Meta-Scaling Laws: On the Invariance and Generality of Neural Scaling

Xing Han
Johns Hopkins University
Baltimore, MD 21218, USA
xhan56@jhu.edu

Liu Ziyin
NTT Research, MIT
Cambridge, MA 02139, USA
ziyinl@mit.edu

Suchi Saria
Johns Hopkins University, Bayesian Health
Baltimore, MD 21218, USA
ssaria1@jhu.edu

Paul Pu Liang
MIT
Cambridge, MA 02139, USA
ppliang@mit.edu

Abstract

Neural scaling laws establish a predictable relationship between model performance and data or compute, offering crucial guidance for resource allocation in new domains and tasks. Yet such laws are most needed precisely where they are hardest to obtain: fitting one for a new model–task pair demands expensive sweeps that typically exhaust the very compute budget the law is meant to economize. We study *meta-scaling laws*: relations that characterize when neural scaling behavior is preserved and predict how it changes under data transformations, enabling forecasts across domains when the shift can be approximated by a compatible transformation. We address this by identifying the right invariants: scaling laws are preserved under bijective (information-preserving) transformations of the data and modified in predictable, information-theoretically grounded ways under non-bijective transformations that lower its information resolution ρ: a single axis along which a law fit in one domain can be transported to another. We validate this across language, vision, and speech, and demonstrate two cross-domain applications: predicting scaling for language models trained on electronic health records from laws fit on general text, and predicting time-series classification scaling under varying levels of noise injection, recovering the data-scaling exponents to within 3% error.

1 Introduction

The development of foundation models has been governed by the empirical regularities known as Neural Scaling Laws (NSLs). These laws establish a predictable power-law relationship between a model's performance and the physical resources consumed during training, namely parameter count (N), dataset size (D), and compute budget. First formalized for language modeling [30] and subsequently extended to vision [53], multimodal learning [1], and beyond, these laws typically take the form $L \approx AN^{-\alpha} + BD^{-\beta}$, where L is the pretraining loss and A, B are coefficients, indicating that the pretraining loss decreases predictably as parameters or training data are added.

A key limitation of this framework is that fitting such a law is itself prohibitively expensive: for instance, deriving a scaling law for language models [30] requires training a sweep of models spanning orders of magnitude in N and D and evaluating across multiple downstream tasks. This is infeasible in most real-world settings, where data, compute, or task diversity are all limited: yet these are precisely the settings in which practitioners most need to anticipate scaling behavior be-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

fore committing to large-scale data collection or training runs [54, 20, 11, 42]. A medical team, for example, may want to predict how a language model will scale on electronic health records without incurring the cost of training large models from scratch, a need that becomes acute when the target domain has limited data availability or when computational constraints preclude extensive experiments [26, 37, 23, 52]. This motivates *meta-scaling laws*: relations that characterize invariance and predict changes in neural scaling under data transformations. Such relations enable predictions from a source-domain scaling law when the target shift can be approximated by a compatible transformation, reducing the need for repeated scaling sweeps. While recent work has begun to explore how data quality [48], volume [23], and diversity [8] affect scaling, these studies remain confined to a single domain and task, and model data degradation solely as a reduction in effective sample size. As a consequence, they predict the same irreducible loss floor regardless of data quality, at odds with information-theoretic limits when task-relevant information is irreversibly destroyed.

In this work, we develop an information-theoretic framework for meta-scaling laws. We use data transformations to relate source and target domains through the task-relevant information they preserve, providing a basis for predicting how scaling behavior changes. We distinguish between **bijective transformations**, which are information-preserving and (we prove) leave the scaling exponents invariant, and **non-bijective transformations**, which lower the data's information content and degrade scaling through two mechanisms: *variance inflation*, which raises the sample count needed for equivalent statistical precision, and *optimal loss shift*, which raises the irreducible loss floor itself. These two mechanisms yield a new scaling law parameterized by the information resolution $\rho(T) = I(X';Y)/I(X;Y)$, the fraction of task-relevant mutual information that survives a transformation T from clean input X to its transformed counterpart X', with $\rho = 1$ recovering the standard Chinchilla form and $\rho < 1$ capturing principled, information-theoretically grounded deviations from it. Our experiments demonstrate the accuracy of this law across language, vision, and speech domains under a range of bijective and non-bijective transformations, recovering the data-scaling exponents β to within 3% error when transferred across transformation types. We further show two cross-domain applications: predicting scaling for language models trained on electronic health records using a law fit on general text, and predicting how time-series foundation models scale under varying levels of real-world noise injection, in both cases obtaining accurate forecasts without training models at the target scale.

Related Work. Neural scaling laws were first formalized for language [30, 27] and extended to vision [53, 2], multimodal [25, 9, 1, 47], and time-series [18] settings, generally treating data as a monolithic quantity. A complementary line predicts scaling behavior without exhaustive training, via observational capability spaces [44], intermediate checkpoints [10], loss-to-loss transfer [6], or domain-weighted mixtures [47], but offers no principled account of when scaling exponents should change versus remain invariant. Data-dependent formulations move closer to this question, refining laws to account for noise [5], quality-conditioned filtering [23, 33], diversity and syntheticity [7], and density-driven sub-scaling [8], yet model degradation only as effective-sample-size reduction and predict a quality-independent loss floor. Finally, data transformations - bijective ones underlying normalizing flows and ICA [43, 40, 28], and non-bijective ones such as quantization and dimensionality reduction [12, 34, 3, 32] are pervasive in practice but have not been connected to scaling theory. Our framework unifies these threads through information resolution.

2 Meta-Scaling Laws via Information Resolution

In this section, we propose that NSLs are fundamentally governed by the **information resolution** $\rho(T) = I(X';Y)/I(X;Y)$ of the dataset: the fraction of task-relevant mutual information that a transformation $T : X \mapsto X'$ preserves about the target Y. This single quantity links the two transformation classes we study: a transformation is bijective when $\rho(T) = 1$ (all task-relevant information is preserved, regardless of how the input is re-encoded), and non-bijective when $\rho(T) < 1$ (some task-relevant information is irreversibly destroyed). In Section 2.1, we show that NSLs are invariant under bijective transformations, since $\rho = 1$ leaves the law's data-dependent term unchanged. In Section 2.2, we characterize how $\rho < 1$ modifies NSLs through two distinct mechanisms — variance inflation and Bayes-risk elevation. To support our framework, in Appendix C.1, we discuss how ρ can be estimated in practice and used to predict scaling behavior across domains.

2.1 NSLs are Robust to Bijective Transformations

Real-world model development routinely subjects data to transformations: tokenizers re-encode text, frequency-domain conversions re-represent audio, color-space changes re-parameterize images, and normalizing flows re-distribute features. Each one alters the surface form of the input while, in many cases, preserving every bit of information relevant to the downstream task. We formalize this class first, since it provides the natural baseline against which the more interesting non-bijective case is compared: any change in scaling behavior under a bijective transformation must come from optimization dynamics, not from a change in what the data fundamentally encodes.

Definition 1 (Bijective Transformation)**.** A transformation $T : \mathcal{X} \to \mathcal{X}'$ is *bijective* if it satisfies:

1. **Injectivity**: For all $x_1, x_2 \in \mathcal{X}$, $T(x_1) = T(x_2)$ implies $x_1 = x_2$.
2. **Surjectivity**: For every $x' \in \mathcal{X}'$, there exists $x \in \mathcal{X}$ such that $T(x) = x'$.

Equivalently, T is bijective if and only if there exists an inverse mapping $T^{-1} : \mathcal{X}' \to \mathcal{X}$ such that $T^{-1}(T(x)) = x$ for all $x \in \mathcal{X}$ and $T(T^{-1}(x')) = x'$ for all $x' \in \mathcal{X}'$.

We consider two classes of bijective transformations.

Linear Transformations. Given a dataset with representations in $\mathbb{R}^d$, a linear transformation is defined as: $T_{\text{linear}} : \mathbb{R}^d \to \mathbb{R}^d,\ x \mapsto Wx$ where $W \in \mathbb{R}^{d\times d}$ is a full-rank matrix with $\det(W) \neq 0$. The bijectivity follows from the existence of the inverse transformation $T^{-1}_{\text{linear}}(x') = W^{-1}x'$.

Nonlinear Bijective Transformations. A nonlinear transformation $T_{\text{nonlinear}} : \mathbb{R}^d \to \mathbb{R}^d$ is bijective if it is a diffeomorphism: a smooth bijection with a smooth inverse. Such transformations can be constructed through compositions of bijective layers: $T_{\text{nonlinear}} = f_L \circ f_{L-1} \circ \cdots \circ f_1$ where each f_ℓ is an bijective function. The composition of bijective functions is bijective, since $(f_L \circ \cdots \circ f_1)^{-1} = f_1^{-1} \circ \cdots \circ f_L^{-1}$. A transformation T is information-preserving with respect to a random variable X if the entropy is conserved: $H(T(X)) = H(X)$. Both discrete and continuous random variables under diffeomorphisms satisfy this property [38, 41]. We now provide theoretical grounding for why NSLs remain invariant under bijective transformations.

Lemma 1 (Mutual Information Invariance)**.** Let X be the input data and Y be the target labels. For any bijective transformation T, the mutual information is preserved: $I(T(X); Y) = I(X; Y)$.

Proof. By the data processing inequality, for any T, we have $I(T(X); Y) \leq I(X; Y)$. Since T is bijective, we can also apply the transformation T^{-1} to obtain $I(X; Y) = I(T^{-1}(T(X)); Y) \leq I(T(X); Y)$. Combining both inequalities yields $I(T(X); Y) = I(X; Y)$. □

By mutual information definition: $I(X; Y) = H(Y) - H(Y|X)$, since $H(Y)$ depends only on the target distribution and $I(T(X); Y) = I(X; Y)$ by Lemma 1, we also have $H(Y|T(X)) = H(Y|X)$.

Connection to Model Performance. The Chinchilla Scaling Law [27] specifies that

$$L(N, D) = \frac{A}{N^\alpha} + \frac{B}{D^\beta} + E, \tag{1}$$

where N is the number of training parameters, D is the number of training tokens, α/β is the exponent for model/data scaling, and E is the irreducible loss. Lemma 1 has direct implications for several learning-theoretic quantities below that govern NSL behavior.

Proposition 1 (Bayes Risk Invariance)**.** Let $T : \mathcal{X} \to \mathcal{X}'$ be a bijective transformation. The Bayes-optimal risk satisfies:

$$R^*_{T(X)} := \inf_{f:\mathcal{X}'\to\mathcal{Y}} \mathbb{E}[\ell(f(T(X)), Y)] = \inf_{g:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(g(X), Y)] =: R^*_X \tag{2}$$

That is, the irreducible loss E in the scaling law remains unchanged under bijective transformations.

Proposition 2 (Sample Complexity Preservation)**.** The sample complexity for achieving ϵ-excess risk is preserved: $n_\epsilon(T(X), Y) = n_\epsilon(X, Y)$, where n_ϵ denotes the minimum number of samples required to achieve excess risk at most ϵ with high probability. Consequently, the data scaling exponent β remains invariant.

Table 1: Comparison of our framework with two recent scaling law formulations under data degradation: Quality-Aware scaling laws [48] and Sub-Optimal [8] scaling laws.

Aspect	Quality-Aware	Sub-Optimal	Ours (Info-Resolution)
Formulation	$\frac{A}{N^\alpha} + \frac{B}{D^\beta Q^\gamma} + E$	$\frac{\lambda_N R_N}{N^{\alpha_N}} + \frac{\lambda_D R_D}{D^{\alpha_D}} + E$	$\frac{A}{N^\alpha} + \frac{B}{D^\beta} \cdot \rho(T)^{-\nu} + E + \kappa(1-\rho(T))^\mu$
Mechanisms	Single (effective data)	Single (decay factors)	Dual (inflation + shift)
Irreducible loss E	Constant	Constant	ρ-dependent: $E + \kappa(1-\rho)^\mu$
Degradation parameter	Data quality Q	Data density ρ_{density}	Transformation property $\rho(T)$
Interpretation	Fewer useful samples	Redundancy $\rightarrow$ diminishing returns	Variance inflation + optimal loss shift
Asymptotic ($D \to \infty$)	$L \to E$	$L \to E$	$L \to E + \kappa(1-\rho(T))^\mu$

Proposition 3 (Model Capacity Preservation)**.** The model capacity to approximate the optimal predictor is preserved, leaving the parameter scaling exponent α invariant.

Lemma 2 (NSL Invariance under Bijective Transformations)**.** Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a dataset drawn from distribution P_{XY}. Define the transformed dataset $\mathcal{D}' = \{(T(x_i), y_i)\}_{i=1}^n$. If the test loss on $\mathcal{D}$ follows the scaling law in Eq.(1), then the test loss on $\mathcal{D}'$ follows: $L'(N, D) = \frac{A'}{N^\alpha} + \frac{B'}{D^\beta} + E$ with identical exponents (α, β) and irreducible loss E.

The core intuition is that bijectivity allows lossless translation between predictor spaces: any predictor f on X maps to $g = f \circ T^{-1}$ on $T(X)$ with identical risk, preserving Bayes risk (E), sample complexity (β), and approximation rates (α). Lemma 2 combines these invariances directly. Proofs of Proposition 1 - 3 can be found in Appendix B.

2.2 NSLs are Dependent on Information Resolution

In practice, many of the transformations applied to training data are not bijective. Quantization collapses continuous values onto a finite grid [34]; low-rank projection discards directions in feature space [35]; aggressive deduplication, summarization, or filtering removes content entirely [21]; and stochastic corruption [48] injects ambiguity into the input-output relationship. Each of these operations can irreversibly destroy task-relevant information about Y, driving $\rho(T) < 1$ and altering the scaling law in ways that bijective re-encodings cannot. We now define this class formally and characterize how it modifies scaling behavior.

Definition 2 (Non-Bijective Transformation)**.** A transformation $T : \mathcal{X} \to \mathcal{X}'$ is *non-bijective* if and only if at least one of the following holds:

1. **Non-injectivity**: There exist distinct $x_1, x_2 \in \mathcal{X}$ such that $T(x_1) = T(x_2)$.
2. **Non-surjectivity**: There exists $x' \in \mathcal{X}'$ such that $x' \notin \text{range}(T)$.

We define the *Information Resolution* for non-bijective transformations as $\rho(T) := \frac{I(X';Y)}{I(X;Y)} \in (0, 1]$, to measure the fraction of task-relevant information preserved by T. The Data Processing Inequality guarantees $\rho(T) \leq 1$ with strict inequality for non-invertible T. Our key insight here is that such transformations degrade model performance through two distinct mechanisms: (1) **Variance Inflation:** information reduction increases the variance of parameter estimates, requiring more samples to achieve equivalent statistical precision. (2) **Optimal Loss Shift:** discarded information creates an irreducible performance gap: a hard limit that no amount of data can overcome. Specifically, for (1), information loss inflates the variance of learning. For unbiased estimators under corrupted data, variance scales inversely with usable information: $\text{Var}(\hat{\theta}_{\text{corrupted}}) \propto \frac{1}{D \cdot \rho(T)^\nu}$, where $\nu > 0$ depends on the corruption type, which controls how severely information loss penalizes statistical efficiency: larger ν means variance grows faster as $\rho(T)$ decreases. In our empirical evaluations, ν falls within the range $(0, 1)$ (e.g., $\nu \approx 0.19$ for next-token prediction tasks), reflecting the influence of the redundancy and noise present in real-world data. Recent work [48] has discovered similar mechanisms. However, it doesn't capture mechanism (2), where the optimal pretraining loss on corrupted data is strictly worse than on clean data.

Lemma 3 (Information Ceiling Bound)**.** Let L_X^* and $L_{X'}^*$ denote optimal pretraining losses for predicting Y from X and $X' = T(X)$ respectively. Then:

$$L_{X'}^* - L_X^* \geq \Phi\left(I(X;Y) - I(X';Y)\right) = \Phi\left(I(X;Y)(1-\rho(T))\right) \tag{3}$$

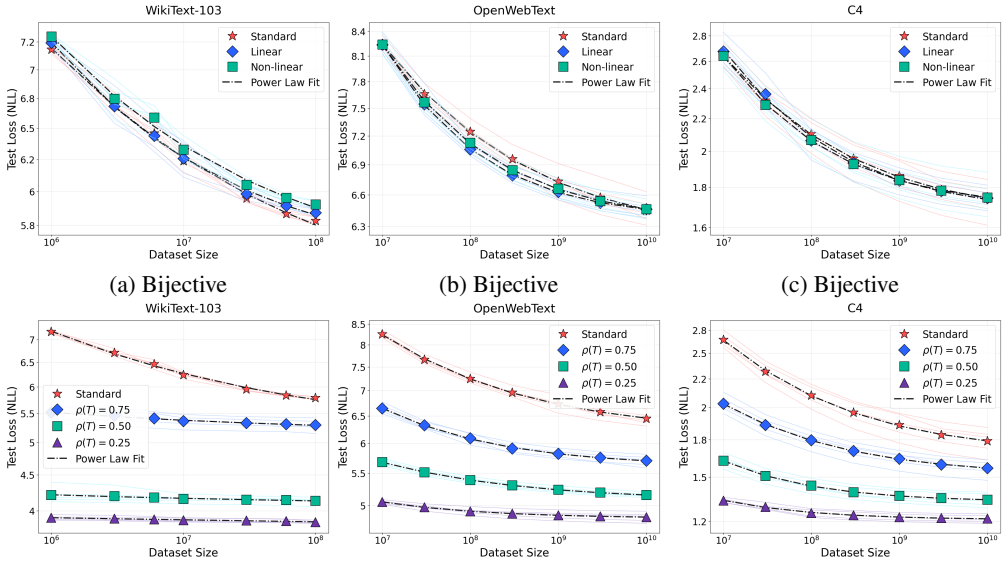


Figure 1: NSLs for language modeling on the WikiText (left), OpenWebText (middle), and C4 (right) datasets. The first row presents results under bijective transformations, the second row corresponds to non-bijective transformations (quantization). In the figures, the thin colored lines represent the test loss curves for individual parameter sizes, while the scaling law estimates (black dotted lines) are fitted using the average test loss values across parameter sizes. Results on negative log-likelihood (NLL) indicate that bijective transformations do not alter the scaling behavior; in contrast, as the degree of information loss increases, the scaling behavior progressively weakens.

where $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is a monotonically increasing function with $\Phi(0) = 0$.

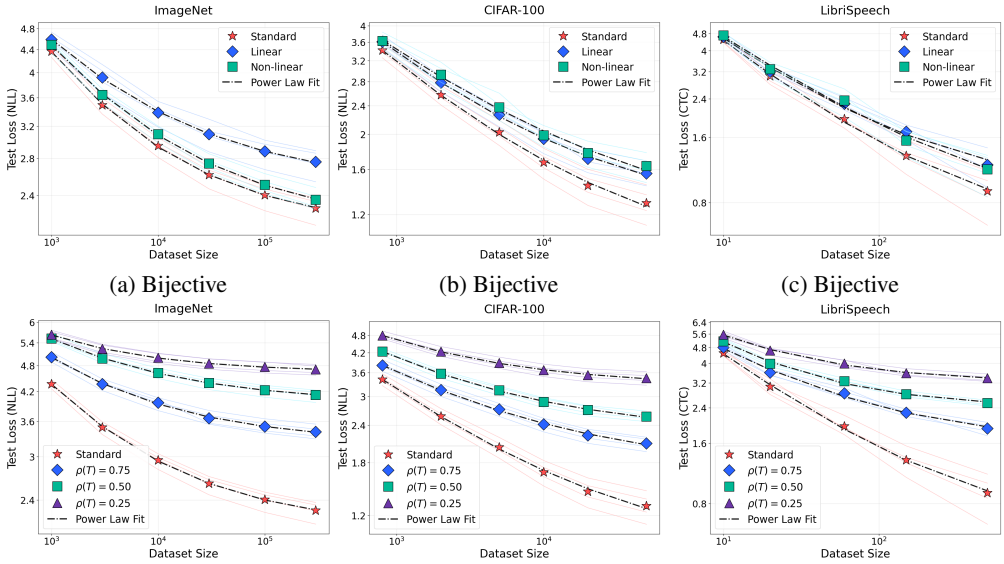
The proof of Lemma 3 can be found at Appendix B. It states that when applying a non-bijective transformation, the best achievable loss on the transformed data is strictly worse than on the original data, which creates an information ceiling bound dependent on ρ even as $D \to \infty$. We now propose an updated scaling law that incorporates both mechanisms.

Proposition 4 (Information-Resolution Scaling Law)**.** Let T be a non-bijective transformation with information resolution $\rho(T) = I(T(X); Y)/I(X; Y) \in (0, 1]$. The expected test loss follows:

$$L(N, D, \rho) = \frac{A}{N^\alpha} + \underbrace{\frac{B}{D^\beta} \cdot \rho(T)^{-\nu}}_{\text{Variance Inflation } \phi_{\text{VI}}} + E + \underbrace{\kappa(1 - \rho(T))^\mu}_{\text{Optimal Loss Shift } \phi_{\text{LS}}}, \tag{4}$$

where $\nu > 0$ governs the variance inflation term $\phi_{\text{VI}} = (B/D^\beta) \cdot \rho(T)^{-\nu}$, which means when $\rho(T) < 1$, more samples are needed to achieve the same estimation precision; E is the original irreducible loss on clean data; $\phi_{\text{LS}} = \kappa(1 - \rho(T))^\mu$ is the shift of optimal loss, where $\kappa > 0$ is the scaling coefficient, and $\mu \in (0, 1]$ governs how quickly the optimal loss shift can be reached. If the transformation is deterministic, then $\mu = 1$ according to Lemma 3; for stochastic corruptions, $\mu < 1$ because averaging over samples could partially recover obscured information.

Equation (4) provides a meta-scaling law: given a source calibration, it predicts the loss surface under changes in information resolution. Table 1 contrasts our formulation with other scaling law formulations under data degradation. Specifically, our formulation satisfies two boundary conditions: (1) when $\rho = 1$ (bijective transformation), it reduces to the standard Chinchilla law; (2) when $\rho \to 0$, the ceiling term dominates and $L \to A/N^\alpha + E + \kappa$. A critical distinction is that, as the amount of data increases, prior quality-aware scaling laws converge to the same loss E regardless of quality, while our formulation predicts a *ρ-dependent loss*. This aligns with information-theoretic limits: no amount of degraded data can recover information that was discarded.



(a) Bijective (b) Bijective (c) Bijective

(d) Non-Bijective (e) Non-Bijective (f) Non-Bijective

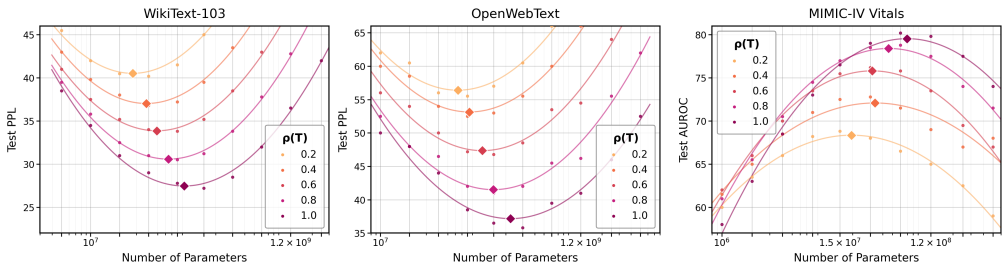
Figure 2: NSLs for vision and speech tasks on the ImageNet-1K (left column), CIFAR-100 (middle column), and LibriSpeech (right column) datasets. Performance is evaluated using negative log-likelihood (NLL) and Connectionist Temporal Classification (CTC) loss. Similarly, the non-bijective transformation (low-rank) shows a much stronger influence on scaling behavior.

3 Primary Empirical Observations

We conduct comprehensive empirical evaluations to validate our hypothesis and proposed information–resolution scaling law framework, aiming to answer the following questions: (**RQ1**) Do standard NSLs remain invariant under bijective transformations? (**RQ2**) How do non-bijective transformations that alter the information resolution of a dataset affect NSLs, and how effectively does our framework capture these effects? (**RQ3**) How can these insights be leveraged in real-world cross-domain applications? We evaluate scaling behavior across diverse modality benchmarks, including language modeling, vision, speech, and medical records. Our implementation is available at `https://github.com/aaronhan223/scaling_law`.

Experimental Setup. We evaluate our framework across four modalities. Language: causal Llama-3-style Transformers [17] with the GPT-2 tokenizer (context 1024) trained on WikiText-103 [36], OpenWebText [22], and C4 [15] for next-token prediction and translation. Vision: ViTs [16] on ImageNet-1K [13] and ResNets [24] on CIFAR-100 [31]. Speech: Wav2Vec2 [4] on LibriSpeech [39]. Clinical time-series: MIMIC-IV [29] ICU vitals and labs. Each scaling-law fit uses 4–6 model capacities and 5–6 dataset sizes. We probe two bijective transformation families: linear (full-rank random matrix on token IDs) and nonlinear (affine coupling layers from the normalizing flows literature [14]), and two non-bijective families: quantization (Eq. (15)) and low-rank projection (Eq. (17)) of embeddings at multiple intensity levels. Full details are deferred to Appendix E.

NSLs Under Various Transformations. Figure 1 (a)–(c) shows that NSLs remain largely invariant under bijective transformations, with linear and nonlinear re-encodings tracing nearly identical scaling curves to the untransformed baseline. In contrast, Figure 1 (d)–(f) shows that progressively reducing $\rho(T)$ via quantization weakens the scaling curves and induces a clear saturation effect at each ρ level, with the asymptote tracking the ρ-dependent loss floor predicted by Eq. (4). Note that transformations such as quantization map multiple token embeddings to the same token ID, effectively reducing the overall vocabulary size. This leads to a decrease in NLL, while perplexity correspondingly increases. Figure 2 shows that the same qualitative pattern holds for vision (ImageNet, CIFAR) and speech (LibriSpeech): bijective transformations of pixel or spectrogram inputs preserve scaling exponents within fitting noise, while non-bijective transformations shift the asymptotic floor in a ρ-dependent manner. Notably, the magnitude of degradation is smaller than in language modeling: classification tasks tolerate substantial information loss before scaling visibly



(a) (b) (c)

Figure 3: Optimal model capacity as a function of $\rho(T)$. Each curve is an iso-ρ slice: model capacity is swept across ten scales at fixed dataset size and fixed $\rho(T)$, and we plot test loss on the y-axis. The U-shape roughly reflects the underfitting-to-overfitting tradeoff; the minimum of each curve marks the compute-optimal capacity $N^*(\rho)$. As $\rho(T)$ decreases, the optimal model size shrinks, consistent with our prediction from Eq. (5). $\rho(T)$ is varied via embedding quantization for the language tasks and quantile binning of clinical time-series for MIMIC-IV.

Algorithm 1 Procedure of Cross-domain Scaling Law Prediction

Require: Source-domain losses $\{L_s(N, D, \rho)\}$ across varying ρ; transformation T.

Ensure: Predicted target-domain exponents $\hat{\beta}_t$ and irreducible loss $\hat{E}_t$.

1: **Estimate info-resolution parameters from source.** Fit Eq. (4) to source-domain losses across different levels of ρ to obtain the source-domain scaling law and transformation-type parameters $\hat{\nu}, \hat{\mu}, \hat{\kappa}$, together with $A_s, \alpha_s, B_s, \beta_s, E_s$.

2: **Compute target information resolution analytically.** Estimate $\rho(T)$ via the transformation-specific formula (e.g. Eqs. (15)-(19)).

3: **Predict target loss surface.** Substitute $\rho(T)$ and $\hat{\nu}, \hat{\mu}, \hat{\kappa}$ into the info-resolution scaling law: $L_t(N, D) = \frac{A_s}{N^{\alpha_s}} + \frac{B_s}{D^{\beta_s}} \cdot \rho(T)^{-\hat{\nu}} + E_s + \hat{\kappa}\big(1 - \rho(T)\big)^{\hat{\mu}}$.

4: **Extract effective Chinchilla-form exponents.** Fit $L_t = \frac{A_t}{N^{\hat{\alpha}_t}} + \frac{B_t}{D^{\hat{\beta}_t}} + \hat{E}_t$ to the analytically predicted surface to obtain $\hat{\beta}_t$ and $\hat{E}_t$.

5: **return** $\hat{\beta}_t,\ \hat{E}_t$.

suffers, likely because the prediction target is a coarse label rather than a high-entropy next-token distribution, and because these tasks rely less on long-range sequential context that quantization and projection most severely disrupt. The redundancy of natural images and the slot-like structure of speech tokens further mean that much of the discarded information is task-irrelevant. In this moderate-resolution regime, both variance inflation and Bayes-risk shift contribute non-negligibly to the observed scaling behavior.

4 Predicting Scaling Across Domains

4.1 Estimating Optimal Model Capacity

Determining the optimal model capacity for a given dataset is one of the key insights provided by scaling laws. Here, we investigate how changes in information resolution affect this optimum, offering practical guidance for model design under varying levels of distortion. Our information-resolution scaling law admits a closed-form prediction for this quantity. Under compute-optimal allocation $C = 6ND$ [27], minimizing Eq. (4) over N yields

$$N^*(C, \rho) = N^*(C, 1) \cdot \rho^{\nu/(\alpha+\beta)}, \qquad D^*(C, \rho) = D^*(C, 1) \cdot \rho^{-\nu/(\alpha+\beta)}, \tag{5}$$

where $N^*(C, 1)$ is the standard Chinchilla-optimal model size on clean data. The optimal model size therefore decreases as a power law in ρ, while the optimal token budget grows by the inverse factor. To validate this prediction empirically, we train ten model capacities at each of several $\rho(T)$ levels on three benchmarks: language modeling on WikiText-103 and OpenWebText, and in-hospital mortality prediction on MIMIC-IV. Dataset size is held fixed so that capacity is the sole varying factor. For

Table 2: Predicted vs. empirical NSL parameters β and E across two cross-domain prediction tasks. "Real" denotes empirically fitted values; Quality-Aware [48] and Sub-Optimal [8] are baselines. Both baselines (approximately) predict $E_t = E_s$ and therefore systematically underestimate the loss-floor shift, while we recover E within a small margin. See Appendix D for full results.

Task / Transformation	ν	$\rho(T)$	**Method**	β	E
		0.75	Real	0.2843	2.6533
			Quality-Aware [48]	0.2975	1.8314
			Sub-Optimal [8]	0.2933	1.8285
			Ours	**0.2961**	**2.7842**
(i) Within Dataset: In-hospital Mortality / Noise Injection	0.12	0.50	Real	0.2441	3.7833
			Quality-Aware [48]	0.2535	1.9768
			Sub-Optimal [8]	0.2541	1.9344
			Ours	**0.2512**	**3.7814**
		0.25	Real	0.1927	5.1371
			Quality-Aware [48]	0.2153	2.3907
			Sub-Optimal [8]	0.2221	2.2415
			Ours	**0.1821**	**4.7839**
(ii) Cross Corpus: Next Token Prediction / Quantization	0.19	0.54	Real	0.3254	4.4132
			Quality-Aware [48]	0.3335	3.3854
			Sub-Optimal [8]	0.3619	3.4662
			Ours	**0.3212**	**4.0012**

language modeling, $\rho(T)$ is varied via quantization of token embeddings (consistent with Figure 1); for mortality prediction, via quantile binning of clinical time-series at increasing coarseness.

Figure 3 confirms a clean monotonic relationship between optimal capacity and $\rho(T)$ across all three tasks, with the empirical curves closely tracking the $\rho^{\nu/(\alpha+\beta)}$ form predicted by Eq. (5). The direction of the effect is theoretically expected: when each sample carries less usable information, the compute-optimal strategy shifts toward processing more samples with a smaller model, rather than allocating compute to a larger model that cannot recover information already destroyed by the transformation. What is more interesting is the magnitude of the shift: even moderate drops in ρ produce substantial reductions in optimal capacity, suggesting that practitioners working with degraded or compressed data may be systematically over-provisioning model size.

4.2 Cross-domain Scaling Prediction

Forecasting scaling behavior in a new domain from a calibrated source law is the central application of our meta-scaling framework. This is most valuable in medical domain where privacy prevents access to abundant training data and compute resources are scarce. We apply Algorithm 1 to two prediction tasks that sit at different points along the applicability spectrum of our framework. **(i) Within dataset noise injection:** predicting scaling laws for 48-hour in-hospital mortality on MIMIC-IV time series under varying levels of injected noise, given the law fit on the clean signals. **(ii) Cross-corpus next-token prediction (more difficult):** predicting scaling laws for MIMIC-IV discharge summaries given the law fit on OpenWebText. The first shares dataset with the source and only varies a known corruption mechanism; the second requires assuming the corpus shift can itself be characterized as a reduction in $\rho(T)$. We discuss the key procedures (steps 2 and 3) below.

Step 2: Computing $\rho(T)$. For *(i) Noise Injection*, the noise distribution is known by construction, so $\rho(T)$ follows in closed form from Eq. (19) by plugging the resulting signal-to-noise ratio into the SNR-based formula: no estimation required. For *(ii) Corpus-Transfer*, we view MIMIC-IV discharge summaries as a *sublanguage* of OpenWebText: same English grammar but a restricted vocabulary and formulaic clinical templates that collapse the diverse expressions of web text into a smaller set of recurring patterns, which acts as a quantization of OpenWebText's token distribution and lets Eq. (15) apply with V and q corresponding to the source and target effective vocabularies. To estimate $\rho(T)$ in this case, we use a compression-based estimator $\hat{\rho}_{\text{compress}} = (C(\mathcal{D}')/|\mathcal{D}'|)/(C(\mathcal{D})/|\mathcal{D}|)$ with `gzip` [7], which jointly captures vocabulary, frequency, syntax, local semantics, and repetition structure, where $C(\cdot)$ denotes compressed size (in bits), $|\mathcal{D}|$ denotes raw corpus size, and $C(\mathcal{D})/|\mathcal{D}|$ is the compression ratio. Details for this and alternative estimators are deferred to Appendix C.

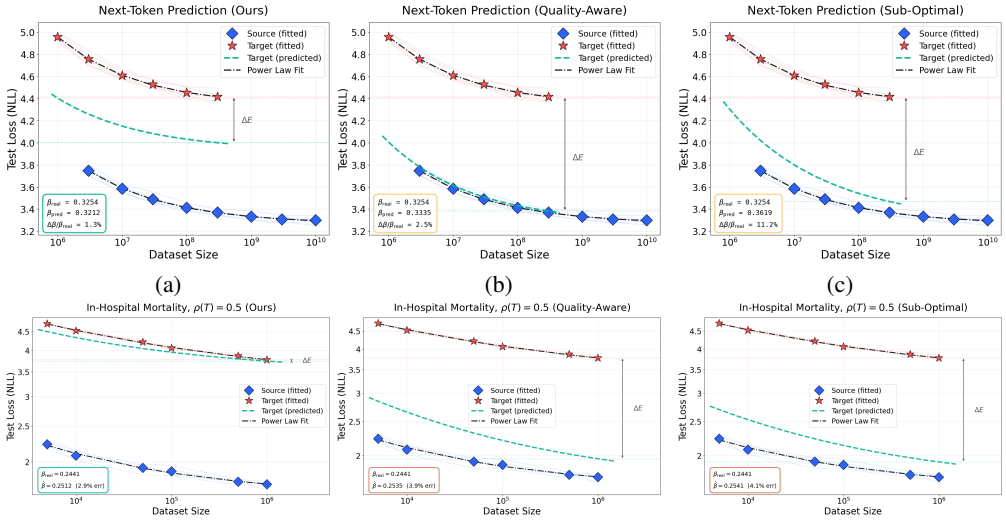


Figure 4: Visualization of cross-domain scaling predictions compared with other baselines. We provide an accurate estimate (within 3%) of both the data scaling exponent β and the irreducible loss E induced by information loss. The Quality-Aware [48] scaling law partially captures the impact of data quality on the scaling exponent, but fails to account for the loss gap, which results in a larger discrepancy between the fitted and predicted scaling law curves. In contrast, the Sub-Optimal [8] baseline shows higher estimation error in both β and E. Our framework adapts to the loss shift as changing the corpus domain or the noise injection level. In contrast, the baselines rely on a constant-E assumption, making them incapable of capturing such variations.

Step 3: Transferability of NSL Parameters. Only $\rho(T)$ is re-estimated per domain; the parameters ν, μ are properties of the transformation type, not the dataset. Specifically, $\mu = 1$ for any deterministic transformation, since deterministic information loss induces a linear relationship between lost mutual information and Bayes-risk elevation (Lemma 3); ν is determined by how transformation-induced distortion propagates through parameter estimation, grounded in the Cramér-Rao bound; and from the proof of Lemma 3, the loss shift takes the form $\Phi\big(I(X;Y)(1-\rho)\big)$, so κ implicitly absorbs the source-domain mutual information $I(X;Y)$. For (i), we reuse $(\hat{\nu}, \hat{\mu}, \hat{\kappa})$ across noise levels. For (ii), transferring κ from OpenWebText is reasonable as both corpora share an autoregressive prediction structure over natural language, making their context–target mutual information comparable. We fit $(\hat{\nu}, \hat{\mu}, \hat{\kappa})$ once on the source-domain ρ-sweep, a leave-one-out fold across noise levels on full MIMIC-IV vital signs for (i), and the OpenWebText quantization sweep for (ii), and reuse them on the target. Besides, the model-capacity exponent α is invariant to data transformations (Prop. 3) and is inherited from the source; only β and E require prediction. The resulting values $\hat{\beta}_t$ and $\hat{E}_t$ obtained in Step 4 already account for the effects of variance inflation and optimal-loss shift.

Results. Table 2 and Figure 4 reports predicted versus fitted β and E for both settings, alongside the Quality-Aware [48] and Sub-Optimal [8] baselines. Our framework recovers the data-scaling exponent β to within 3% relative error, substantially outperforming both baselines. In addition, neither of baselines model a quality-dependent loss floor and accordingly predict an unchanged E.

4.3 Limitations of Current Framework

Our meta-scaling framework is most reliable when the source-to-target shift can be meaningfully cast as a reduction in information resolution, and its assumptions weaken as one moves outside this regime. We highlight three main limitations. **(i) Cross-domain transferability of ν, μ, κ.** The framework treats these as transformation-specific properties: ν from how distortion propagates through estimator variance (Cramér-Rao), $\mu = 1$ for any deterministic transformation, and κ as a loss-shift coefficient that implicitly absorbs the source-domain mutual information $I(X;Y)$. Transferring κ across corpora therefore assumes comparable task-relevant mutual information between source and target: reasonable for OpenWebText→MIMIC-IV (both natural language, both

autoregressive prediction) but increasingly tenuous for unrelated subdomains (e.g., natural text $\rightarrow$ source code), where shared information is minimal. **(ii) Estimating $\rho(T)$.** When the transformation is known by construction (noise injection, explicit quantization), $\rho(T)$ admits a closed form. For implicit cross-corpus shifts, $\rho(T)$ must be estimated from data, and our compression- and n-gram-based estimators only approximate the true ρ_{true}: they may fail when the target introduces new information absent from the source, or when information loss is non-uniform and concentrated on task-critical features. **(iii) Cross-modality scaling.** Our framework currently assumes a single shared modality between source and target; cross-modality prediction (e.g., text$\rightarrow$image) lacks a natural transformation T within the present formulation, and the underlying loss functions are likely incomparable. A promising extension is to define $\rho(T)$ within a shared embedding space such as CLIP, but this remains future work.

5 Conclusion and Future Work

Neural scaling laws are most valuable precisely where they are hardest to fit: in new domains, on new tasks, under data and compute constraints that preclude exhaustive sweeps. This paper addresses that gap through meta-scaling laws that characterize when scaling remains invariant and predict how it changes under data transformations. We show that scaling laws are preserved under bijective transformations of the training data and modified under non-bijective ones in information-theoretically grounded ways. Such modifications are through two distinct mechanisms, variance inflation and Bayes-risk elevation, that together yield the *Information-Resolution Scaling Law*. Within its applicability regime, the framework predicts target-domain scaling from a source calibration and an estimate of $\rho(T)$. Empirically, this enables a law fit on general-purpose text or clean physiological signals to predict scaling behavior on clinical notes and noisy mortality data, recovering data-scaling exponents to within 3% and the ρ-dependent loss floor that prior quality-aware formulations cannot capture. Three directions appear most promising: extending ρ estimation to compositional transformations whose effects do not factor cleanly; deriving cross-modality ρ within shared embedding spaces (e.g., CLIP) to bridge the current single-modality boundary; and using the law's compute-optimal predictions $N^*(C, \rho)$ to guide curation policies that explicitly trade data quality for model size.

Acknowledgments and Disclosure of Funding

Xing Han and Suchi Saria were funded by the Gordon and Betty Moore Foundation grant #12128, and National Science Foundation grant #1840088. The authors declare no competing interests.

References

[1] A. Aghajanyan, L. Yu, A. Conneau, W.-N. Hsu, K. Hambardzumyan, S. Zhang, S. Roller, N. Goyal, O. Levy, and L. Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR, 2023.

[2] I. M. Alabdulmohsin, X. Zhai, A. Kolesnikov, and L. Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425, 2023.

[3] A. Alqahtani, M. Ali, X. Xie, and M. W. Jones. Deep time-series clustering: A review. *Electronics*, 10(23):3001, 2021.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[5] Y. Bansal, B. Ghorbani, A. Garg, B. Zhang, C. Cherry, B. Neyshabur, and O. Firat. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pages 1466–1482. PMLR, 2022.

[6] D. Brandfonbrener, N. Anand, N. Vyas, E. Malach, and S. Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *arXiv preprint arXiv:2411.12925*, 2024.

[7] E. Chang, M. Paltenghi, Y. Li, P.-J. Lin, C. Zhao, P. Huber, Z. Liu, R. Rabatin, Y. Shi, and V. Chandra. Scaling parameter-constrained language models with quality data. *arXiv preprint arXiv:2410.03083*, 2024.

[8] Z. Chen, S. Wang, T. Xiao, Y. Wang, S. Chen, X. Cai, J. He, and J. Wang. Revisiting scaling laws for language models: The role of data quality and training strategies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23881–23899, 2025.

[9] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

[10] L. Choshen, Y. Zhang, and J. Andreas. A hitchhiker's guide to scaling law estimation. *arXiv preprint arXiv:2410.11840*, 2024.

[11] Y. Chu, Z. Han, G. Luo, and X. Gao. Revisiting data scaling law for medical segmentation. *arXiv preprint arXiv:2511.13883*, 2025.

[12] P. C. Cosman, K. L. Oehler, E. A. Riskin, and R. M. Gray. Using vector quantization for image processing. *Proceedings of the IEEE*, 81(9):1326–1341, 2002.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[15] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

[16] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[18] T. D. Edwards, J. Alvey, J. Alsing, N. H. Nguyen, and B. D. Wandelt. Scaling-laws for large time-series models. *arXiv preprint arXiv:2405.13867*, 2024.

[19] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[20] N. C. Frey, R. Soklaski, S. Axelrod, S. Samsi, R. Gomez-Bombarelli, C. W. Coley, and V. Gadepally. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, 2023.

[21] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):4–es, 2007.

[22] A. Gokaslan, V. Cohen, E. Pavlick, and S. Tellex. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

[23] S. Goyal, P. Maini, Z. C. Lipton, A. Raghunathan, and J. Z. Kolter. Scaling laws for data filtering–data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22702–22711, 2024.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[26] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

[27] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models (2022). *arXiv preprint arXiv:2203.15556*, 2022.

[28] A. Hyvärinen, I. Khemakhem, and H. Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10), 2023.

[29] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[30] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[31] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[32] J. Lang, Z. Guo, and S. Huang. A comprehensive study on quantization techniques for large language models. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 224–231. IEEE, 2024.

[33] R. Li, Y. Wei, M. Zhang, N. Yu, H. Hu, and H. Peng. Scalingfilter: Assessing data quality through inverse utilization of scaling laws. *arXiv preprint arXiv:2408.08310*, 2024.

[34] X. Lu, H. Wang, W. Dong, F. Wu, Z. Zheng, and G. Shi. Learning a deep vector quantization network for image compression. *IEEE Access*, 7:118815–118825, 2019.

[35] I. Markovsky and K. Usevich. *Low rank approximation*, volume 139. Springer, 2012.

[36] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models, 2016.

[37] N. Muennighoff, A. Rush, B. Barak, T. Le Scao, N. Tazi, A. Piktus, S. Pyysalo, T. Wolf, and C. A. Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

[38] C. Nair, B. Prabhakar, and D. Shah. On entropy for mixtures of discrete and continuous variables. *arXiv preprint cs/0607075*, 2006.

[39] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[40] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[41] L. Parra, G. Deco, and S. Miesbach. Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation*, 8(2):260–269, 1996.

[42] C. Peng, K. Zhang, M. Lyu, H. Liu, L. Sun, and Y. Wu. Scaling up biomedical vision-language models: Fine-tuning, instruction tuning, and multi-modal learning. *arXiv preprint arXiv:2505.17436*, 2025.

[43] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[44] Y. Ruan, C. J. Maddison, and T. B. Hashimoto. Observational scaling laws and the predictability of langauge model performance. *Advances in Neural Information Processing Systems*, 37:15841–15892, 2024.

[45] J. Scarlett and V. Cevher. An introductory guide to fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*, 2019.

[46] R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581, 2023.

[47] M. Shukor, E. Fini, V. G. T. da Costa, M. Cord, J. Susskind, and A. El-Nouby. Scaling laws for native multimodal models. *arXiv preprint arXiv:2504.07951*, 2025.

[48] A. Subramanyam, Y. Chen, and R. L. Grossman. Scaling laws revisited: modeling the role of data quality in language model pretraining. *arXiv preprint arXiv:2510.03313*, 2025.

[49] H. L. Van Trees. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.

[50] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

[51] D. Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94:103–114, 2017.

[52] J. Ye, P. Liu, T. Sun, J. Zhan, Y. Zhou, and X. Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*, 2024.

[53] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[54] S. Zhang, Q. Liu, N. Usuyama, C. Wong, T. Naumann, and H. Poon. Exploring scaling laws for ehr foundation models. *arXiv preprint arXiv:2505.22964*, 2025.

Supplementary Material for "Meta-Scaling Laws: On the Invariance and Generality of Neural Scaling"

A Extended Related Works

Scaling Laws Across Modalities. Neural scaling laws have become foundational to understanding and predicting model behavior, originating from systematic characterizations in language modeling [30, 27] and subsequently extending across diverse data modalities. In computer vision, analogous power-law relationships have been established for Vision Transformers, with further refinements showing that compute-optimal configurations require joint optimization of width and depth rather than scaling either dimension in isolation [53, 2]. The multimodal setting has received particularly extensive investigation: autoregressive generative modeling across images, video, and image-text data exhibits nearly universal exponents relating optimal model size to compute [25], while contrastive language-image pre-training follows reproducible power-law scaling for zero-shot classification and retrieval [9]. Mixed-modal scaling laws have been derived across seven modalities, explicitly modeling both synergy and competition between modalities as additive terms to unimodal laws [1], with recent work on native multimodal models revealing that early-fusion architectures outperform late-fusion approaches at lower parameter counts [47]. Beyond vision and language, decoder-only transformers for time series forecasting exhibit power-law scaling analogous to language models across diverse temporal domains [18]. In this work, we demonstrate the application of our proposed framework across multiple modalities, highlighting its generalizability to diverse domains.

Predicting Scaling Laws. A growing body of research has developed methods to predict scaling behavior without exhaustive training. Observational approaches demonstrate that model performance can be characterized through low-dimensional capability spaces, bypassing training entirely [44]. Complementary strategies leverage existing models and training trajectories: fitting to intermediate checkpoints improves prediction accuracy, and scaling parameters often transfer across model families [10]. Transfer-based prediction has been extended through loss-to-loss strategies using shifted power laws, enabling prediction across pretraining datasets and to downstream tasks [6]. Beyond model-centric approaches, recent work incorporates data composition, predicting loss as a joint function of model size, training tokens, and domain weights to optimize data mixtures from small-scale runs [47]. However, these methods treat data as a monolithic quantity characterized by volume or coarse domain labels, without principled account of why scaling exponents change across settings, or when they should remain invariant.

Data-Dependent Scaling Laws. Recent work has extended classical scaling laws to account for data quality and composition, revealing several key phenomena. Early investigations into noise and architecture effects on data scaling found that quality degradation minimally impacts scaling exponents, suggesting compensation through increased data volume [5]. However, this view has been refined by studies demonstrating that data utility is heterogeneous: quality subsets exhibit differing scaling behavior, data repetition yields diminishing returns, and optimal filtering strategies must be tailored to available compute budgets [23]. Complementary work has leveraged scaling laws as diagnostic tools for data quality itself, using perplexity differences between models of varying sizes to reveal that scaling exponents vary predictably with data characteristics [33]. More granular analyses have incorporated effective training tokens based on text diversity and syntheticity [7], while large-scale empirical studies across hundreds of models have identified high data density and suboptimal resource allocation as primary drivers of sub-scaling phenomena [8]. Despite these advances, existing formulations lack a unified theoretical framework explaining why different data properties modulate scaling behavior.

Data Transformations and Applications Transforming data into different representations has long been a fundamental operation in machine learning. Invertible transformations, which enable exact reconstruction—including linear mappings and bijective nonlinear functions—have been widely exploited in diverse settings such as normalizing flows [43, 40] and independent component analysis [28]. In contrast, non-invertible transformations inherently discard information, as in quantization, clustering, and dimensionality reduction. Such methods are commonly employed to improve efficiency and reduce memory consumption, and have been extensively used in language modeling

[28, 32], image processing [12, 34], and time-series forecasting [3]. In the following sections, we categorize canonical transformations and analyze how they affect the components of scaling laws.

B Proofs of Propositions and Lemma

Proof of Proposition 1. For any predictor f on original data X, define $g = f \circ T^{-1}$ on transformed data $T(X)$. Then:

$$\mathbb{E}[\ell(g(T(X)), Y)] = \mathbb{E}[\ell(f(T^{-1}(T(X))), Y)] = \mathbb{E}[\ell(f(X), Y)] \tag{6}$$

Since T is invertible, this mapping $f \mapsto f \circ T^{-1}$ is a bijection between hypothesis classes that preserves risk. Taking infimum over all predictors:

$$R^*(T(X)) = \inf_g \mathbb{E}[\ell(g(T(X)), Y)] = \inf_f \mathbb{E}[\ell(f(X), Y)] = R^*(X) \tag{7}$$

Thus the irreducible loss $E = R^*$ is invariant. □

Proof of Proposition 2. Sample complexity lower bounds depend on mutual information $I(X; Y)$ [50]:

$$n \geq \Omega\left(\frac{I(X;Y)}{\epsilon^2}\right) \tag{8}$$

By Lemma 1, $I(T(X); Y) = I(X; Y)$ for invertible T, so the bound is identical.

For the upper bound, any learning algorithm A on original data can be simulated on transformed data: given samples $\{(T(x_i), y_i)\}$, compute $\{(T^{-1}(T(x_i)), y_i)\} = \{(x_i, y_i)\}$ and run A. This simulation has identical sample complexity, so:

$$n^*(\epsilon; T(X)) = n^*(\epsilon; X) \tag{9}$$

Thus the data scaling exponent β is invariant. □

Proof of Proposition 3. The Bayes optimal predictor for transformed data is $g^* = f^* \circ T^{-1}$.

Linear case ($T(x) = Wx$): For a neural network approximating f^*, the composition $f^* \circ W^{-1}$ can be realized by replacing the first layer weights $W_1 \to W_1 W^{-1}$. This requires identical parameter count.

Nonlinear case (T a smooth diffeomorphism): If $f^* \in C^s$ (s-times differentiable), then $g^* = f^* \circ T^{-1} \in C^s$ since composition preserves smoothness. Neural network approximation rates depend only on smoothness [51]:

$$\inf \|f - \mathcal{N}\|_\infty = \Theta(N^{-s/d}) \tag{10}$$

Since f^* and g^* have identical smoothness, they require identical model capacity. Thus the parameter scaling exponent α is invariant. □

Proof of Lemma 3. We prove the result in two steps: deriving the gap for log-loss, then establishing irreversibility.

Step 1: Bayes-optimal loss gap. For log-loss, the Bayes-optimal loss given observation Z equals the conditional entropy: $L^*_Z = H(Y|Z)$. Using $H(Y|Z) = H(Y) - I(Z; Y)$:

$$\begin{aligned} L^*_{X'} - L^*_X &= H(Y|X') - H(Y|X) \\ &= [H(Y) - I(X'; Y)] - [H(Y) - I(X; Y)] \\ &= I(X; Y) - I(X'; Y) = I(X; Y)(1 - \rho(T)) \end{aligned} \tag{11}$$

For general losses, Fano's inequality [45] implies $L^*_{X'} - L^*_X \geq \Phi(I(X; Y) - I(X'; Y))$ where Φ is monotonically increasing with $\Phi(0) = 0$.

Step 2: Irreversibility via Data Processing Inequality. The DPI guarantees $I(X'; Y) \leq I(X; Y)$ for any transformation, with strict inequality when T is non-invertible. Crucially, no post-processing can recover lost information: for any function g, $I(g(X'); Y) \leq I(X'; Y)$. Therefore, no learning

algorithm on X' can achieve loss below $L^*_{X'}$, and the gap $\Phi(I(X;Y)(1-\rho))$ persists regardless of dataset size D or model capacity N.

This distinguishes the ceiling from statistical efficiency loss: efficiency loss vanishes as $D \to \infty$, while the ceiling is a hard information-theoretic limit that no amount of data can overcome. □

Proof of Proposition 4. The result follows from decomposing the expected loss into three components.

(1) Bayes-optimal loss shift. By proof of Lemma 3, the Bayes-optimal loss on transformed data is:

$$L^*_{X'} = L^*_X + \Phi(I(X;Y)(1-\rho)) = E + \kappa(1-\rho)^\mu \tag{12}$$

where $E = L^*_X$ is the clean-data Bayes risk and $\Phi(\cdot) = \kappa(\cdot)^\mu$ for appropriate κ, μ.

(2) Statistical efficiency degradation. Information loss inflates estimator variance. By the Cramér-Rao bound [49], variance scales inversely with Fisher information, which is proportional to $I(X';Y) = \rho \cdot I(X;Y)$. For D samples:

$$\text{Var}(\hat{f}) \propto \frac{1}{D \cdot \rho^\nu} \tag{13}$$

where ν depends on corruption type ($\nu = 1$ for additive noise, $\nu = 2$ for label noise).

(3) Combining terms. Standard bias-variance decomposition gives $\mathbb{E}[L] = L^* + \text{estimation error} + \text{approximation error}$. With sufficient model capacity, approximation error scales as A/N^α. Estimation error scales as $(B/D^\beta) \cdot \rho^{-\nu}$. The Bayes risk is $L^*_{X'} = E + \kappa(1-\rho)^\mu$. Combining:

$$L(N, D, \rho) = \frac{A}{N^\alpha} + \frac{B}{D^\beta} \cdot \rho^{-\nu} + E + \kappa(1-\rho)^\mu \tag{14}$$

The boundary conditions follow directly: $\rho = 1$ gives $\rho^{-\nu} = 1$ and $(1-\rho)^\mu = 0$; $\rho \to 0$ gives $\rho^{-\nu} \to \infty$ (but dominated by ceiling for large D) and $(1-\rho)^\mu \to 1$. □

C Additional Methodological Details

C.1 Information Resolution for Specific Transformations

We discuss practical usages of our proposed framework and its potential applications for predicting scaling laws in different domains. We first derive the corresponding scaling parameters for two common non-bijective transformations. More detailed discussions can be found in Appendix.

Quantization. Consider q-level quantization of a discrete source with vocabulary $\mathcal{V}$, $|\mathcal{V}| = V$. If the original tokens are approximately uniformly distributed, then $H(X) \approx \log V$. After quantization to q centroids, the maximum entropy is $H(X') \leq \log q$. By the data processing inequality applied to the Markov chain $Y \to X \to X'$:

$$\rho_{\text{quant}} = \frac{I(X';Y)}{I(X;Y)} \leq \frac{H(X')}{H(X)} \approx \frac{\log q}{\log V}. \tag{15}$$

The variance inflation term is $\phi_{\text{VI}} = \rho^{-\nu}$ with the exponent ν determined by how quantization error propagates through learning. The optimal loss shift arises because quantization irreversibly merges V/q tokens per cluster on average, discarding fine-grained distinctions:

$$\phi_{\text{LS}} = \kappa\left(1 - \frac{\log q}{\log V}\right) = \kappa(1-\rho). \tag{16}$$

Low-Rank Projection. Let $X \in \mathbb{R}^d$ with covariance $\Sigma = \sum_{i=1}^d \lambda_i u_i u_i^\top$ (eigenvalues sorted descending). Projection onto the top-k subspace, $X' = U_k U_k^\top X$, discards variance in the trailing (d-k) directions. For jointly Gaussian (X, Y), mutual information decomposes across independent components, yielding:

$$\rho_{\text{rank}} = \frac{I(X';Y)}{I(X;Y)} = \frac{\sum_{i=1}^k \lambda_i \cdot r_i^2}{\sum_{i=1}^d \lambda_i \cdot r_i^2} \tag{17}$$

where r_i^2 is the squared correlation between the i-th principal component and Y. When Y loads uniformly across components, this simplifies to the variance ratio $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$. The optimal loss shift is exactly the information in discarded dimensions:

$$\phi_{\text{LS}} = \kappa(1-\rho) = \kappa \cdot \frac{\sum_{i=k+1}^{d} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \tag{18}$$

Since projection is a linear operation, the variance inflation exponent is $\nu = 1$.

Additive Noise. For $X' = X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$ independent of X, the mutual information under a joint Gaussian (X, Y) satisfies:

$$I(X'; Y) = \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_\epsilon^2}\right) = \frac{1}{2} \log(1 + \text{SNR}) \tag{19}$$

Thus $\rho_{\text{noise}} = \log(1 + \text{SNR}) / \log(1 + \text{SNR}_0)$ where SNR_0 is the baseline. In the high-SNR regime, $\rho \approx \text{SNR}/\text{SNR}_0$. From the Cramér-Rao bound, estimator variance scales inversely with Fisher information [49], which is proportional to SNR: this gives $\phi_{\text{VI}} = \rho^{-1}$. Unlike deterministic transformations, additive noise creates a soft ceiling: information is obscured rather than deleted, so ϕ_{LS} grows more gradually, scaling as $\kappa(1-\rho)^\mu$ with $\mu < 1$ reflecting partial recovery through averaging.

C.2 Information Resolution Estimators for Cross-domain Scaling Predictions.

The naive vocabulary-size estimator $\hat{\rho}_{\text{vocab}}(T) = \log q / \log V$ from Eq. (15) provides a useful starting point but rests on two unrealistic assumptions: (i) tokens are approximately uniformly distributed, and (ii) vocabulary size is the dominant axis of cross-corpus variation. Real corpora violate both: word frequencies are heavy-tailed, and shifts between domains involve syntax, local semantics, and repetition structure in addition to vocabulary. Because $\hat{\rho}_{\text{vocab}}$ provides an *upper bound* on the true $\rho(T)$, it systematically underestimates the loss shift $\hat{E}_t = E_s + \hat{\kappa}(1-\hat{\rho})^{\hat{\mu}}$, propagating bias into downstream predictions. We therefore consider two refinements.

Estimator 1: Empirical n-gram entropy ratio. Rather than counting vocabulary, we estimate the entropy of the empirical token distribution directly:

$$\hat{\rho}_{\text{emp}}(T) = \frac{H(\hat{p}_{\mathcal{D}'})}{H(\hat{p}_{\mathcal{D}})}, \tag{20}$$

where $\hat{p}_{\mathcal{D}}$ and $\hat{p}_{\mathcal{D}'}$ are unigram distributions on the source and target corpora. To capture sequential dependencies, we extend this to the 3-gram conditional entropy:

$$\hat{\rho}_{\text{3-gram}}(T) = \frac{\hat{H}_3(\mathcal{D}')}{\hat{H}_3(\mathcal{D})}, \quad \hat{H}_3(\mathcal{D}) = -\sum_{w_1,w_2,w_3} \hat{p}(w_1, w_2, w_3) \log \hat{p}(w_3 \mid w_1 w_2). \tag{21}$$

Pros: captures frequency skew, token diversity, and local syntactic structure simultaneously, so a clinical corpus with formulaic templates registers as lower-ρ even when its raw vocabulary size matches another domain's. In practice, 3- or 4-gram statistics already account for most local syntactic effects. **Cons:** requires computing joint distributions over $|\mathcal{V}|^n$ cells, which becomes memory- and sample-bound for large vocabularies; estimates degrade for rare n-grams and require smoothing; and it remains blind to long-range semantic structure beyond the n-gram window.

Estimator 2: Compression-based. Inspired by recent work on data quality in language models [7], we use a general-purpose compressor (gzip) as a model-free probe of information content. The compression ratio is

$$\hat{\rho}_{\text{compress}}(T) = \frac{C(\mathcal{D}')/|\mathcal{D}'|}{C(\mathcal{D})/|\mathcal{D}|}, \tag{22}$$

where $C(\cdot)$ denotes compressed size in bits and $|\cdot|$ denotes raw corpus size. A more compressible corpus is, by Kolmogorov-complexity intuition, lower in usable information per token and hence lower in ρ.

Pros: simultaneously reflects vocabulary, frequency, syntax, local semantics, and repetition structure without explicitly modeling any of them; computationally cheaper than constructing high-order

n-gram tables; entirely non-parametric, so no smoothing or vocabulary-size hyperparameter is required. **Cons:** the compressor's inductive bias is fixed and may not align with task-relevant information: gzip rewards local repetition but is insensitive to long-range semantic structure that a downstream language model would actually exploit; absolute compression ratios depend on corpus size and the compressor's dictionary, so only *ratios* are interpretable.

Empirical comparison. Table 5 reports $\hat{\rho}(T)$ from all three estimators on the OpenWebText→MIMIC-IV discharge-summary transfer, alongside the resulting predicted loss shift $\hat{E}_t$ and its gap to the empirical E_t. The vocabulary-based estimator yields $\hat{\rho}_{\text{vocab}} \approx 0.71$ and the largest $\hat{E}_t$ gap, consistent with its upper-bound bias: by inflating ρ, it shrinks $(1-\hat{\rho})^{\hat{\mu}}$ and understates the loss-floor elevation. The 3-gram estimator returns a lower $\hat{\rho}$: clinical text's templated structure makes its conditional entropy meaningfully smaller than OpenWebText's, and tightens the $\hat{E}_t$ gap accordingly. The compression-based estimator produces the lowest $\hat{\rho}$ and the smallest gap, reflecting its sensitivity to the formulaic repetition that dominates discharge summaries (e.g., section headers, recurring drug names and abbreviations), exactly the redundancy that n-gram statistics partially miss because such patterns span variable distances. Across estimators, the ordering of $\hat{E}_t$ accuracy tracks the ordering of $\hat{\rho}$ richness: the more dimensions of the cross-corpus shift the estimator captures, the closer the predicted loss floor sits to the empirical one. The remaining gap under $\hat{\rho}_{\text{compress}}$ is small enough to be consistent with our theoretical framework's assumption that ν, μ, κ transfer cleanly between deterministic transformations on natural-language corpora.

C.3 Robustness of the Information Resolution Scaling Law

Holding the calibrated coefficients fixed, an error in $\hat{\rho}$ affects both the data term and the predicted loss floor. For $0 < \hat{\rho} < 1$, Eq. (4) gives

$$\frac{\partial \hat{E}_t}{\partial \hat{\rho}} = -\kappa\mu(1-\hat{\rho})^{\mu-1}, \qquad \frac{\partial \log \phi_{\text{VI}}}{\partial \log \hat{\rho}} = -\nu, \tag{23}$$

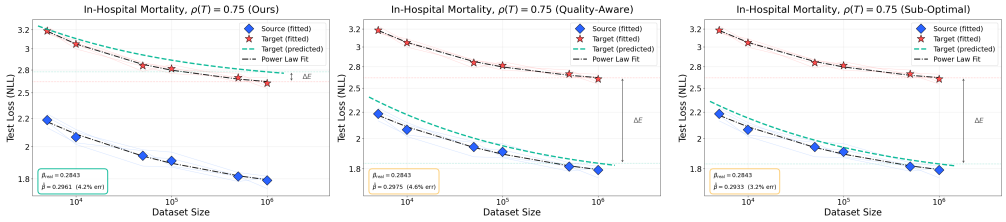
where $\hat{E}_t = E_s + \kappa(1-\hat{\rho})^{\mu}$ and $\phi_{\text{VI}} = (B/D^{\beta})\hat{\rho}^{-\nu}$. Thus, a small relative error $\epsilon = \Delta\hat{\rho}/\hat{\rho}$ induces a relative data-term error of approximately $-\nu\epsilon$. For $\mu = 1$, the loss-floor error is exactly $\Delta\hat{E}_t = -\kappa\Delta\hat{\rho}$. Using the cross-corpus coefficients in Table 5 ($\nu = 0.19$, $\kappa = 2.61$, $\mu = 1$, $\hat{\rho} = 0.54$), a 10% relative perturbation in $\hat{\rho}$ changes the data term by approximately 1.9% and the predicted floor by 0.141 in absolute loss. Increasing D reduces the absolute error in the data term, but does not remove the floor error. Sensitivity can be larger near the boundaries: the data term grows rapidly as $\hat{\rho} \to 0$, and for $\mu < 1$ the floor derivative diverges as $\hat{\rho} \to 1$.

These expressions quantify sensitivity of the predicted loss surface. At fixed ρ, the factor $\rho^{-\nu}$ rescales B without changing β: $BD^{-\beta}\rho^{-\nu} = (B\rho^{-\nu})D^{-\beta}$. Sensitivity of an effective exponent obtained through a constrained finite-range fit therefore depends on the fitting protocol; it cannot be inferred from ρ alone. The estimator comparison in Table 5 complements this local analysis by showing how different proxies for ρ affect the predicted loss floor.

C.4 Cost of Obtaining Coefficients of Information Resolution Scaling Law

The framework replaces repeated target-domain scaling sweeps with a source calibration that can be reused under compatible transformations. Algorithm 1 requires a source sweep over model size, dataset size, and transformation strength to fit the base law and its information-resolution coefficients. Parameters describing the distortion mechanism can be reused when the transformation family and modeling assumptions remain applicable. A justified closed-form expression for ρ further avoids estimating it from the target corpus, but does not eliminate the initial calibration cost.

The loss-shift coefficient κ requires particular care. For the exact log-loss identity with task-specific mutual information, $\mu = 1$ and $\kappa = I(X;Y)$, so κ depends on the source distribution. Transferring it assumes comparable task-relevant information in the source and target; distant transfers may instead require target-side calibration across transformation strengths. The computational benefit is therefore greatest when several predictions share a valid source calibration. Residual loss-floor errors can reflect either an inaccurate $\hat{\rho}$ or a change in κ, and do not by themselves distinguish these explanations.



(a) (b) (c)

Figure 5: We compare our method with baselines under different noise injection levels for the in-hospital mortality prediction task on MIMIC-IV. The reported results for our method follow the same experimental settings as in Table 2 of the main paper. Aside from the gap induced by the irreducible loss, all three frameworks achieve similarly strong estimation performance on β.

C.5 Additional Discussions

Scope of scalar sufficiency. For a fixed source distribution and prediction target Y, information resolution exactly determines the Bayes loss-floor shift under log-loss:

$$L^*_{X'} - L^*_X = H(Y \mid X') - H(Y \mid X) = I(X;Y)(1 - \rho(T)). \tag{24}$$

Two transformations with the same task-specific ρ therefore induce the same Bayes loss shift in this setting. This identity holds for both deterministic and stochastic input transformations. It does not establish that ρ alone determines the full finite-scale scaling law, which also depends on the calibrated coefficients and assumptions about approximation and estimation error. In particular, a fitted $\mu \neq 1$ should be interpreted as a phenomenological extension when using an information proxy or modeling finite-scale behavior, rather than as a consequence of stochasticity alone.

Task dependence and composition. A transformation can preserve information relevant to one target while discarding features needed for another, so ρ is task-relative. For a pipeline $X' = T_1(X)$, $X'' = T_2(X')$ satisfying $Y \to X \to X' \to X''$, data processing implies $\rho(T_2 \circ T_1) \leq \rho(T_1)$. However, resolutions measured separately on the original input distribution do not determine the composite resolution: the second transformation acts on the information retained by the first. Evaluating the second ratio on that intermediate distribution gives the exact product $\rho(T_2 \circ T_1) = \rho(T_1)I(X'';Y)/I(X';Y)$ when $I(X';Y) > 0$. A pipeline therefore requires information about the intermediate representation, beyond standalone transformation summaries.

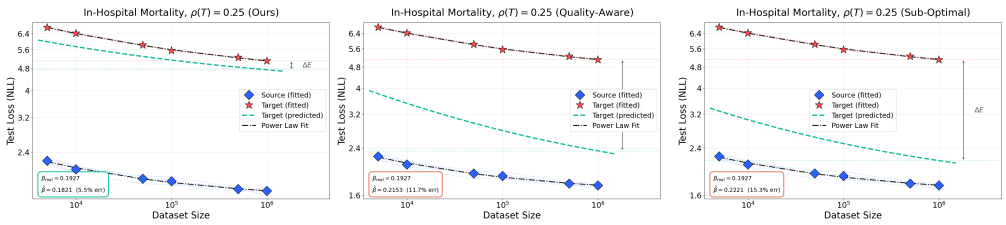
Emergent behavior. Nonlinear or discontinuous evaluation metrics can produce apparent emergence from smoothly improving model predictions [46]. Such changes need not contradict a smooth loss prediction. A genuine transition in the loss-scaling regime, however, invalidates the corresponding base-law approximation. Our framework inherits the validity range of the underlying Chinchilla-form law and does not independently account for such transitions.

D Additional Results

In this section, we present additional experimental results supplementary to the main experiments. Figures 5 and 6 illustrate the cross-corpus next-token prediction experiments when $\rho(T)$ is set to 0.75 and 0.25, respectively. Tables 3 and 4 present scaling law prediction results on the ImageNet-1K dataset under two transformation types: image quantization and noise injection. Tables 5 through 8 report the complete information-resolution scaling law parameter results for the scaling law prediction experiments supplementary to Table 2.

E Experiment Details

All experiments are run on $8\times$ NVIDIA A5500 GPUs with 24 GB memory per GPU.



(a) (b) (c)

Figure 6: At higher noise injection levels, our method still provides reasonable estimates (within an acceptable error range) for the key scaling law exponents. In contrast, for the baselines, not only does the irreducible loss persist, but the estimation of β also shifts under these conditions.

Table 3: Cross-domain prediction results on the ImageNet-1K dataset. We evaluate scaling law parameter estimation under image quantization transformations and report all relevant parameters in the corresponding results.

Parameter	Formula	Values	Ground Truth
ρ	Pixel number to point	0.25	—
$\rho^{-\nu}$	—	1.233	—
B_{eff}	$B_s \cdot \rho^{-\nu}$	43.16	—
$\kappa(1-\rho)^{\mu}$	—	2.025	—
$\hat{E}_t$	$E_s + \kappa(1-\rho)^{\mu}$	4.315	4.46
$\hat{\beta}_t$	refitted	0.225	0.197
Fixed parameters (from source domain fitting)			
$E_s = 2.29$ (irreducible loss)			
$\beta_s = 0.28$ (source data-scaling exponent)			
$\nu = 0.15$, $\kappa = 2.70$, $\mu = 1.0$ (transformation-type parameters)			
$\alpha_s = 0.31$, $A_s = 20.03$, $B_s = 34.87$			

E.1 Dataset and Task Information.

We first evaluate scaling behavior on standard language modeling tasks, including next-token prediction and neural machine translation. For next-token prediction, we use two widely adopted benchmarks: WikiText-103 [36] and OpenWebText [22]. WikiText-103 comprises over 100 million tokens drawn from verified "Good" and "Featured" Wikipedia articles. OpenWebText is substantially larger, containing over 10 billion tokens across more diverse domains, and consists of high-quality web content curated from a large collection of URLs. For the neural machine translation task, we use subsets of the Colossal Clean Crawled Corpus (C4) [15], which contains over 156 billion tokens collected from more than 365 million domains and has been used to train models such as T5 and the Switch Transformer [19]. We evaluate scaling behavior for vision models using image classification tasks on standard benchmarks such as ImageNet-1K [13] and CIFAR-100 [31]. In addition, we evaluate scaling behavior on a speech recognition task using the LibriSpeech dataset [39], a large-scale corpus derived from audiobooks that contains approximately 1,000 hours of English speech. We also study a healthcare-domain time-series classification task using the MIMIC-IV dataset [29], a comprehensive database of de-identified electronic health records from patients admitted to medical centers. Our focus is on clinical time-series data, including high-frequency vital signs (e.g., heart rate and blood pressure) and laboratory measurements recorded during ICU stays.

E.2 Implementation Details.

For language modeling, we conduct pretraining experiments using a Llama 3 style causal Transformer architecture [17]. We employ the GPT-2 tokenizer with a context length of 1024. For each scaling law study, we consider 6 dataset sizes and 5 model sizes for each corresponding dataset. We focus on two types of bijective transformations: linear transformations, which apply a random matrix to token IDs, and nonlinear transformations, which apply a multi-layer stochastic neural network

Table 4: Cross-domain prediction results on the ImageNet-1K dataset. We evaluate scaling law parameter estimation under noise injection and report all relevant parameters in the corresponding results. These experiments demonstrate that Info-Resolution Scaling Law provides reasonably accurate predictions for the key scaling law exponents under both types of transformations.

Parameter	Formula	ρ	Real
ρ	SNR from Eq. (19)	0.75	—
$\rho^{-\nu}$	—	1.044	—
B_{eff}	$B_s \cdot \rho^{-\nu}$	29.95	—
$\kappa(1-\rho)^{\mu}$	—	0.588	—
$\hat{E}_t$	$E_s + \kappa(1-\rho)^{\mu}$	2.928	2.732
$\hat{\beta}_t$	refitted	0.277	0.252
Fixed parameters (from source domain fitting)			
$E_s = 2.34$ (irreducible loss)			
$\beta_s = 0.31$ (source data-scaling exponent)			
$\nu = 0.13$, $\kappa = 2.35$, $\mu = 1.0$ (transformation-type parameters)			
$\alpha_s = 0.27$, $A_s = 21.25$, $B_s = 28.69$			

Table 5: Cross-corpus scaling prediction (setting ii) results for the next-token prediction task, supplementary to Table 2. We additionally compare different ρ estimators against the ground truth.

Parameter	Formula	ρ vocab	ρ 3-gram	ρ compress	Ground Truth
ρ	estimated	0.71	0.57	0.54	—
$\rho^{-\nu}$	—	1.066	1.112	1.124	—
B_{eff}	$B_s \cdot \rho^{-\nu}$	47.97	50.04	50.58	—
$\kappa(1-\rho)^{\mu}$	—	0.757	1.122	1.201	—
$\hat{E}_t$	$E_s + \kappa(1-\rho)^{\mu}$	3.557	3.922	4.001	4.413
ΔE	$\lvert\hat{E} - E_{\text{real}}\rvert$	0.856	0.491	0.412	—
$\hat{\beta}_t$	refitted	0.298	0.319	0.321	0.325
Fixed parameters (from source domain fitting)					
$A_s = 24.96$, $\alpha_s = 0.35$ (model-capacity terms, invariant)					
$B_s = 45.02$, $\beta_s = 0.33$ (source data-scaling)					
$E_s = 2.80$ (source irreducible loss)					
$\nu = 0.19$, $\kappa = 2.61$, $\mu = 1.0$ (transformation-type parameters)					

to the resulting token embeddings before they are fed into the model; note that, the transformation network is not trained jointly with the model. For non-bijective transformations, we apply quantization and low-rank projection at varying levels to the token embeddings. For ImageNet-1K, we use Vision Transformer [16] models with four different capacities and six dataset sizes. For CIFAR-100, we instead use ResNet architectures [24] with four different capacities and six dataset sizes. For the LibriSpeech dataset, we employ Wav2Vec2 [4] models with four different model sizes and five dataset sizes.

E.3 Hyper-parameters and Model Configurations

Tables 9 and 10 summarize the dataset configurations across different modalities used in the scaling law experiments. Table 11 provides comprehensive details of the model architectures and training configurations employed across modalities.

F Broader Impact

This work aims to make neural scaling laws transferable: fit once on a well-resourced source domain and applied to settings where running a full sweep is infeasible. The most immediate beneficiaries are domains operating under structural data and compute constraints: clinical AI, scientific model-

Table 6: Within dataset noise injection ($\rho(T) = 0.75$) scaling prediction (setting i) results for the in-hospital mortality task, supplementary to Table 2.

Parameter	Formula	Values	Ground Truth
ρ	SNR from Eq. 16	0.75	—
$\rho^{-\nu}$	—	1.035	—
B_{eff}	$B_s \cdot \rho^{-\nu}$	12.42	—
$\kappa(1-\rho)^{\mu}$	—	1.05	—
$\hat{E}_t$	$E_s + \kappa(1-\rho)^{\mu}$	2.54	2.36
ΔE	$\lvert\hat{E} - E_{\text{real}}\rvert$	0.18	—
$\hat{\beta}_t$	refitted	0.296	0.284
Fixed parameters (from source domain fitting)			
$E_s = 1.491$ (irreducible loss)			
$\beta_s = 0.322$ (source data-scaling exponent)			
$\nu = 0.121$, $\kappa = 3.102$, $\mu = 0.730$ (transformation-type parameters)			

Table 7: Within dataset noise injection ($\rho(T) = 0.5$) scaling prediction (setting i) results for the in-hospital mortality task, supplementary to Table 2.

Parameter	Formula	Values	Ground Truth
ρ	SNR from Eq. 16	0.5	—
$\rho^{-\nu}$	—	1.087	—
B_{eff}	$B_s \cdot \rho^{-\nu}$	13.04	—
$\kappa(1-\rho)^{\mu}$	—	1.878	—
$\hat{E}_t$	$E_s + \kappa(1-\rho)^{\mu}$	3.427	3.378
ΔE	$\lvert\hat{E} - E_{\text{real}}\rvert$	0.049	—
$\hat{\beta}_t$	refitted	0.251	0.244
Fixed parameters (from source domain fitting)			
$E_s = 1.493$ (irreducible loss)			
$\beta_s = 0.320$ (source data-scaling exponent)			
$\nu = 0.124$, $\kappa = 3.101$, $\mu = 0.730$ (transformation-type parameters)			

ing, low-resource languages, and applications in regulated industries where privacy or cost prevents the kind of large-scale sweeps that yielded existing scaling laws. By providing principled forecasts of how performance will scale before training, our framework can reduce wasted compute, lower the energy footprint of model development, and democratize access to scaling-informed model design beyond labs with frontier-scale resources. We see particular value in healthcare, where over- or under-provisioned models carry direct patient-safety consequences and where data access is strictly capped by privacy regulations. Potential risks are modest but worth noting: predicted scaling curves are point estimates with their own uncertainty, and over-reliance on them, particularly in safety-critical domains, could substitute extrapolation for empirical validation. We therefore view our framework as a planning tool that complements, rather than replaces, careful evaluation at deployment scale.

G Dataset Licenses and Terms of Use

All datasets used in this work are publicly available and were used in accordance with their respective licenses and terms. **WikiText-103** [36] is released under the Creative Commons Attribution-ShareAlike (CC BY-SA 3.0) license. **OpenWebText** [22] is distributed under the Creative Commons CC0 1.0 Universal (public domain) license, with the underlying content sourced from publicly accessible Reddit-linked URLs. **C4** [15] is released by AllenAI under the Open Data Commons Attribution License (ODC-BY), with the underlying Common Crawl data subject to Common Crawl's terms of use. **ImageNet-1K** [13] is available for non-commercial research and educational use under the ImageNet terms of access; we use it solely for academic research consistent with

Table 8: Within dataset noise injection ($\rho(T) = 0.25$) scaling prediction (setting i) results for the in-hospital mortality task, supplementary to Table 2.

Parameter	Formula	Values	Ground Truth
ρ	SNR from Eq. 16	0.25	—
$\rho^{-\nu}$	—	1.178	—
B_{eff}	$B_s \cdot \rho^{-\nu}$	14.14	—
$\kappa(1-\rho)^{\mu}$	—	2.479	—
$\hat{E}_t$	$E_s + \kappa(1-\rho)^{\mu}$	3.933	4.146
ΔE	$\lvert\hat{E} - E_{\text{real}}\rvert$	0.213	—
$\hat{\beta}_t$	refitted	0.182	0.193
Fixed parameters (from source domain fitting)			
$E_s = 1.487$ (irreducible loss)			
$\beta_s = 0.323$ (source data-scaling exponent)			
$\nu = 0.119$, $\kappa = 3.104$, $\mu = 0.730$ (transformation-type parameters)			

Table 9: Language modeling dataset configurations.

Property	Value
WikiText-103	
Source	HuggingFace `wikitext-103-v1`
Tokenizer	GPT-2
Vocabulary size	50,257
Sequence length	1,024 tokens (chunked, non-overlapping)
Train sizes tested	100K, 300K, 1M, 3M, 10M, 30M, 100M, 300M, 1B, 3B tokens
Train batch size	32
OpenWebText	
Source	HuggingFace `openwebtext`
Split	98% train / 1% val / 1% test
Tokenizer	GPT-2
Vocabulary size	50,257
Sequence length	1,024 tokens
Train sizes tested	3M, 10M, 30M, 100M, 300M, 1B, 3B, 10B tokens

these terms. **CIFAR-100** [31] is released for research purposes by the Canadian Institute for Advanced Research and is freely available for academic use. **LibriSpeech** [39] is distributed under the Creative Commons Attribution 4.0 (CC BY 4.0) license, derived from public-domain LibriVox audiobooks. **MIMIC-IV** [29] is released under the PhysioNet Credentialed Health Data License 1.5.0 and requires completion of CITI "Data or Specimens Only Research" training and a signed data use agreement; all authors who accessed MIMIC-IV completed the required credentialing, and no protected health information is reproduced in this paper. We use each dataset solely for the non-commercial academic research purposes permitted by its license.

Table 10: Vision and speech dataset configurations.

Property	Value
ImageNet	
Classes	1,000
Image size	224×224
Patch size (ViT)	$16 \times 16 \rightarrow 196$ patches $+1$ CLS $= 197$ tokens
Normalization	mean $= [0.485, 0.456, 0.406]$, std $= [0.229, 0.224, 0.225]$
Train transform	RandomResizedCrop(224) + RandomHorizontalFlip + Normalize
Val/test transform	Resize(256) + CenterCrop(224) + Normalize
Train sizes tested	1K, 3K, 10K, 30K, 100K, 300K samples
Batch size	128 (train and eval)
Epochs	100
Sampling	Stratified, nested (smaller sets are true subsets of larger)
LibriSpeech	
Train splits	train-clean-100 + train-clean-360 + train-other-500
Dev / test splits	dev-clean / test-clean
Sample rate	16,000 Hz
Max audio length	20.0 seconds (truncated)
Avg utterance duration	6.5 seconds
CTC vocabulary	32 (character-level)
CNN downsampling	$320\times$
Train sizes tested	10h $\approx$ 5,538 samples; 50h $\approx$ 27,692; 100h $\approx$ 55,384; 500h $\approx$ 276,923
Batch size	32
Metric	WER + CTC loss

Table 11: Model architecture and training configurations across modalities.

Parameter / Property	Text (LM)	Image (ViT)	Speech (Wav2Vec2)
Model Architecture			
Size Index 0	Hidden 16, Layers 2, Heads 2, FFN 64, Params ∼827K	ViT-Tiny: Embed 192, Depth 12, Heads 3, FFN 768, Params ∼5.5M	wav2vec2-tiny: Hidden 256, Layers 4, Heads 4, FFN 1024, Params ∼4M
Size Index 1	Hidden 32, Layers 4, Heads 4, FFN 128, Params ∼1.7M	ViT-Small: Embed 384, Depth 12, Heads 6, FFN 1536, Params ∼22M	wav2vec2-small: Hidden 384, Layers 6, Heads 6, FFN 1536, Params ∼13M
Size Index 2	Hidden 64, Layers 6, Heads 8, FFN 256, Params ∼3.6M	ViT-Base: Embed 768, Depth 12, Heads 12, FFN 3072, Params ∼86M	wav2vec2-medium: Hidden 512, Layers 8, Heads 8, FFN 2048, Params ∼30M
Size Index 3	Hidden 128, Layers 8, Heads 8, FFN 512, Params ∼8.1M	ViT-Large: Embed 1024, Depth 24, Heads 16, FFN 4096, Params ∼307M	wav2vec2-large: Hidden 640, Layers 10, Heads 10, FFN 2560, Params ∼55M
Size Index 4	Hidden 256, Layers 10, Heads 8, FFN 1024, Params ∼21M	—	wav2vec2-xlarge: Hidden 768, Layers 12, Heads 12, FFN 3072, Params ∼91M
Size Index 5	Hidden 512, Layers 12, Heads 12, FFN 2048, Params ∼64M	—	—
Size Index 6	Hidden 640, Layers 24, Heads 16, FFN 2560, Params ∼151M	—	—
Size Index 7	Hidden 768, Layers 36, Heads 16, FFN 3072, Params ∼294M	—	—
Size Index 8	Hidden 1024, Layers 48, Heads 32, FFN 4096, Params ∼657M	—	—
Training Hyperparameters			
Optimizer	AdamW	AdamW	AdamW
Learning rate	5×10^{-4}	1×10^{-4}	5×10^{-4}
β_1, β_2	0.9, 0.95	0.9, 0.999	0.9, 0.95
Weight decay	0.01	0.05	0.01
Grad clip	1	1	1
LR schedule	Linear warmup + linear decay to 0	Linear warmup + linear decay to 0	Linear warmup + linear decay to 0
Warmup steps	1,000	1,000	1,000
Precision	FP16	FP16	FP16
Max steps / epochs	1000 epochs	100 epochs	1000 epochs

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe Info-Resolution Scaling Law's contributions (identify key factors that alter scaling laws, propose a general formulation, and leverage such formulation for several downstream applications), all of which are directly supported by experiments in Sections 3. No claim exceeds what is demonstrated in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4.3 discusses the limitations of scaling law predictions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The detailed assumptions and proofs can be found at Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix E discusses all the experimental details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets are publicly accessible to credentialed researchers. Detailed dataset informtion in Appendix E.1. Code is available at Section 3.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix E discusses all related experimental setting/details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The scaling law experiments are expensive and time-consuming to run, so no error bars are reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E discussed this information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research does not involve human subjects or participant interaction. All experiments use de-identified datasets accessed under data use agreements of corresponding platform, which we comply with fully, including not transmitting data to external APIs and not releasing personally identifiable information. We do not foresee harmful societal consequences from this work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix F discusses the potential societal impacts of the work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new pretrained model or dataset. The released code trains transformer models for various tasks on existing credentialed datasets, which pose no high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The dataset creators are properly cited in Appendix E.1 and Section 3. Licenses and terms of use for them are listed in Appendix G. No code bases are released for baseline models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The experiment code is released via an anonymized repository. No new datasets or model weights are released. We included a README file in the repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments use pre-existing datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects. All experiments use pre-existing datasets, and IRB approval is not required.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are used only for writing, editing, and surveying assistance. This usage did not influence the scientific methodology, experimental design, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.