and Frankle, J. Beyond chinchilla-optimal: accounting for inference in language model scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage?, 2023. URL https://arxiv.org/abs/2304.15004.

Shazeer, N. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M. Position: will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Xue, F., Fu, Y., Zhou, W., Zheng, Z., and You, Y. To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36:59304–59322, 2023.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., and et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Ye, J., Liu, P., Sun, T., Zhan, J., Zhou, Y., and Qiu, X. Data mixing laws: Optimizing data mixtures by predicting language modeling performance, 2025. URL https://arxiv.org/abs/2403.16952.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

A. Training Dataset

We use the English portion of the Common Crawl Dataset (Common Crawl Foundation), utilizing 96 of the snapshots, from CC-MAIN-2013-20 to CC-MAIN-2024-18. Following Bi et al. (2024), we ran a global fuzzy deduplication across all snapshots, resulting in a total dataset with 3.7T tokens.

B. Justification for the Normalization Term $\log(K)$

In Equation 3 , we incorporate a normalization term $\log(K)$ into the decay function to model the interaction between repetition decay and the total token budget. We selected this logarithmic form after rigorously evaluating alternative formulations. Specifically, we compared our chosen decay term against constant normalization and power-law normalization:

- **Constant Normalization:** Assuming the decay rate is independent of the dataset scale:

$$\text{Decay}(t) \propto e^{-\lambda(N)t} \tag{10}$$

- **Power-Law Normalization:** Assuming the decay scales polynomially with the token budget:

$$\text{Decay}(t) \propto e^{-\frac{\lambda(N)t}{K^{\alpha}}} \tag{11}$$

- **Logarithmic Normalization (Ours):**

$$\text{Decay}(t) \propto e^{-\frac{\lambda(N)t}{\log(K)}} \tag{12}$$

While we omit the visual plots for brevity, our preliminary experiments demonstrated that the alternative forms failed to unify the scaling behaviors across different token budgets K:

1. **Failure of Constant Normalization:** This formulation fails to account for the scaling properties of information density. Empirically, we observed that it systematically overestimates the accumulated Information for large models when trained with larger token budgets. Consequently, this leads to overly optimistic loss predictions that deviate significantly from the actual experimental results.
2. **Failure of Power-Law Normalization:** We found this formulation to be fundamentally unsuitable. It resulted in a complete failure to fit the relationship between Information and Validation Loss. The data points derived using power-law normalization remained scattered without exhibiting the necessary power-law correlation, rendering it impossible to derive a valid scaling law.

In contrast, the $\log(K)$ term was the only formulation that minimized the alignment error—successfully collapsing diverse configurations of (w, K, S) onto a single unified power-law curve (as shown in Figure 3f)—and maintained a low extrapolation error across the full range of model scales (252M to 7B). This suggests that the marginal utility of repeated data diminishes logarithmically relative to the total training budget.

C. LayerMix Sampling Function

We show the detail of LayerMix sampling function in Algorithm 1.

$$\sqrt{m} = \frac{N_{opt}}{N} = \frac{D}{D_{opt}} \tag{13}$$

A value of $m = 1$ indicates a compute-optimal training run, while $m > 1$ signifies that the model is overtrained relative to its compute budget.

D. Training

The model structures used in LayerMix are illustrated in Table 3. We train all the model with 2048 as the max sequence length, we use a cosine decay scheduler and the initial learning rate calculated by $lr = round(0.3118 \cdot C^{-0.1250}, 8)$, the warm up ratio is set 0.5%. We use AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay$= 0.1$.

Algorithm 1 LayerMix Sampling Function $H(w, K, S, B)$

Function H (w, K, S, B) **:**

Input : w: list of target proportions for six buckets, $w = [w_0, \ldots, w_5]$, $\sum w_d = 1$ K: total number of tokens for the final training dataset S: total number of tokens in the entire source corpora B: source distribution proportions $B = [0.05, 0.15, 0.2, 0.2, 0.2, 0.2]$

Output : D_{train}: final packed training dataset M: list of unique token counts per layer, $M = [M_0, \ldots, M_5]$ R: list of average repetition counts per layer, $R = [R_0, \ldots, R_5]$

Initialize empty training dataset $D_{train} \leftarrow \emptyset$ Initialize empty statistics lists $M \leftarrow []$, $R \leftarrow []$

for $d \leftarrow 0$ **to** 5 **do**

`// Iterate through each quality bucket`

$K_{needed} \leftarrow K \times w_d$ `// tokens needed from bucket` d `for the target mix`

$S_d \leftarrow S \times B[d]$ `// source tokens available in bucket` d

$Ratio_d \leftarrow K_{needed}/S_d$ `// sampling ratio for current bucket`

`// Detailed sampling process for bucket` d

Initialize empty temporary set $D_{sampled\_d} \leftarrow \emptyset$ **foreach** *data point* x *in bucket* d **do**

`// 1. Deterministic copy for the integer part of the ratio`

for $i \leftarrow 1$ **to** $\lfloor Ratio_d \rfloor$ **do**

Add x to $D_{sampled\_d}$

`// 2. Probabilistic sampling for the fractional part`

if $Ratio_d - \lfloor Ratio_d \rfloor > 0$ ***and*** $random() < (Ratio_d - \lfloor Ratio_d \rfloor)$ **then**

Add x to $D_{sampled\_d}$

Append all data from $D_{sampled\_d}$ to D_{train}

$M_d \leftarrow \min(K_{needed}, S_d)$ `// unique tokens for bucket` d

Append M_d to M $R_d \leftarrow K_{needed}/M_d$ `// average repetition count`

Append R_d to R

return D_{train}, M, R `// dataset and statistics`

E. Supplementary Analysis of Repetition Effects

Notation. IST (Infinite Source Tokens) denotes $S \gg K$, where repetition is negligible; LST (Limited Source Tokens) denotes $S = K$, where repetition is induced by the sampling weights. HQ/MQ refer to the LayerMix preset recipes in Table 1. Figure 5(a) provides an additional sanity check for the loss–C_m behavior under different repetition regimes, while Figure 5(b) shows the corresponding training-time dynamics motivating a saturation/decay model.

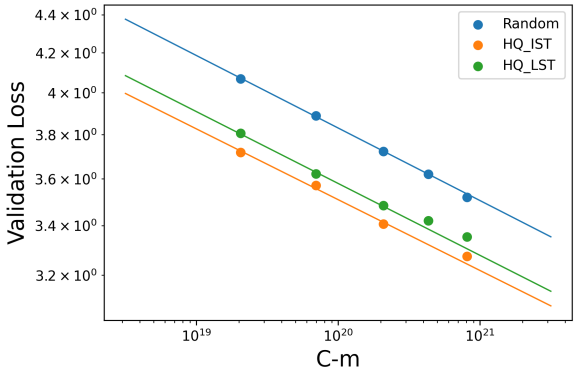
F. The relationship between benchmark validation loss and performance

Our InfoLaw focus on predicting the evaluation loss on downstream benchmarks. However, it also represents for the actual downstream performance. Figure 6 shows a near-linear relationship between validation loss and downstream performance on our evaluation tasks, and Table 4 shows the spearman corelation between validation loss and downstream performance. Lower loss consistently corresponds to higher performance within the operating regime of our models. This indicates that improvements in loss provide reliable signals for expected gains in downstream performance.

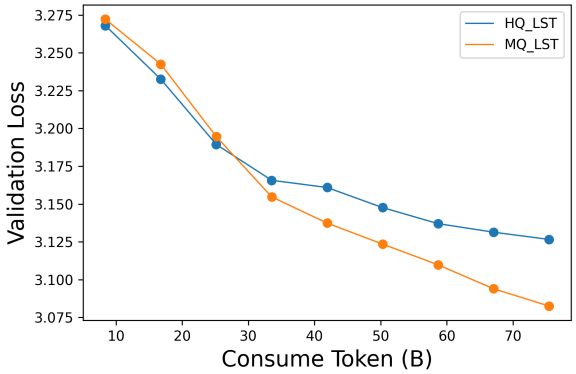
G. Alternative Fits for λ

In Section 5.2, we model the relationship between non-embedding FLOPs/token N and hyperparameter λ. Our primary specification adopts the logarithmic form Equation 9. Beyond this baseline, we also evaluated alternative function families, including an exponential form:

$$\lambda(x; a, b, c) = a \cdot \left(1 - e^{-bx+c}\right) \tag{14}$$



(a) Loss–C_m curves under different data regimes. Random: large source ($S \gg K$) with negligible repetition. HQ_IST: LayerMix with the HQ recipe and $S \gg K$ (negligible repetition). HQ_LST: the same HQ recipe but $S = K$, inducing repetition.



(b) Training-time evaluation loss for two 850M runs (HQ_LST vs MQ_LST), illustrating late-stage slowdown under heavier repetition.

Figure 5. **Supplementary evidence for repetition effects.** (a) In the loss–C_m view, repetition induces systematic deviation from a single power-law trend. (b) Heavier repetition leads to slower late-stage improvement and worse final loss, consistent with diminishing returns.

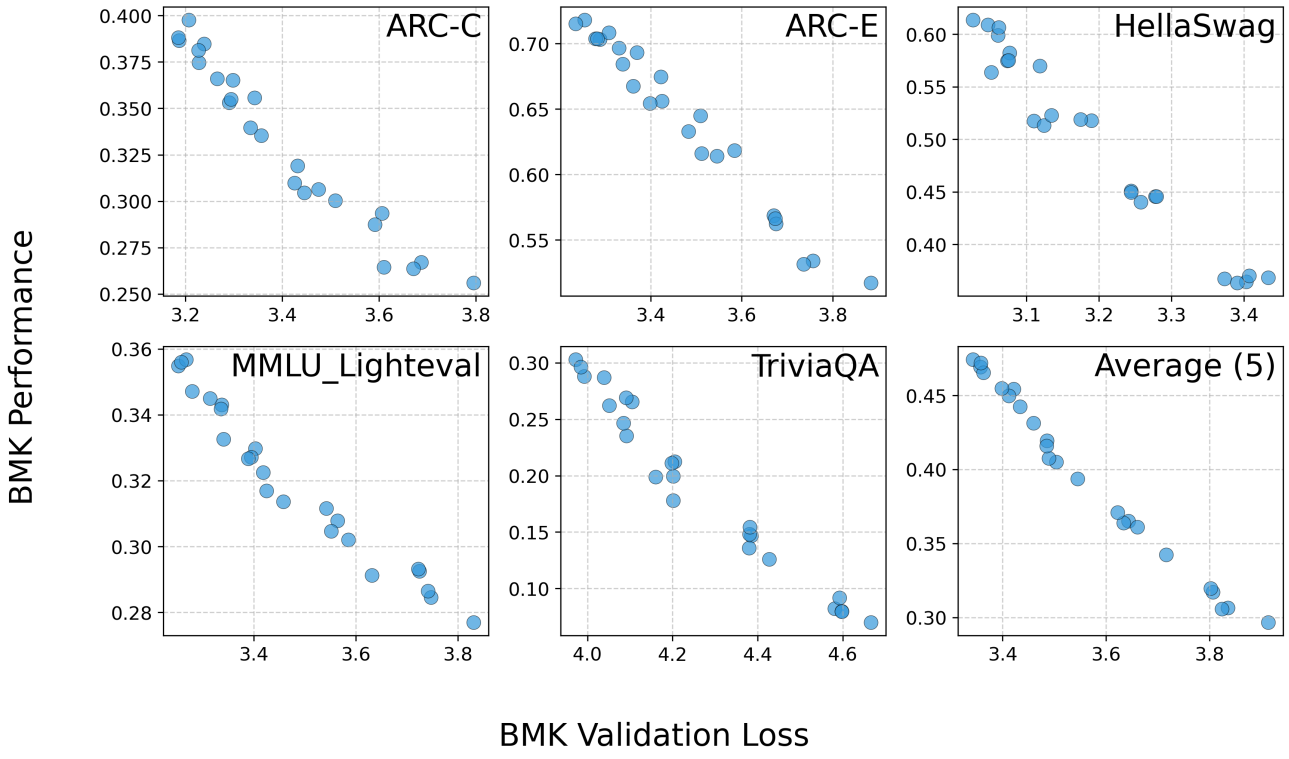


Figure 6. Validation loss versus downstream performance across benchmarks (ARC-C, ARC-E, HellaSwag, MMLU-Lighteval, TriviaQA) and their average.

Algorithm 2 Calculation of Overtrain Degree and Optimal Tokens

Function `CalculateOvertrainExtrapolation`$(model_{curr}, D_{curr}, models_{target})$**:**

Input : $model_{curr}$: size of the current model configuration D_{curr}: number of tokens used to train the current model $model_{target}$: size of the target model configuration

Output : m: calculated overtrain degree for the current configuration D_{target}: train tokens of target model under the same overtrain degree

`// Part 1: Calculate overtrain degree` m `from the current configuration`

$N_{curr} \leftarrow$ `Get_N`$(model_{curr})$ `// non-embedding FLOPs/token for current model`

$C \leftarrow N_{curr} \times D_{curr}$ `// total compute budget`

$N_{opt} \leftarrow 0.06085 \times C^{0.5445}$ `// Chinchilla-optimal` N `for budget` C

$D_{opt} \leftarrow 16.4326 \times C^{0.4555}$ `// Chinchilla-optimal tokens for budget` C

$\sqrt{m} \leftarrow N_{opt}/N_{curr}$ `// overtrain degree` m `(equiv.` $\sqrt{m} = D_{curr}/D_{opt}$`)`

`// Part 2: Extrapolate to target model while keeping` m `constant`

foreach $model_t$ **to** $[model_{curr}] + models_{target}$ **do**

  $N_t \leftarrow$ `Get_N`$(model_t)$ `// non-embedding FLOPs/token for target model`

  $N'_{opt} \leftarrow N_t \times \sqrt{m}$ `// corresponding optimal` N `for the target`

  $C_{new} \leftarrow (N'_{opt}/0.06085)^{1/0.5445}$ `// derive new compute budget`

  $D'_{opt} \leftarrow 16.4326 \times C_{new}^{0.4555}$ `// optimal tokens for the new budget`

  $D_{target} \leftarrow D'_{opt} \times \sqrt{m}$ `// required tokens for the target model`

return m, D_{target} `// overtrain degree and target train tokens at same` m

and a power-law form:

$$\lambda(x; a, b) = a \cdot x^b \tag{15}$$

As shown in Figure 7, the logarithmic model achieves the best fit to the $N - \lambda$ relationship, outperforming the exponential and power-law alternatives. Accordingly, we adopt function 9 as the final parameterization.

H. Deviation of Traditional Scaling Law

We show all $Loss$-C curve of different LayerMix sampling weights with IST and LST in Figure 8 and Figure 9, they all exhibit a clear deviation from the traditional scaling law, which is fitted from the first three data points.

I. Quality Score

We show some data samples in different Quality buckets in Figure 10. This figure indicates that high-score samples under our merged FineWebEdu and DCLM scores are more coherent and instructional. By contrast, low-score cases predominantly consist of advertisements or low-information content, offering little substantive value.

Table 5 reports four benchmark results for training a 1.2B model from scratch on 30B tokens using three datasets: the top 5% and top 20% selected by the FineWebEdu classifier, and a random sample, all from Penedo et al. (2023). High-quality data selected by FineWebEdu outperforms the random baseline, and higher-quality subsets yield better results.

J. Optimizing Token Mix with InfoLaw

We present the detailed optimal LayerMix parameters (or token-mix ratios) for different models and training budgets predicted by InfoLaw in Table 6. This table shows that small models or small training budgets prioritize quality, while large models or large training budgets prioritize diversity.

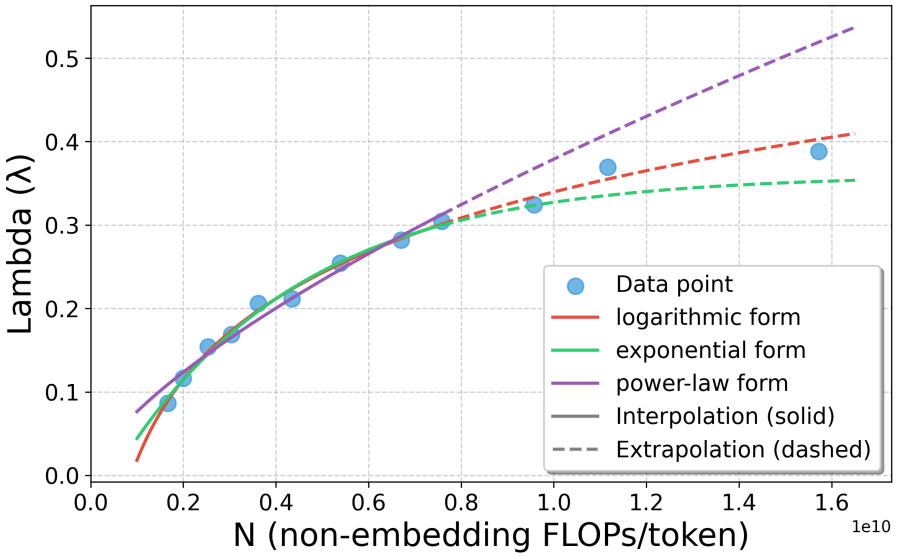


Figure 7. Comparison of functional fits for λ as a function of N (non-embedding FLOPs/token). The logarithmic form provides the best in-domain fit and extrapolation behavior compared with the exponential and power-law alternatives. Solid lines denote interpolation over observed N; dashed lines indicate extrapolation beyond the observed range.

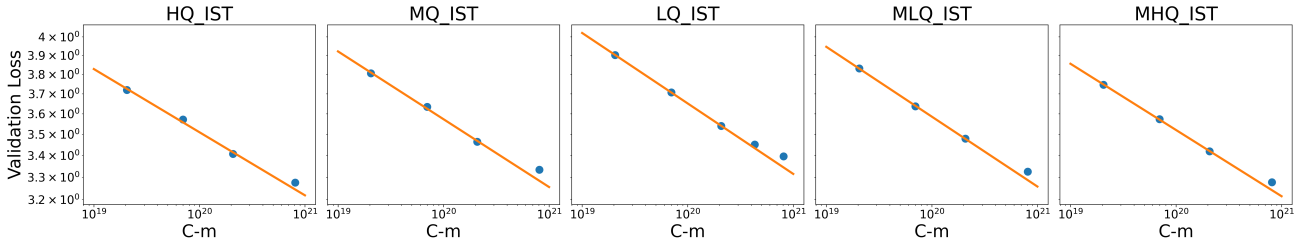


Figure 8. $Loss$ and C_m Curve of different LayerMix IST experiments

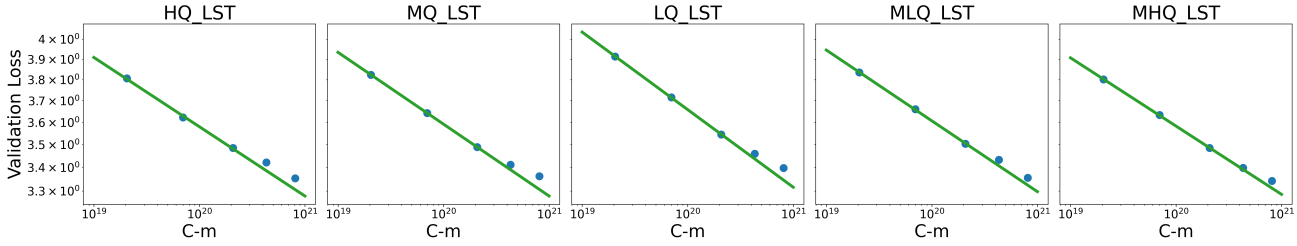


Figure 9. $Loss$ and C_m Curve of different LayerMix LST experiments

Quality_Range 0-5%	Quality_Range 80-100%
Celebrate your way Whether you are having a picnic with your family, a barbeque with friends in the backyard or follow the Australian of the Year awards, Australia Day on the 26th of January is an occasion to come together as a nation and celebrate what's great about Australia. On Australia Day we celebrate the past, present and future of the country. It is a commemoration of the day that the First Fleet landed in Sydney Cove in 1788, as well as a celebration of all the achievements of our country. The tradition of having Australia Day as a national holiday on 26 January is actually a pretty recent one. Not until 1935 did all Australian states and territories use that name to mark that date, and only in 1994 did they begin to celebrate Australia Day as a public holiday on that date. Today, Australia Day has grown to be a community day which is embraced by most Australians. Apart from the formal ceremonies around the country such as flag raising, citizenship ceremonies and the presentation of community awards, there are a wider range of festivities that encourage the participation of all family and community members.	This message was posted by The Dumper, posted on January 05, 2002 at 03:06:18 coming from 209.204.139 This message is a reply to Will BUY snes copier posted from Spongebob posted at January 05, 2002 at 02:25:29\n> Looking for an snes copier. Preferrably the Super Wildcard DX 2. Will pay $$$$$ or it.
March 7, 2012 (Shirley Allen)\nMonthly home prices in the United States increased by a seasonally adjusted 0.7 percent in December which follows a similarly revised 0.7 percent gain in November according to the Federal Housing Finance Agency's (FHFA) monthly House Price Index (HPI).\nDecember's home prices were still 0.8 percent lower than they were a year ago and since the market peak in April 2007, home prices have declined over 18 percent and are at roughly the same levels last seen in March of 2004.\n\nSix of the nine Census Divisions posted monthly price gains in December with the Mountain Division recording the most improvement of 2.5 percent. Two Divisions posted declines in home prices while one Division, the Middle Atlantic, remained unchanged from the previous month. Of the two Divisions that posted declines, the West North Central Division posted the largest decline of 0.9 percent.\nSeven of the Divisions registered year-over-year price declines with the Pacific Division posting the largest decline of 3.8 percent. The only two Divisions that posted an increase in annual home prices were the East South Central Division and the West South Central Division which posted increases of 3.0 and 1.7 percent, respectively.	these adorable little wall hangings feature bright pops of colour, tassley texture and pretty little berry knots. How can you resist?\nPlus...you get to choose from our gorge range of colours!\nMade from 100% recycled cotton and mounted on a Tasmanian Oak dowel.\nMeasures approximately 16cm wide by 33cm long (including hanger).\n*This item is hand made with love and ready to ship. Colours vary from screen to screen.\n*Looking for something similar? Custom orders are available – price available upon request.\n*Outside Australia? Please contact us for country specific freight charges.'

Figure 10. Case study contrasting data quality. Left (0–5% quality range): coherent, informational, and instructional passages. Right (80–100% quality range): low-information, ad-like content with minimal reasoning or educational value.

Table 3. Structure of models used in LayerMix.

Model	Hidden dim. (C)	MLP dim. (D)	Layers (L)	Heads
252M	1024	2752	20	16
302M	1024	2752	24	16
392M	1280	3392	20	20
470M	1280	3392	24	20
566M	1536	4096	20	24
680M	1536	4096	24	24
850M	1792	4800	22	28
1B	1920	5120	24	30
1.2B	2048	5440	24	16
1.5B	2304	6144	24	36
1.8B	2304	6144	28	36
2.5B	2560	6848	32	40
7.7B	4096	14336	32	32

Table 4. Spearman correlation between validation loss and performance across benchmarks

Benchmark	Spearman r_s	p-value
ARC-C	-0.979	1.02×10^{-16}
ARC-E	-0.982	2.72×10^{-17}
HellaSwag	-0.942	6.13×10^{-12}
MMLU-LightEval	-0.989	1.26×10^{-19}
TriviaQA	-0.970	4.53×10^{-15}
Average (5)	-0.996	3.54×10^{-24}

K. Generalization to Refinedweb

To evaluate the robustness and generalization capability of the InfoLaw across different data distributions, we conducted an additional series of verification experiments on the RefinedWeb dataset (Penedo et al., 2023).

Experimental Setup. We followed the identical data preprocessing, LayerMix sampling, and training procedures described in Section 3.1 and Section 5.1, with the sole exception of replacing the source corpus with RefinedWeb. Due to time and computational constraints, we limited the scope of this study to three LayerMix sampling configurations: HQ (High Quality) and LQ (Low Quality) were used for parameter fitting (interpolation), while MLQ (Medium-Low Quality) was held out for extrapolation testing. For each configuration, we trained models at three specific scales: 302M, 566M, and 1.2B parameters.

Fitting and Extrapolation. We applied the fitting methodology outlined in Section 5.2. Our analysis yielded two key observations:

- **Consistency of Quality Density (f):** The fitted values for the quality density function f_d were numerically very close to those derived from our primary dataset. Specifically, the fitted parameter θ is 0.93 for RefinedWeb, which is remarkably close to the value of 0.92 obtained from our primary dataset. We attribute this similarity to the fact that RefinedWeb (Penedo et al., 2023) is also derived from Common Crawl (Common Crawl Foundation); despite employing different filtering strategies, the shared underlying data source results in a comparable information density distribution.
- **Optimization of $\lambda(N)$:** In the main experiments, we modeled the relationship between the parameter $\lambda(N)$ and model

Table 5. FineWebEdu-selected subsets vs. random data for training a 1.2B model on 30B tokens

Model	Data	ARC-C	HellaSwag	TriviaQA	MMLU-LightEval	avg
1.2B	Random 30B	28.50%	51.56%	15.55%	30.23%	31.46%
1.2B	FWE-top20% 30B	34.30%	55.26%	20.05%	32.82%	35.61%
1.2B	FWE-top5% 30B	37.20%	55.14%	19.25%	34.50%	36.52%

Table 6. The detailed best layer token mix for different models and train token

Model	**Train Token**	**Source Token**	$w0$	$w1$	$w2$	$w3$	$w4$	$w5$
7B	200B	500B	0.619	0.376	0.004	0.001	0.000	0.000
	300B	500B	0.548	0.444	0.004	0.003	0.002	0.000
	400B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	500B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	600B	500B	0.491	0.487	0.017	0.005	0.000	0.000
	700B	500B	0.439	0.430	0.130	0.001	0.000	0.000
	800B	500B	0.439	0.430	0.130	0.001	0.000	0.000
	900B	500B	0.404	0.403	0.183	0.006	0.003	0.000
	1000B	500B	0.395	0.387	0.214	0.003	0.001	0.000
1.8B	200B	500B	0.825	0.165	0.005	0.004	0.001	0.000
	300B	500B	0.619	0.376	0.004	0.001	0.000	0.000
	400B	500B	0.548	0.444	0.004	0.003	0.002	0.000
	500B	500B	0.548	0.444	0.004	0.003	0.002	0.000
	600B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	700B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	800B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	900B	500B	0.491	0.487	0.017	0.005	0.000	0.000
	1000B	500B	0.491	0.487	0.017	0.005	0.000	0.000
1.2B	200B	500B	0.926	0.066	0.006	0.002	0.000	0.000
	300B	500B	0.758	0.229	0.012	0.001	0.000	0.000
	400B	500B	0.619	0.376	0.004	0.001	0.000	0.000
	500B	500B	0.619	0.376	0.004	0.001	0.000	0.000
	600B	500B	0.548	0.444	0.004	0.003	0.002	0.000
	700B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	800B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	900B	500B	0.496	0.492	0.007	0.003	0.002	0.000
	1000B	500B	0.496	0.492	0.007	0.003	0.002	0.000

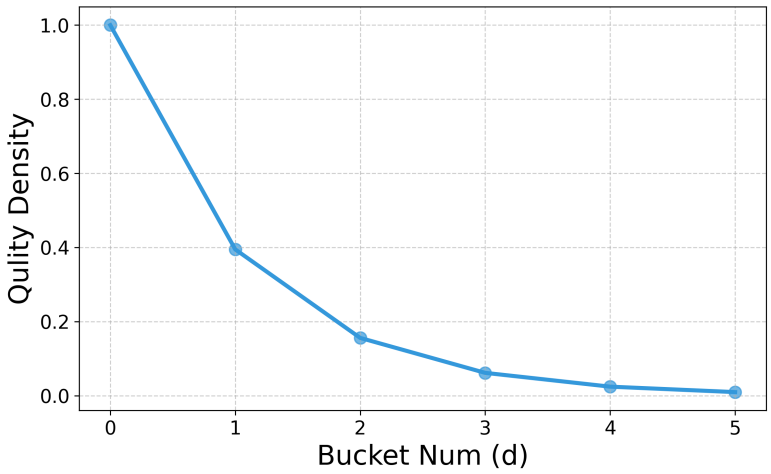


Figure 11. The fitted quality density function f_d on the RefinedWeb dataset.

scale N using a logarithmic curve. However, due to the limited number of data points in this verification set (only three distinct model scales), fitting a robust $\lambda(N) - N$ curve was not feasible. Consequently, we skipped the curve fitting step for $\lambda(N)$ and directly searched for the optimal λ values corresponding to the specific model sizes (302M, 566M, and 1.2B).

Results. Using the parameters fitted on the HQ and LQ configurations, we predicted the validation loss for the unseen MLQ configuration, as illustrated in Figure 12. The InfoLaw demonstrated strong predictive accuracy on the RefinedWeb dataset, achieving a maximum absolute error of 0.36% and a mean absolute percentage Error 0.24% on the extrapolated MLQ experiments. These results further corroborate that the InfoLaw effectively captures the fundamental trade-offs between data quality, repetition, and compute scale, independent of the specific underlying data source.

L. Limitation

We note several limitations of our work. Our data bucketing is based on a fixed, empirical heuristic. We have not performed ablation studies to determine the optimal number or boundaries of these quality tiers. A more systematic approach to data partitioning could further improve the model's predictive accuracy. And while we observe that the overtrain degree m systematically shifts the scaling law curve, a theoretical explanation for this behavior is still needed. These areas present clear avenues for future work.

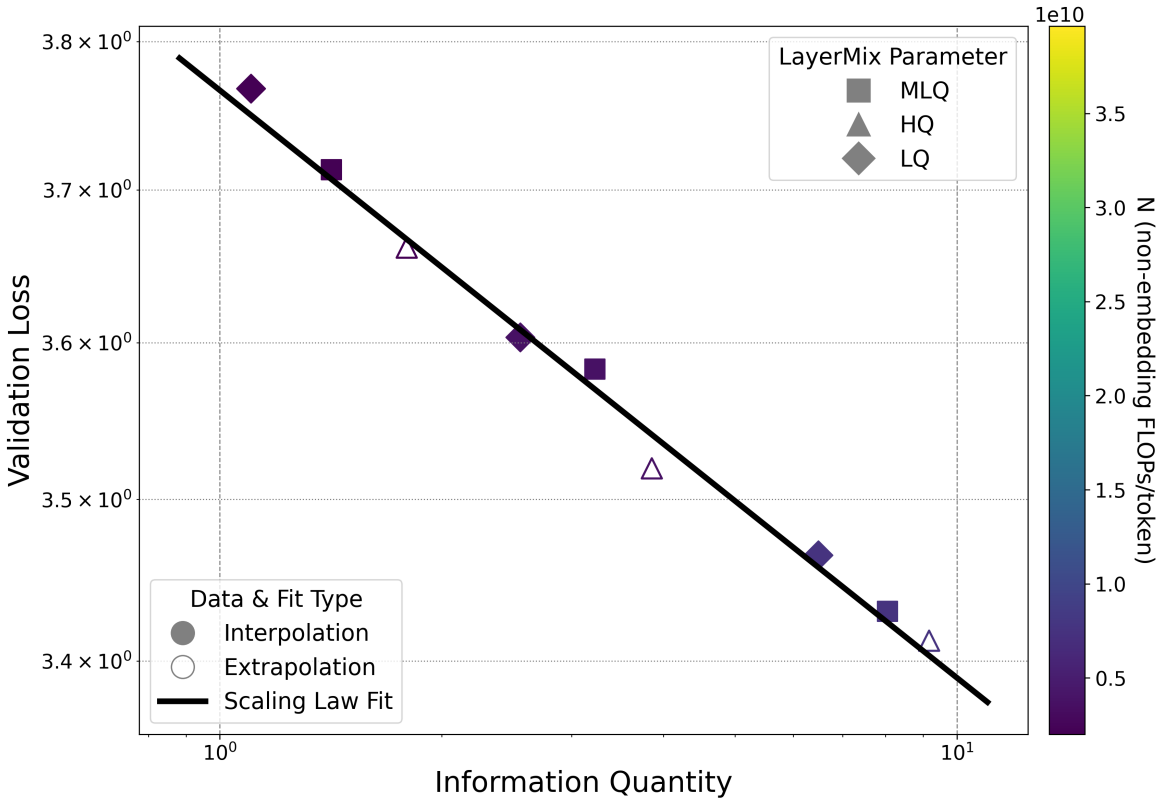


Figure 12. The Unified Information-Loss Scaling Law on the RefinedWeb dataset.