LEARNING SHRINKS THE HARD TAIL: TRAINING-DEPENDENT INFERENCE SCALING IN A SOLVABLE LINEAR MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

We analyze neural scaling laws in a solvable model of last-layer fine-tuning where targets have intrinsic, instance-heterogeneous difficulty. In our Latent Instance Difficulty (LID) model, each input's target variance is governed by a latent "precision" drawn from a heavy-tailed distribution. While generalization loss recovers standard scaling laws, our main contribution connects this to inference. The pass@k failure rate exhibits a power-law decay, $k^{-\beta_{\text{eff}}}$, but the observed exponent β_{eff} is training-dependent. It grows with sample size N before saturating at an intrinsic limit β set by the difficulty distribution's tail. This coupling reveals that learning shrinks the "hard tail" of the error distribution: improvements in the model's generalization error steepen the pass@k curve until irreducible target variance dominates. The LID model yields testable, closed-form predictions for this behavior, including a compute-allocation rule that favors training before saturation and inference attempts after. We validate these predictions in simulations and on CIFAR-10H, where human-label variance provides a realistic difficulty measure.

1 INTRODUCTION

The remarkable success of large-scale machine learning models is tightly linked to empirically observed and theoretically understood scaling laws, which characterize performance improvements with increasing data, model size, or training time (Kaplan et al., 2020; Hestness et al., 2017; Rosenfeld et al., 2019; Bahri et al., 2024; Maloney et al., 2022; Bordelon et al., 2024). These laws have been crucial for predicting learning curves and optimizing resource allocation in the large dataset size N and number of parameters P regime. Most classical analyses focus on the *training* side, relating generalization loss $\mathcal{L}_{\text{gen}}$ to N (and sometimes P) through properties of the data distribution such as spectral decay (Bartlett et al., 2020; Hastie et al., 2020; Bordelon et al., 2020) or tractable model classes (Maloney et al., 2022; Tay et al., 2022).

A complementary paradigm emphasizes *inference-time* compute: repeated attempts, best-of-N, and search can produce substantial gains on difficult reasoning tasks even without further training, typically evaluated with pass@k under a verifier (Snell et al., 2024; Brown et al., 2024). While practical methods for exploiting inference-time compute are rapidly evolving (see Sec. 2), and some explanations for their success have been given (Levi, 2024; Schaeffer et al., 2025), a basic theoretical question remains: *how does training progress shape performance scaling at inference time?*

In many real-world settings the input–output relationship is not deterministic; some instances are intrinsically more variable than others (Arpit et al., 2017; Northcutt et al., 2021). Such instance-level heterogeneity affects both stages: during training we observe only a single noisy realization per input, and during inference we may compare a model prediction to multiple fresh realizations under pass@k. This motivates a minimal model that explicitly ties training and inference through *instance difficulty*.

Setting. We adopt a deliberately simple but analyzable framework that mirrors common practice in fine-tuning: *last-layer (linear) regression on fixed features* in high dimension, with intrinsically stochastic targets. Each instance $\mathbf{x}$ carries a latent *precision* $\tau_{\mathbf{x}}$ (its "difficulty"), drawn from a heavy-tailed distribution, which controls the variance of its target around the mean $\mathbf{x}^\top \boldsymbol{\theta}^*$. Training observes one realization $y \sim Y^*_{\mathbf{x}}$ per input and fits a linear head (ridge/OLS). At inference, we evaluate pass@k with a perfect verifier by drawing k fresh realizations per test input and asking whether at least one lies within a fixed tolerance of the model prediction.

Overview and contributions.

- **A solvable LID model for linear fine-tuning.** We formalize the Latent Instance Difficulty (LID) model in the high-dimensional linear setting. Training with a single realization per input reduces to ridge/OLS and recovers established generalization scaling with respect to sample size N, dimension d, and spectral exponent α (including the $1/N$ tail in the classical regime when the average target variance is finite, i.e., $\beta > 2$).
- **Training–inference coupling via a two-tail law.** We show that the distribution of *single-trial success probabilities* under pass@k acquires two regularly varying components: an *intrinsic* tail determined by the latent difficulty distribution (exponent β) and a *finite-N* tail governed by the model's error relative to the mean target (exponent $\gamma(N) \propto 1/\mathcal{L}_{\text{gen}}(N)$). Averaging over trials yields a mixture pass@k law from which the *effective* inference exponent $\beta_{\text{eff}}(N) = \min\{\beta, \gamma(N)\}$ emerges. Hence the observed pass@k slope *increases with N* and *saturates* at the intrinsic difficulty index β.
- **Predictions and implications.** The theory predicts (i) a crossover surface in (N, k) separating a finite-N (bias-dominated) region from an intrinsic-tail region; (ii) a saturating $\beta_{\text{eff}}(N)$ curve; and (iii) continued prefactor improvements with N even after the slope has plateaued. These lead to a simple compute-allocation rule: invest in training until $\beta_{\text{eff}}(N)$ is near β, then prioritize inference attempts.
- **Evidence in simulation and a real-data proxy.** Controlled simulations confirm the $1/N$ training tail (with the correct intercept), the steepening of pass@k with N, and a saturating $\hat{\beta}_{\text{eff}}(N)$. A CIFAR-10H last-layer fine-tuning experiment, where human-label variance provides realistic instance difficulty, exhibits analogous behavior.

Taken together, these results provide a clean, testable baseline that unifies training and inference scaling in a setting that captures a widely used fine-tuning regime. The model makes explicit *when* test-time compute should help, *how far* its benefits can go, and *how* those benefits depend on training.

2 Related Work

Due to space constraints, we defer a detailed literature review to App. A and focus here on the prior research most central to our contribution.

Generalization Scaling Laws A large body of work has established that the generalization loss of deep networks often scales as a predictable power law with resources like dataset size N or model parameters (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann and et al., 2022). Theoretical frameworks seek to explain these laws by appealing to properties of the data, such as its spectral decay, or the model architecture (Bahri et al., 2021; Maloney et al., 2022). Our work builds on the standard scaling of generalization loss with N, which serves as the "training" component of our unified model.

Inference-Time Scaling A parallel line of work has shown that performance on difficult reasoning tasks can be dramatically improved by increasing compute at inference time, even without further training (Snell et al., 2024; Brown et al., 2024). The dominant methods involve generating multiple candidate solutions and selecting the best one, often evaluated with the pass@k metric. While some theoretical models for this phenomenon have been proposed (Levi, 2024; Schaeffer et al., 2025), they typically analyze inference in isolation.

To our knowledge, a simple, solvable model that analytically connects the progress of training (i.e., the decrease in generalization error) to the scaling of inference performance is still missing. Our work aims to provide exactly this unified view.

The rest of the paper is organized as follows: In Sec. 3, we present the LID model. We provide our main results for training and inference scaling laws in Sec. 4. In Sec. 5 we present an example of regression performed on an inherently stochastically labeled dataset, showing that the LID is a reasonable proxy for a real-world task. We conclude in Sec. 6.

3 The Latent Instance Difficulty Setting

Real-world datasets exhibit significant instance-level heterogeneity: some image labels are ambiguous, incurring higher annotator disagreement (Peterson et al., 2019; Northcutt et al., 2021), and some reasoning problems are intrinsically harder than others, leading to more variable outputs. Standard

homogeneous-noise assumptions overlook this complexity. Addressing this, and connecting it to the distinct scaling behaviors observed during training versus inference (especially when multiple inference attempts can be verified against a correct solution; cf. Section 1), motivates our setting. Intuitively, factors that increase training difficulty (harder to learn the mean) also shape inference reliability (harder to match a fresh realization).

Last-layer fine-tuning view. We instantiate *Latent Instance Difficulty* (LID) within the common fine-tuning regime: a frozen representation produces features $\mathbf{x} \in \mathbb{R}^d$ and we learn a linear head $\mathbf{x}^\top \boldsymbol{\theta}$. Each instance carries a latent *precision* (its "easiness") $\tau_{\mathbf{x}}$ that controls the variance of its target around the mean $\mathbf{x}^\top \boldsymbol{\theta}^*$. Training observes *one* realization y per $\mathbf{x}$; at inference, pass@k compares the model's prediction to k fresh realizations from the same instance-specific target distribution.

Definition 3.1 (Latent Instance Difficulty (LID) Model)**.** The data generation process for an observation $(\mathbf{x}, y)$ is:

1. **Features.** An input feature vector $\mathbf{x} \in \mathbb{R}^d$ is drawn from a distribution $p(\mathbf{x})$ with zero mean $\mathbb{E}[\mathbf{x}] = 0$ and covariance $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$. We assume eigenvalues σ_j^2 exhibit power-law decay $\sigma_j^2 \propto j^{-(1+\alpha)}$ for $j = 1, \ldots, d$, with $\alpha > 0$.

2. **Latent difficulty.** Associated with $\mathbf{x}$ is a latent *difficulty precision* $\tau_{\mathbf{x}} \in (0, \infty)$, drawn independently of $\mathbf{x}$ from
$$\tau_{\mathbf{x}} \sim \mathrm{Gamma}(\text{shape} = \beta/2,\ \text{rate} = 1), \tag{1}$$
where $\beta > 0$ controls the near-zero tail. Smaller β increases the mass at very low precision (high intrinsic variance), corresponding to "hard" instances.

3. **Stochastic target.** Conditional on $(\mathbf{x}, \tau_{\mathbf{x}})$, the instance target is Gaussian around the mean relationship $f^*(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}^*$ with variance inversely proportional to $\tau_{\mathbf{x}}$:
$$Y_{\mathbf{x}}^* \sim \mathcal{N}\big(\text{mean} = \mathbf{x}^T\boldsymbol{\theta}^*,\ \text{variance} = \sigma_\eta^2/\tau_{\mathbf{x}}\big), \quad \text{where } \sigma_\eta^2 > 0 \text{ is a global scale.} \tag{2}$$

4. **Training labels.** The observed label is a single realization
$$y \sim Y_{\mathbf{x}}^* \quad \text{equivalently} \quad y = \mathbf{x}^T\boldsymbol{\theta}^* + \eta,\ \ \eta \sim \mathcal{N}(0,\ \sigma_\eta^2/\tau_{\mathbf{x}}). \tag{3}$$

Comments. (i) The power-law spectrum in Item 1 abstracts benign-feature regimes observed in practice and used in linear scaling analyses (Maloney et al., 2022; Levi and Oz, 2023; 2024). (ii) The Gamma family in Eq. (1) is chosen for analytic convenience; all results that govern inference scaling depend only on the *near-zero* tail $\Pr(\tau_{\mathbf{x}} \le t) \asymp t^{\beta/2}$, so other distributions with the same tail index yield the same exponents (Levi, 2024). (iii) Independence of $\tau_{\mathbf{x}}$ and $\mathbf{x}$ simplifies exposition; allowing correlation is a natural extension and would primarily affect constants and crossover locations rather than exponents. (iv) We will denote the model's *bias relative to the mean target* on a fresh test feature by $\mathcal{E}_{\text{gen}}(\mathbf{x}) = \mathbf{x}^\top\hat{\boldsymbol{\theta}} - \mathbf{x}^\top\boldsymbol{\theta}^*$; its distribution is governed by the training procedure and sample size N (see Sec. 4).

Corollary 1. The average variance of the target around its mean is
$$\mathbb{E}\big[(Y_{\mathbf{x}}^* - \mathbf{x}^T\boldsymbol{\theta}^*)^2\big] = \mathbb{E}_{\mathbf{x}}[\mathrm{Var}(Y_{\mathbf{x}}^* \mid \mathbf{x})] = \sigma_\eta^2\,\mathbb{E}\left[\frac{1}{\tau_{\mathbf{x}}}\right]. \tag{4}$$

For $\tau_{\mathbf{x}} \sim \mathrm{Gamma}(\beta/2, 1)$, $\mathbb{E}[1/\tau_{\mathbf{x}}] = \frac{\Gamma(\beta/2-1)}{\Gamma(\beta/2)} = \frac{2}{\beta-2}$ when $\beta > 2$.

Assumption 3.2 (Finite average target variance)**.** We assume $\mathbb{E}[(Y_{\mathbf{x}}^* - \mathbf{x}^T\boldsymbol{\theta}^*)^2] < \infty$, i.e., $\beta > 2$.

Assumption 3.2 enables standard high-dimensional ridge/OLS analyses of $\mathcal{L}_{\text{gen}}$ (training/generalization scaling). Importantly, our inference-time results (pass@k scaling) rely only on the small-τ tail and continue to hold in the sense of exponents even when $\beta \le 2$ (though constants and the onset of asymptotics may change), provided the learned predictor is consistent.

3.1 Training Setup

We consider learning the mean relationship $f^*(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\theta}^*$ from a dataset $\mathcal{D}_N = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where each y_i is a *single* realization sampled according to Def. 3.1. In the fine-tuning view, $\mathbf{x}_i$ are

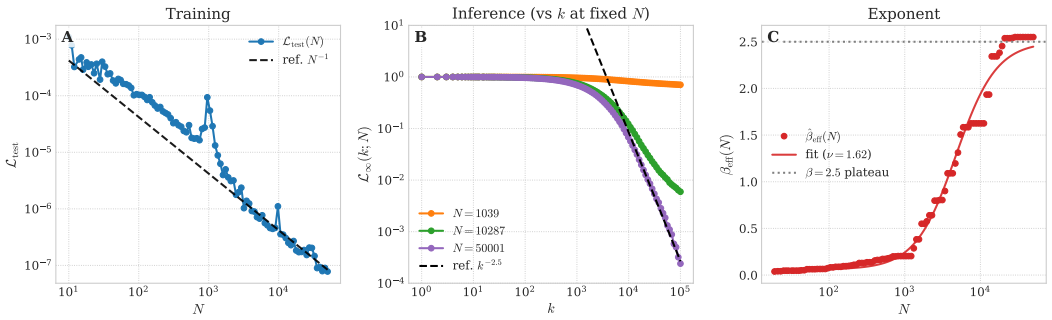

Figure 1: **Training and inference-time scaling laws in the LID setting.** *Left:* Generalization error $\mathcal{L}_{\text{gen}}$ vs. N, showing double descent and the two classical regimes. *Center:* Pass@k inference failure rate $\mathcal{L}_{\text{inf}}(k)$ vs. k for several $N > d$, with asymptotic slope $-\beta$ (dashed black) once the mean is well learned. *Right:* The effective inference exponent $\beta_{\text{eff}}(N)$ extracted from the local log–log slope in a fixed k-window. The dotted line marks the asymptote β; the solid curve is the fit $\beta_{\text{eff}}(N) = \beta - \Delta/(1 + c_\beta N^\nu)$. We use $\lambda = 10^{-9}$ and $\sigma_\eta = 10^{-3}$.

frozen features from a pretrained backbone and we train only a linear head. The learner observes $(\mathbf{x}_i, y_i)$ and estimates $\hat{\boldsymbol{\theta}}$ by minimizing a ridge objective against the *realized* labels while evaluation is always against the *mean* target:

$$\mathcal{L}_{\text{train}}(\hat{\boldsymbol{\theta}}) \;=\; \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\theta}})^2 \;+\; \lambda\|\hat{\boldsymbol{\theta}}\|_2^2, \tag{5}$$

where $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ (last-layer parameters) and $\lambda \geq 0$ is a small regularizer.

The minimizer admits the standard closed form

$$\hat{\boldsymbol{\theta}}_\lambda \;=\; \arg\min_{\hat{\boldsymbol{\theta}}} \mathcal{L}_{\text{train}}(\hat{\boldsymbol{\theta}}) \;=\; (N^{-1}X^\top X + \lambda I_d)^{-1}\, N^{-1}X^\top \mathbf{y}, \tag{6}$$

where $X \in \mathbb{R}^{N\times d}$ stacks the features and $\mathbf{y} \in \mathbb{R}^N$ the realized labels from equation 3. We will take the ridgeless limit $\lambda \to 0$ when well-posed; in the overparameterized case $N < d$ we use the standard scaling $\lambda = \tilde{\lambda}/N$ to recover the minimum-norm interpolator (see Sec. 4 and App.B).

Bias relative to the mean target. Throughout, we evaluate generalization against the *mean* signal $\mathbf{x}^\top \boldsymbol{\theta}^*$, and denote the model's instancewise deviation by

$$\mathcal{E}_{\text{gen}}(\mathbf{x}) \;:=\; \mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top \boldsymbol{\theta}^*. \tag{7}$$

The test (generalization) loss is $\mathcal{L}_{\text{gen}}(N, \lambda) \;=\; \mathbb{E}_{\mathbf{x},\mathcal{D}_N}\big[\mathcal{E}_{\text{gen}}(\mathbf{x})^2\big]$. Under Ass. 3.2, the effective training noise has finite average variance $\sigma^2_{\text{noise}} = \sigma_\eta^2\, \mathbb{E}[1/\tau_{\mathbf{x}}] = \frac{2\sigma_\eta^2}{\beta-2}$, so standard high-dimensional ridge/OLS tools apply. In Sec. 4 we analyze how $\mathcal{L}_{\text{gen}}$ scales with N, d, and the feature spectrum exponent α, and how the resulting distribution of $\mathcal{E}_{\text{gen}}(\mathbf{x})$ controls the finite-N inference behavior (pass@k) and the effective inference exponent $\beta_{\text{eff}}(N)$.

4 Theoretical Analysis of Scaling Laws

We analyze two coupled laws: (i) the dependence of the generalization loss $\mathcal{L}_{\text{gen}}$ on sample size N, and (ii) the pass@k inference failure $\mathcal{L}_{\text{inf}}$ on the number of trials k. In our fine-tuning view (last-layer regression on frozen features), $\mathcal{L}_{\text{gen}}$ controls the distribution of the instancewise deviation $\mathcal{E}_{\text{gen}}(\mathbf{x}) = \mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top \boldsymbol{\theta}^*$, which in turn governs the finite-N behavior of $\mathcal{L}_{\text{inf}}$. As N grows, $\mathcal{L}_{\text{gen}}$ shrinks and the $\mathcal{E}_{\text{gen}}(\mathbf{x})$-induced penalty in $\mathcal{L}_{\text{inf}}$ recedes; the observed inference slope transitions from a finite-N regime to the latent-difficulty asymptote $-\beta$. We quantify this transition empirically via $\beta_{\text{eff}}(N)$ (Fig. 1, right).

4.1 Training Scaling Law ($\mathcal{L}_{\text{GEN}}$ vs. N)

We evaluate generalization against the mean target, so the test loss is

$$\mathcal{L}_{\text{gen}}(N, \lambda) \;=\; \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\,\mathcal{D}_N}\big[\,\mathcal{E}_{\text{gen}}(\mathbf{x})^2\,\big] \;=\; \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),\,\mathcal{D}_N}\big[\,\big(\mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top \boldsymbol{\theta}^*\big)^2\,\big]. \tag{8}$$

In high dimensions, $\mathcal{L}_{\text{gen}}$ depends critically on the N–d ratio and the spectrum of $\mathbf{\Sigma}$; we follow the standard decomposition into regimes (Belkin et al., 2019; Hastie et al., 2020). Under Ass. 3.2, the effective label noise has finite average variance $\sigma^2_{\text{noise}} = \sigma_\eta^2\,\mathbb{E}[1/\tau_{\mathbf{x}}]$, and classical ridge/OLS results apply (see App.B for a concise derivation in our notation).

Overparameterized ($N < d$). With ridgeless (or lightly regularized) interpolation, the error is governed by the minimum-norm bias and the data spectrum exponent α (Belkin et al., 2018; Bartlett et al., 2020; Mei and Montanari, 2020; Cohen et al., 2021; Hastie et al., 2020; Wu and Xu, 2020; Maloney et al., 2022). For power-law spectra $\sigma_j^2 \propto j^{-(1+\alpha)}$,

$$\mathcal{L}_{\text{gen}}(N) \;\propto\; P_N\, N^{-\alpha}, \qquad (N < d), \tag{9}$$

where P_N is a benign prefactor encapsulating spectrum and teacher alignment. This captures the structured-data difficulty along trailing eigendirections.

Underparameterized ($N > d$). When samples exceed parameters, the variance term dominates and decays at the parametric rate,

$$\mathcal{L}_{\text{gen}}(N) \;\propto\; \sigma_\eta^2\,\mathbb{E}[1/\tau_{\mathbf{x}}]\;\, d/N, \qquad (N \gg d), \tag{10}$$

with the usual $1/N$ slope and a noise-controlled constant that reflects the mean target being learned from single realizations.

Transition ($N \approx d$). Near interpolation the estimator is ill-conditioned and $\mathcal{L}_{\text{gen}}$ exhibits a peak ("double descent"), whose height and width depend on λ and the spectrum (Belkin et al., 2019).

The left panel of Fig. 1 confirms these scalings in our LID setting, and provides the $\mathcal{L}_{\text{gen}}(N)$ baselines used to interpret the finite-N inference behavior discussed next.

4.2 Inference Scaling Law ($\mathcal{L}_{\text{INF}}$ vs. k)

We analyze inference through the pass@k metric (Snell et al., 2024; Brown et al., 2024), i.e., the probability that at least one of k independent trials matches the target within tolerance. In the LID setting, for a test instance $\mathbf{x}$ with latent precision $\tau_{\mathbf{x}}$,

$$Y^*_{\mathbf{x}} \sim \mathcal{N}\big(\mathbf{x}^\top \boldsymbol{\theta}^*,\, \sigma_\eta^2/\tau_{\mathbf{x}}\big), \tag{11}$$

and the model outputs $\hat{y} = \mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda$. In trial j we draw $y_j \sim Y^*_{\mathbf{x}}$ so that

$$e_j \;=\; \hat{y} - y_j \;=\; \big(\mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top \boldsymbol{\theta}^*\big) - \eta_j \;=\; \mathcal{E}_{\text{gen}}(\mathbf{x}) - \eta_j, \qquad \eta_j \sim \mathcal{N}\big(0,\, \sigma_\eta^2/\tau_{\mathbf{x}}\big). \tag{12}$$

A trial succeeds if $|e_j| \le \delta$.

Assumption 4.1 (Perfect Verification)**.** We assume a perfect verifier that declares overall success iff at least one of the k trials satisfies $|e_j| \le \delta$.

Let $p(\mathbf{x}, \tau_{\mathbf{x}}) = \mathbb{P}_{\eta \sim \mathcal{N}(0, \sigma_\eta^2/\tau_{\mathbf{x}})}\big(|\mathcal{E}_{\text{gen}}(\mathbf{x}) - \eta| > \delta\big)$ be the single-trial failure probability. Assuming independence across trials, the pass@k failure is

$$\mathcal{L}_{\text{inf}}(k) \;=\; \mathbb{E}_{\mathbf{x},\,\tau_{\mathbf{x}},\,\mathcal{D}_N}\big[\,p(\mathbf{x}, \tau_{\mathbf{x}})^{\,k}\,\big] \;=\; 1 - \text{pass@}k. \tag{13}$$

Remark (model-draw vs. target-draw pass@k). In practical pass@k for LLMs one draws k *model* outputs and verifies them against a fixed target; here we draw k *target* realizations and compare them to a fixed predictor $\hat{y} = x^\top \hat{\theta}_\lambda$. For our scaling claims these protocols are equivalent under a small tolerance window and smooth local densities: the single-try success is (2δ) times the *local density at the gap* $B_N(x) = x^\top(\hat{\theta}_\lambda - \theta^*)$, so exchanging where the randomness lives (model vs. target) leaves the k-dependence and the small-τ_x control unchanged (constants may differ). We formalize this equivalence and its conditions at the start of App. B.4.

Asymptotic tail (bias-free reference). When the mean is learned well so that $|\mathcal{E}_{\text{gen}}(\mathbf{x})|$ is negligible relative to $\sigma_\eta/\sqrt{\tau_{\mathbf{x}}}$ for the small-$\tau_{\mathbf{x}}$ instances that dominate inference failures, a small-window expansion gives

$$1 - p(\mathbf{x}, \tau_{\mathbf{x}}) \;\approx\; c_\delta\,\sqrt{\tau_{\mathbf{x}}}, \qquad c_\delta \;=\; 2\delta/\sqrt{2\pi}\,\sigma_\eta. \tag{14}$$

Then $p(\mathbf{x}, \tau_{\mathbf{x}})^{\,k} \approx \exp\big(-k c_\delta \sqrt{\tau_{\mathbf{x}}}\big)$ and averaging over $\tau_{\mathbf{x}} \sim \text{Gamma}(\beta/2, 1)$ yields (see App. B.3)

$$\mathcal{L}_{\text{inf}}(k) \;\sim\; \tilde{P}_{\beta,\delta,\sigma_\eta}\, k^{-\beta}, \qquad \tilde{P}_{\beta,\delta,\sigma_\eta} \;=\; \frac{2\,\Gamma(\beta)}{\Gamma(\beta/2)\,(c_\delta)^\beta}, \qquad k \to \infty. \tag{15}$$

This reproduces the $-\beta$ slope governed by the small-τ tail of the LID prior, independent of d/N.

Finite-N correction and $\beta_{\text{eff}}(N)$ evolution. For moderate $\tau_{\mathbf{x}} = \Theta(1)$, finite-$N$ mean error $\mathcal{E}_{\text{gen}}(\mathbf{x})$ suppresses the per-trial success probability. A Gaussian CDF expansion gives, for small δ,

$$1 - p(\mathbf{x}, \tau_{\mathbf{x}}) \;\approx\; \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\delta}{\sigma_\eta} \sqrt{\tau_{\mathbf{x}}}\, \exp\Big(- \frac{\mathcal{E}_{\text{gen}}(\mathbf{x})^2\, \tau_{\mathbf{x}}}{2\sigma_\eta^2} \Big). \tag{16}$$

Thus very small success probability arises either from $\tau_{\mathbf{x}} \downarrow 0$ (intrinsic difficulty) *or* from large $|\mathcal{E}_{\text{gen}}(\mathbf{x})|$ at typical $\tau_{\mathbf{x}}$. Under standard linear generalization, $B_N(\mathbf{x}) := \mathcal{E}_{\text{gen}}(\mathbf{x})$ is approximately Gaussian with $\mathrm{Var}[B_N] = \Theta(\mathcal{L}_{\text{gen}}(N))$.

Using equation 16 and Tauberian arguments for Laplace–Stieltjes transforms (under the conditions in the assumptions below), we obtain the two-tail mixture law (details in App. C).

$$\mathcal{L}_{\text{inf}}(k; N) \;=\; \tilde{P}\, k^{-\beta} \;+\; \tilde{P}_N(N)\, k^{-\gamma(N)}\, (1 + o(1)), \qquad k \to \infty, \tag{17}$$

where $\tilde{P} > 0$ depends on $(\beta, \delta, \sigma_\eta)$ as in equation 15, $\tilde{P}_N(N) > 0$, and

$$\gamma(N) \;=\; \Theta\Big(\frac{1}{\mathrm{Var}[B_N]}\Big) \;=\; \Theta\Big(\frac{1}{\mathcal{L}_{\text{gen}}(N)}\Big). \tag{18}$$

Assumptions for the Tauberian step. The derivation relies only on mild regular-variation and smoothness conditions, summarized in App. B.3 (regularly varying $\Pr(\tau_x \le t)$ near 0, a uniform small-window expansion for the single-try success probability, sub-Gaussian $B_N(x)$ with $\mathrm{Var} B_N \asymp \mathcal{L}_{\text{gen}}(N)$, and conditional independence of trials).

Proposition 4.2 (Training-dependent effective exponent)**.** *Fix a k-window $[k_1, k_2]$ for which the local slope is constant. For sufficiently large k_1 and any N, the effective exponent for $\mathcal{L}_{inf}(k; N)$ satisfies*

$$\beta_{\text{eff}}(N; [k_1, k_2]) \;\approx\; \min\big\{\beta,\; \gamma(N)\big\}, \tag{19}$$

with $\gamma(N)$ given by equation 18. Consequently, as N increases and $\mathcal{L}_{gen}(N)$ decreases, $\gamma(N)$ grows and $\beta_{\text{eff}}(N)$ monotonically approaches the intrinsic LID exponent β.

In practice we summarize this monotone saturation by the empirical fit

$$\beta_{\text{eff}}(N) \;=\; \beta - \Delta/(1 + c_\beta\, N^\nu), \tag{20}$$

where (Δ, c_β, ν) depend weakly on the k-window and the estimator. The center and right panels of Fig. 1 illustrate: (i) for fixed $N > d$, $\mathcal{L}_{\text{inf}}(k)$ exhibits a slope that steepens with N; (ii) $\beta_{\text{eff}}(N)$ increases with N and plateaus at β.

4.3 Compute Allocation Tradeoff (with training-dependent $\beta_{\text{eff}}(N)$)

Having derived the finite N inference scaling behavior, we can incorporate both training scaling and inference scaling to study the optimal budget allocation between the two. We consider a fixed compute budget C that must be split between training samples (cost c_N each) and inference trials (cost c_k each) $C = Nc_N + kc_k$, s.t. $k = \frac{C - \tilde{N}}{c_k}$, $\tilde{N} := Nc_N$. We minimize a weighted objective

$$\mathcal{L}_{\text{tot}}(\tilde{N}) \;=\; R\, \mathcal{L}_{\text{gen}}(N) \;+\; \mathcal{L}_{\text{inf}}\Big(k = \tfrac{C - \tilde{N}}{c_k}\,;\, N\Big), \quad \text{training/inference weight} = R. \tag{21}$$

In the classical (under-parameterized) regime $N \gg d$ we have $\mathcal{L}_{\text{gen}}(N) \approx P_N N^{-\gamma}$ with $\gamma = 1$ (Sec. 4.1); more generally take $\gamma \in \{1, \alpha\}$ depending on d/N. For inference, Sec. 4.2 established the mixture law leading to a training-dependent effective slope; over the practical k-window used for evaluation we model

$$\mathcal{L}_{\text{inf}}(k; N) \;\approx\; \tilde{P}(N)\, k^{-\beta_{\text{eff}}(N)}, \qquad \beta_{\text{eff}}(N) \uparrow \beta \text{ as } N \to \infty, \tag{22}$$

where $\beta_{\text{eff}}(N)$ is monotone increasing and saturating (e.g., $\beta_{\text{eff}}(N) = \beta - \Delta/(1 + CN^\nu)$), and $\tilde{P}(N)$ is a slowly varying prefactor capturing residual bias effects.

With $\tilde{N} = Nc_N$ and $k = (C - \tilde{N})/c_k$, the budget-constrained objective becomes

$$\mathcal{L}_{\text{tot}}(\tilde{N}) \;=\; R\, P_N\, c_N^{\gamma}\, \tilde{N}^{-\gamma} \;+\; \tilde{P}(N)\, c_k^{\beta_{\text{eff}}(N)}\, (C - \tilde{N})^{-\beta_{\text{eff}}(N)}. \tag{23}$$

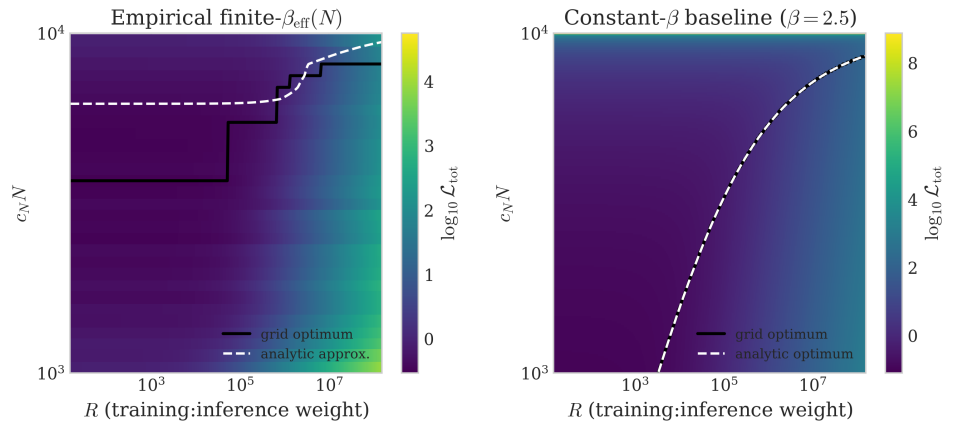

Figure 2: **Compute allocation with a training-dependent inference exponent.** *Left:* empirical finite-N case. We plot $\log_{10} \mathcal{L}_{\text{tot}}(N, k; R)$ with R the training:inference weight ratio and $C = Nc_N + kc_k$. Black: grid optimum $\tilde{N}_\star$; white dashed: analytic approximation from equation 24, using $\beta_{\text{eff}}(N)$ (local log–log slope of $\mathcal{L}_{\text{inf}}$) and its discrete derivative. *Right:* constant-β baseline calibrated from the same data; white dashed is the closed-form analytic optimum. Contours clearly shift toward larger N in the finite-N panel, consistent with the $-\beta'_{\text{eff}}(N) \log(\cdot)$ correction.

Proposition 4.3 (Optimal allocation with training-dependent $\beta_{\text{eff}}(N)$)**.** *Assume $\mathcal{L}_{gen}(N) \approx P_N N^{-\gamma}$ and $\mathcal{L}_{inf}(k; N) \approx \tilde{P}(N) k^{-\beta_{\text{eff}}(N)}$. Let $N = \tilde{N}/c_N$. An interior optimum $\tilde{N}_* \in (0, C)$ satisfies*

$$R\, P_N\, c_N^{\gamma}\, \gamma\, \tilde{N}^{-(\gamma+1)} \;=\; \tilde{P}(N)\, c_k^{\beta_{\text{eff}}(N)}\, (C - \tilde{N})^{-\beta_{\text{eff}}(N)} \left[\frac{\beta_{\text{eff}}(N)}{C - \tilde{N}} \;-\; \frac{\beta'_{\text{eff}}(N)}{c_N} \ln\!\left(\frac{C - \tilde{N}}{c_k} \right) \right]. \tag{24}$$

If $\tilde{P}(N)$ is slowly varying and $|\beta'_{\text{eff}}(N)|$ is small over the relevant range, equation 24 simplifies to the quasi-static balance

$$R\, P_N\, c_N^{\gamma}\, \gamma\, \tilde{N}^{-(\gamma+1)} \;\approx\; \tilde{P}\, c_k^{\beta}\, \beta\, (C - \tilde{N})^{-(\beta+1)}. \tag{25}$$

Interpretation. Compared to the constant-β condition, equation 24 includes a *new logarithmic term* proportional to $-\beta'_{\text{eff}}(N) \ln\big((C - \tilde{N})/c_k\big)$. When the budget is such that $k = (C - \tilde{N})/c_k$ lies *below* the LID-dominated window (so $\beta'_{\text{eff}}(N) > 0$ and $\ln((C - \tilde{N})/c_k) > 0$), this term increases the marginal benefit of training, shifting the optimum towards larger N. Once $\beta_{\text{eff}}(N)$ has saturated (so $\beta'_{\text{eff}}(N) \approx 0$), equation 24 reduces to the constant-exponent balance in equation 25 (recovering the classical tradeoff with β replaced by $\beta_{\text{eff}}(N)$).

Practical regimes. (i) *Under-parameterized, near saturation:* With $\gamma = 1$, $\beta'_{\text{eff}}(N) \approx 0$ and slowly varying $\tilde{P}(N)$, equation 25 gives an accurate allocation rule: allocate training until the marginal N-gain $\propto \tilde{N}^{-2}$ matches the marginal k-gain $\propto (C - \tilde{N})^{-(\beta_{\text{eff}}+2)}$. (ii) *Finite-$N$, sub-asymptotic k:* When $\beta_{\text{eff}}(N)$ is still increasing, the $-\beta'_{\text{eff}}(N) \log(\cdot)$ term in equation 24 makes additional training strictly more valuable than the quasi-static approximation predicts; optimal policies invest more in N until the effective slope stabilizes. (iii) *Boundary optima:* If the right-hand side of equation 24 is always larger (resp. smaller) than the left-hand side over $\tilde{N} \in (0, C)$, the optimum collapses to a boundary solution $\tilde{N}_* \in \{0, C\}$ (all inference or all training), which can be checked by the sign of $d\mathcal{L}_{\text{tot}}/d\tilde{N}$ at the endpoints. This is shown explicitly in Fig. 2.

5 Test Case: LID in CIFAR-10H with Pre-trained Features

To bridge the LID model with a realistic setting, we evaluate on CIFAR-10 (Krizhevsky, 2012) paired with human label distributions from CIFAR-10H (Peterson et al., 2019). We *freeze* a pretrained ResNet-18 (He et al., 2015) and *fine-tune* only a linear head on top of the backbone features, matching the fine-tuning focus of our analysis.

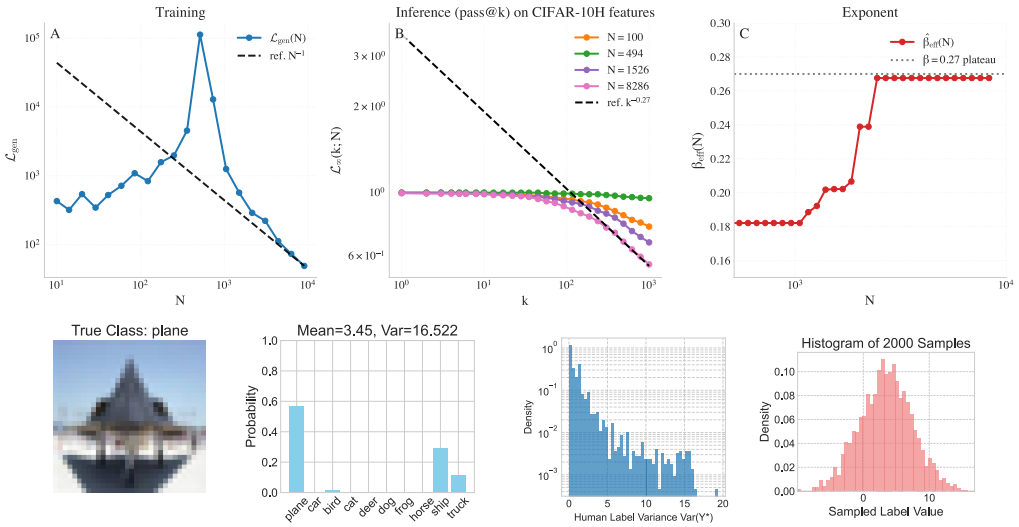

Figure 3: **Training and inference scaling on CIFAR-10H (frozen backbone, linear head). Top row:** *Left:* Generalization loss $\mathcal{L}_{\text{gen}}(N)$ with the N^{-1} tail in the classical regime. *Center:* Pass@k failure $\mathcal{L}_{\text{inf}}(k; N)$ for several N on CIFAR-10H; dashed shows a $k^{-\beta}$ reference anchored on the largest-N curve. *Right:* Effective inference slope $\beta_{\text{eff}}(N)$ estimated from the local log–log slope; the saturating fit $\beta_{\text{eff}}(N) = \beta - \Delta/(1 + c_\beta N^\nu)$ is overlaid, approaching the intrinsic tail index β. **Bottom row:** A CIFAR-10 image, its human label distribution, and the Gaussian PDF used to sample training/inference labels, illustrating instance difficulty via label variance.

5.1 Experimental setup

Feature extraction. We pass each image through a ResNet-18 pretrained on ImageNet and extract the penultimate-layer vector $\mathbf{z}_i \in \mathbb{R}^d$ with $d = 512$. These frozen features play the role of $\mathbf{x}$ in our LID analysis, and we fit only the linear head.

Stochastic Targets from Human Labels. The CIFAR-10H dataset provides, for each image, the distribution of labels assigned by multiple human annotators. For each image $\mathbf{x}_i$ (mapped to feature $\mathbf{z}_i$), we calculate the mean $m_i = \sum_{c=0}^{9} c \cdot P(\text{label} = c|\mathbf{x}_i)$ and variance $v_i = \sum_{c=0}^{9}(c - m_i)^2 \cdot P(\text{label} = c|\mathbf{x}_i)$ of these human label distributions. The key connection to our LID model is made by treating the target for our linear regressor as intrinsically stochastic. In order to control the signal to noise ratio, we introduce a scaling factor s to account for the magnitude of the noise variance, which is controllable in the LID setting. The learning problem is then defined on rescaled quantities

$$\text{Rescaled features: } \mathbf{z}_i^{\text{scaled}} = \mathbf{z}_i/s, \qquad \text{Rescaled mean target: } m_i^{\text{scaled}} = m_i^{\text{orig}}/s. \tag{26}$$

During training, for each rescaled feature vector $\mathbf{z}_i^{\text{scaled}}$, the target label y_i is sampled from a Gaussian distribution centered at its rescaled mean, using the human variance

$$y_i \sim \mathcal{N}(\text{mean} = m_i^{\text{scaled}}, \text{variance} = v_i + \epsilon_v). \tag{27}$$

where m_i^{scaled} is the rescaled mean human label for the instance, v_i is the variance of the human labels, and ϵ_v is a small constant (e.g., 10^{-6}) to ensure non-zero variance. Here, v_i plays a role analogous to $1/\tau_{\mathbf{x}_i}$ in our synthetic LID model, with high human label variance corresponding to a "difficult" instance (low effective precision).

Model and Training. We perform fine tuning by training a linear regression model $\hat{y} = \mathbf{z}^{\text{scaled},T}\hat{\boldsymbol{\theta}}$ to predict a scalar value from the rescaled features $\mathbf{z}^{\text{scaled}}$ extracted from the pretrained ResNet. Given the quadratic loss and linear model, we compute the optimal weights $\hat{\boldsymbol{\theta}}$ analytically using the Ridge regression solution with an effectively scaled regularization parameter $\lambda_{\text{eff}} = \lambda/s^2$

$$\hat{\boldsymbol{\theta}}_\lambda = (N^{-1}\mathbf{Z}_{\text{scaled}}^T\mathbf{Z}_{\text{scaled}} + \lambda_{\text{eff}}\mathbf{I})^{-1}N^{-1}\mathbf{Z}_{\text{scaled}}^T\mathbf{y}, \tag{28}$$

where $\mathbf{Z}_{\text{scaled}}$ is the matrix of rescaled training features, and $\mathbf{y}$ is the vector of sampled noisy labels (from Eq. (27)).

Evaluation metrics. *Generalization loss* $\mathcal{L}_{gen}$. On a held-out set we compute the MSE between $\hat{y}_{\text{val}} = \mathbf{z}_{\text{val}}^{\text{scaled}\,T}\hat{\boldsymbol{\theta}}$ and the rescaled human mean $m_{\text{val}}^{\text{scaled}}$. (To compare across different s, multiply by s^2.) *Inference failure* $\mathcal{L}_{inf}(k; N)$. For each validation point we draw k i.i.d. realizations $y_j \sim \mathcal{N}(m_{\text{val}}^{\text{scaled}}, v_{\text{val}} + \epsilon_v)$ and declare success if $|\hat{y}_{\text{val}} - y_j| \leq \delta_{\text{eff}}$ for any j, with $\delta_{\text{eff}} = \delta/s$. We average the all-fail indicator over the set. This implements the perfect-verification assumption from Sec. 4.2.

5.2 Results and connection to LID theory

Fig. 3 (top) shows $\mathcal{L}_{\text{gen}}$ vs. N and $\mathcal{L}_{\text{inf}}(\cdot; N)$ vs. k (here, $s = 32$ and $\delta = 0.05$). The generalization curve exhibits the familiar transition near $N \approx d$ and the $1/N$ decay for $N \gg d$, consistent with our linear-ridge analysis. For inference, each fixed-N curve follows an approximate power law $k^{-\beta_{\text{eff}}(N)}$; we estimate $\beta_{\text{eff}}(N)$ as the local log–log slope in a central k window. As predicted by Sec. 4.2, $\beta_{\text{eff}}(N)$ *increases with* N and *saturates*, reflecting the crossover from a bias-limited window (finite-N mean error) to the intrinsic LID-dominated tail.

The bottom row illustrates how human disagreement induces instance difficulty: broad label histograms (large v_i) produce wider Gaussians in equation 27, making pass@k success rarer at small k. This matches the mechanism behind the $k^{-\beta}$ law in the synthetic LID model (Sec. 4.2), with the caveat that the empirical difficulty distribution is not exactly Gamma; nevertheless, the slope behavior is robust and governed by the prevalence of high-variance instances.

Takeaways. (i) The linear-head fine-tuning setting reproduces the $\mathcal{L}_{\text{gen}}$ scaling predicted by the ridge analysis. (ii) The pass@k failure curves exhibit log–log linearity with a slope $\beta_{\text{eff}}(N)$ that grows with training data and plateaus, aligning with the finite-N theory and supporting the compute-allocation results in Sec. 4.3. (iii) Treating human uncertainty as instance-dependent noise provides a realistic testbed for LID: the qualitative scaling and the training-dependent inference exponent persist despite deviations from the idealized Gamma prior or independence assumptions between $\tau_{\mathbf{x}}$ and $\mathbf{x}$.

6 Discussion and Conclusion

This work introduced the Latent Instance Difficulty (LID) model, a simple, solvable framework for last-layer fine-tuning that unifies the scaling laws of training and inference. We modeled tasks with intrinsic, instance-heterogeneous difficulty and showed that while the generalization loss, $\mathcal{L}_{\text{gen}}$, follows established scaling with sample size N and data spectrum α, its improvement has a direct and non-trivial impact on inference performance.

Our central contribution is the derivation of a **training-dependent inference exponent**. The pass@k failure rate, $\mathcal{L}_{\text{inf}}(k)$, decays as a power law, but its exponent, $\beta_{\text{eff}}(N)$, is not fixed. It begins small for poorly-trained models and grows with the number of training samples N, eventually saturating at an intrinsic limit, β, determined by the tail of the task's true difficulty distribution. This mechanism reveals how reducing the model's error relative to the mean target makes inference-time compute more effective, up to a point of diminishing returns set by the data's irreducible stochasticity.

This unified view yields actionable insights. It predicts a clear crossover in the optimal resource allocation strategy: when the model is undertrained and $\beta_{\text{eff}}(N) < \beta$, the marginal benefit of acquiring more training data is high. Once the model is well-trained and the inference exponent has saturated, further gains are best sought by investing in more inference-time compute.

The LID model, while simple, provides a valuable theoretical baseline. It cleanly separates the roles of data structure (α) and data heterogeneity (β) while also explaining how they are coupled through the training process. By providing a closed-form, testable theory for when and how much test-time compute should help, it offers a first step toward a more principled understanding of resource allocation in modern machine learning.

Limitations. Our analysis focused on a tractable linear model to derive clear, analytical insights. This necessary simplification comes with limitations. First, while our fine-tuning frame makes the linear model highly relevant, a full theoretical extension to multi-layered, non-linear architectures would be the next step. A potential avenue for this is via random features or kernel models. Second, we assumed that the intrinsic difficulty distribution is a fixed property of the task, independent of the model's architecture. For complex reasoning, a more powerful model might not only learn the mean better but also fundamentally simplify the problem, an effect our current model does not capture. Finally, our work addresses single-output regression, and extending these ideas to structured, auto-regressive outputs, as seen in large language models, is an important challenge for future research.

REFERENCES

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales, 2019. URL https://arxiv.org/abs/1909.12673.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27), June 2024. ISSN 1091-6490. doi: 10.1073/pnas.2311878121. URL http://dx.doi.org/10.1073/pnas.2311878121.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws, 2024. URL https://arxiv.org/abs/2402.01092.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, April 2020. ISSN 1091-6490. doi: 10.1073/pnas.1907378117. URL http://dx.doi.org/10.1073/pnas.1907378117.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation, 2020. URL https://arxiv.org/abs/1903.08560.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q. Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling?, 2022. URL https://arxiv.org/abs/2207.10551.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Noam Levi. A simple model of inference scaling laws, 2024. URL https://arxiv.org/abs/2410.16377.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)?, 2025. URL https://arxiv.org/abs/2502.17578.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks, 2017. URL https://arxiv.org/abs/1706.05394.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021. URL https://arxiv.org/abs/2103.14749.

Jordan Hoffmann and et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust, 2019. URL https://arxiv.org/abs/1908.07086.

Noam Levi and Yaron Oz. The universal statistical structure and scaling laws of chaos and turbulence, 2023. URL https://arxiv.org/abs/2311.01358.

Noam Levi and Yaron Oz. The underlying scaling laws and universal statistical structure of complex datasets, 2024. URL https://arxiv.org/abs/2306.14975.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, July 2019. ISSN 1091-6490. doi: 10.1073/pnas.1903070116. URL http://dx.doi.org/10.1073/pnas.1903070116.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning, 2018. URL https://arxiv.org/abs/1802.01396.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve, 2020. URL https://arxiv.org/abs/1908.05355.

Omry Cohen, Or Malka, and Zohar Ringel. Learning curves for overparametrized deep neural networks: A field theory perspective. *Physical Review Research*, 3(2), April 2021. ISSN 2643-1564. doi: 10.1103/physrevresearch.3.023034. URL http://dx.doi.org/10.1103/PhysRevResearch.3.023034.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression, 2020. URL https://arxiv.org/abs/2006.05800.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020. URL https://arxiv.org/abs/2010.14701.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022. URL http://jmlr.org/papers/v23/20-1111.html.

S Spigler, M Geiger, S d'Ascoli, L Sagun, G Biroli, and M Wyart. A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, October 2019. ISSN 1751-8121. doi: 10.1088/1751-8121/ab4c8b. URL http://dx.doi.org/10.1088/1751-8121/ab4c8b.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws, 2023. URL https://arxiv.org/abs/2210.14891.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *arXiv preprint arXiv:2206.14486*, 2022.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers, 2025. URL https://arxiv.org/abs/2505.01618.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer, 2021. URL https://arxiv.org/abs/2102.01293.

Haowei Lin, Baizhou Huang, Haotian Ye, Qinyu Chen, Zihao Wang, Sujian Li, Jianzhu Ma, Xiaojun Wan, James Zou, and Yitao Liang. Selecting large language model to fine-tune via rectified scaling law, 2024. URL https://arxiv.org/abs/2402.02314.

Robert Irvine, Douglas Boubert, Vyas Raina, Adian Liusie, Ziyi Zhu, Vineet Mudupalli, Aliaksei Korshuk, Zongyi Liu, Fritz Cremer, Valentin Assassi, Christie-Carol Beauchamp, Xiaoding Lu, Thomas Rialan, and William Beauchamp. Rewarding chatbots for real-world engagement with millions of users, 2023. URL https://arxiv.org/abs/2303.06135.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/2501.19393.

Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. Advancing language model reasoning through reinforcement learning and inference scaling, 2025. URL https://arxiv.org/abs/2501.11651.

Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. Evolving deeper llm thinking, 2025. URL https://arxiv.org/abs/2501.09891.

Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D. Goodman. Stream of search (sos): Learning to search in language, 2024. URL https://arxiv.org/abs/2404.03683.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2024. URL https://arxiv.org/abs/2408.00724.

Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. Don't throw away your value model! generating more preferable text with value-guided monte-carlo tree search decoding, 2024. URL https://arxiv.org/abs/2309.15028.

Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B. Tenenbaum, and Chuang Gan. Planning with large language models for code generation, 2023. URL https://arxiv.org/abs/2303.05510.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models, 2024. URL https://arxiv.org/abs/2310.04406.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection, 2023. URL https://arxiv.org/abs/2310.09044.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. Self-evaluation guided beam search for reasoning, 2023. URL https://arxiv.org/abs/2305.00633.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2024a. URL https://arxiv.org/abs/2312.08935.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024b. URL https://arxiv.org/abs/2406.08673.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data, 2024. URL https://arxiv.org/abs/2405.14333.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D. Chang, and Prithviraj Ammanabrolu. Critique-out-loud reward models, 2024. URL https://arxiv.org/abs/2408.11791.

Jack W Silverstein and Z. D. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–199, 1995.

Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022. doi: 10.1017/9781009128490.

A Related Work

Generalization Scaling Laws Neural scaling laws (NSL) have been shown to accurately describe how the generalization loss of deep neural networks improves with scale (model/dataset size, compute) (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann and et al., 2022). Language model cross-entropy loss, for instance, often exhibits power-law scaling over orders of magnitude (Kaplan et al., 2020; Henighan et al., 2020), suggesting quantifiable performance gains with increased resources. Theoretical frameworks aim to explain these observations by identifying distinct scaling regimes (e.g., variance-limited, resolution-limited) (Bahri et al., 2021; Sharma and Kaplan, 2022; Spigler et al., 2019), sometimes linking them to data manifold intrinsic dimension (Sharma and Kaplan, 2022; Bahri et al., 2021). More nuanced models address dynamical scaling evolution with training time (Bordelon et al., 2024) and "broken" neural scaling laws (BNSL) that capture transitions between power-law regimes (Caballero et al., 2023; Nakkiran et al., 2021). Such breaks highlight complexities in extrapolation and suggest plateaus might stem from suboptimal strategies rather than fundamental limits (Sorscher et al., 2022; Dey et al., 2025). Beyond pre-training, scaling laws also govern fine-tuning performance relative to pre-trained model size and fine-tuning data volume (Hernandez et al., 2021; Lin et al., 2024). These studies show pre-training effectively augments fine-tuning data, with transfer benefits following predictable, though task-dependent, patterns (Hernandez et al., 2021), shifting focus towards leveraging pre-training for diverse downstream applications. In this work, we mainly focus on the vanilla NSL setting where the scaling law pertains to generalization loss improvement with increased number of training samples.

Inference Time Scaling Methods for scaling inference-time compute in deep learning often involve generating multiple solution candidates and then selecting the best one based on specific criteria. These criteria include choosing the most frequent response for majority voting or the best response based on an external reward for Best-of-N (Brown et al., 2024; Irvine et al., 2023; Levi, 2024; Muennighoff et al., 2025; Schaeffer et al., 2025). Unlike repeated sampling, previous sequential scaling methods let the model generate solution attempts sequentially, building upon previous attempts and allowing it to refine each attempt based on prior outcomes (Snell et al., 2024; Hou et al., 2025; Lee et al., 2025). Tree-based search methods (Gandhi et al., 2024; Wu et al., 2024) offer a hybrid approach between sequential and parallel scaling. Examples include Monte-Carlo Tree Search (MCTS) (Liu et al., 2024; Zhang et al., 2023; Zhou et al., 2024; Choi et al., 2023) and guided beam search (Xie et al., 2023). REBASE (Wu et al., 2024) employs a process reward model to balance exploitation and pruning during its tree search. Empirically, REBASE has been shown to outperform sampling-based methods and MCTS (Wu et al., 2024). Reward models play a key role in these inference-time scaling methods (Lightman et al., 2023; Wang et al., 2024a;b; Wu et al., 2024; Gandhi et al., 2024; Liu et al., 2024; Zhang et al., 2023; Zhou et al., 2024; Choi et al., 2023; Xie et al., 2023; Xin et al., 2024; Ankner et al., 2024). They generally come in two variants: outcome reward models and process reward models. Outcome reward models (Xin et al., 2024; Ankner et al., 2024) assign a score to complete solutions and are particularly useful in Best-of-N selection. In contrast, process reward models (Lightman et al., 2023; Wang et al., 2024a; Wu et al., 2024) assess individual reasoning steps and are effective in guiding tree-based search methods. Other approaches also explore simple test-time scaling techniques (Muennighoff et al., 2025).

B High Dimensional Ridge Regression with Noisy Labels

In this appendix, we re-derive and analyze the generalization error for high-dimensional Ridge linear regression with additive label noise used in the main text. While these results are well-established in the literature (see, e.g., (Maloney et al., 2022; Hastie et al., 2020) and references therein), we present a concise derivation to highlight the connection to our Latent Instance Difficulty (LID) model and to set the stage for understanding the training scaling laws discussed in Section 4.1.

B.1 Model Setup

We consider a linear model where the learner aims to estimate a true underlying parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ from N training samples $(\mathbf{x}_i, y_i)$. The input features $\mathbf{x}_i \in \mathbb{R}^d$ are drawn IID from a distribution with zero mean and covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$. The observed labels y_i are generated according to

$$y_i = \mathbf{x}_i^T \boldsymbol{\theta}^* + \eta_i, \tag{29}$$

where η_i is the label noise for sample i. In the context of our LID model (Definition 3.1), η_i is a realization from $\mathcal{N}(0, \sigma_\eta^2/\tau_{\mathbf{x}_i})$. For this general derivation, we assume η_i are IID with $\mathbb{E}[\eta_i] = 0$ and $\mathbb{E}[\eta_i^2] = \sigma_{\text{noise}}^2$. If we were directly applying LID, σ_{noise}^2 would be replaced by $\mathbb{E}[\sigma_\eta^2/\tau_{\mathbf{x}}] = \sigma_\eta^2 \mathbb{E}[1/\tau_{\mathbf{x}}]$. The learner estimates $\boldsymbol{\theta}^*$ using Ridge regression by minimizing the loss

$$L(\hat{\boldsymbol{\theta}}) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\theta}})^2 + \frac{\lambda}{2} \|\hat{\boldsymbol{\theta}}\|_2^2, \tag{30}$$

where $\lambda \geq 0$ is the regularization parameter. Let $X \in \mathbb{R}^{N \times d}$ be the matrix of training features (rows $\mathbf{x}_i^T$) and $\mathbf{y} \in \mathbb{R}^N$ be the vector of observed labels. The solution is

$$\hat{\boldsymbol{\theta}}_\lambda = \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \frac{1}{N} X^T \mathbf{y}. \tag{31}$$

B.2 Generalization Error

The generalization error (or test loss) measures the expected squared prediction error on unseen test data $\mathbf{x}_{\text{test}}$ with true mean target $\mathbf{x}_{\text{test}}^T \boldsymbol{\theta}^*$

$$\mathcal{L}_{\text{gen}}(N, \lambda) = \mathbb{E}_{\mathcal{D}_N, \mathbf{x}_{\text{test}}} \left[(\mathbf{x}_{\text{test}}^T \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}_{\text{test}}^T \boldsymbol{\theta}^*)^2 \right]. \tag{32}$$

Let $\Delta\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*$. Then $\mathcal{L}_{\text{gen}} = \mathbb{E}[\Delta\boldsymbol{\theta}^T \boldsymbol{\Sigma} \Delta\boldsymbol{\theta}]$, where the expectation is over the training data $\mathcal{D}_N$. Substituting $\mathbf{y} = X\boldsymbol{\theta}^* + \vec{\eta}$ (where $\vec{\eta}$ is the vector of noise realizations η_i) into Eq. (31)

$$\hat{\boldsymbol{\theta}}_\lambda = \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \frac{1}{N} X^T (X\boldsymbol{\theta}^* + \vec{\eta}) = \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \left(\frac{1}{N} X^T X \boldsymbol{\theta}^* + \frac{1}{N} X^T \vec{\eta} \right).$$

So the error vector $\Delta\boldsymbol{\theta}$ is

$$\begin{aligned}
\Delta\boldsymbol{\theta} &= \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \\
&= \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \left(\frac{1}{N} X^T X \boldsymbol{\theta}^* + \frac{1}{N} X^T \vec{\eta} \right) - \boldsymbol{\theta}^* \\
&= \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \left[\frac{1}{N} X^T X \boldsymbol{\theta}^* + \frac{1}{N} X^T \vec{\eta} - \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right) \boldsymbol{\theta}^* \right] \\
&= \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \left[\frac{1}{N} X^T \vec{\eta} - \lambda \boldsymbol{\theta}^* \right].
\end{aligned} \tag{33}$$

The generalization error can be decomposed into bias and variance terms. Assuming $\boldsymbol{\theta}^*$ is fixed (not random) and $\mathbb{E}[\vec{\eta}] = \mathbf{0}$, $\mathbb{E}[\vec{\eta}\vec{\eta}^T] = \sigma_{\text{noise}}^2 \mathbf{I}_N$ (for IID noise with average variance $\sigma_{\text{noise}}^2 = \mathbb{E}[\eta_i^2]$)

$$\mathbb{E}[\Delta\boldsymbol{\theta}] = \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} (-\lambda \boldsymbol{\theta}^*). \tag{34}$$

This is the bias of the estimator $\hat{\boldsymbol{\theta}}_\lambda$. The covariance of $\Delta\boldsymbol{\theta}$ (which is also $\text{Cov}(\hat{\boldsymbol{\theta}}_\lambda)$) is

$$\begin{aligned}
\text{Cov}(\Delta\boldsymbol{\theta}) &= \mathbb{E}[(\Delta\boldsymbol{\theta} - \mathbb{E}[\Delta\boldsymbol{\theta}])(\Delta\boldsymbol{\theta} - \mathbb{E}[\Delta\boldsymbol{\theta}])^T] \\
&= \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \mathbb{E} \left[\left(\frac{1}{N} X^T \vec{\eta} \right) \left(\frac{1}{N} X^T \vec{\eta} \right)^T \right] \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \\
&= \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \frac{1}{N^2} X^T \mathbb{E}[\vec{\eta}\vec{\eta}^T] X \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \\
&= \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1} \frac{\sigma_{\text{noise}}^2}{N^2} X^T X \left(\frac{1}{N} X^T X + \lambda \mathbf{I}_d \right)^{-1}.
\end{aligned} \tag{35}$$

The generalization error is $\mathcal{L}_{\text{gen}} = \mathbb{E}[\text{Tr}(\boldsymbol{\Sigma}\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T)] = \text{Tr}(\boldsymbol{\Sigma}\mathbb{E}[\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T])$. $\mathbb{E}[\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T] = \text{Cov}(\Delta\boldsymbol{\theta}) + \mathbb{E}[\Delta\boldsymbol{\theta}]\mathbb{E}[\Delta\boldsymbol{\theta}]^T$.

Hence, $\mathcal{L}_{\text{gen}} = \text{Bias}^2 + \text{Variance}$

$$\mathcal{L}_{\text{gen}} = \mathbb{E}[\Delta\boldsymbol{\theta}]^T \boldsymbol{\Sigma} \mathbb{E}[\Delta\boldsymbol{\theta}] + \text{Tr}\left(\boldsymbol{\Sigma}\text{Cov}(\Delta\boldsymbol{\theta})\right) = \lambda^2 \text{Tr}\left[\boldsymbol{\theta}^*(\boldsymbol{\theta}^*)^T \left(\hat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\Sigma} \left(\hat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}\right)^{-1}\right] + \text{Tr}\left(\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}\right)^{-1} \sigma^2_{\text{noise}} \hat{\boldsymbol{\Sigma}} \left(\hat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}\right)^{-1}\right), \tag{36}$$

where $\hat{\boldsymbol{\Sigma}} = \frac{1}{N} X^T X$ is the empirical covariance.

B.2.1 Underparameterized Regime ($N \gg d$)

In the underparameterized limit, $N \gg d$, we can safely take the Ridge parameter to 0, obtaining a simple result

$$\mathcal{L}_{\text{gen}} = \sigma^2_{\text{noise}} \text{Tr}\left(\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Sigma}}^{-1}\right) = \sigma^2_{\text{noise}} \text{Tr}\left(\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\right). \tag{37}$$

In the setting discussed in the main text, the population covariance can be written as a $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\mathbf{I}$, where $\boldsymbol{\Lambda}$ is diagonal with a decaying power law spectrum, as in Item 1. As stated in Silverstein and Bai (1995); Couillet and Liao (2022), the empirical covariance matrix is given by $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}\mathbf{W}$ where $\mathbf{W}$ is a Wishart matrix with parameter $\kappa = d/N$. This definition implies that the generalization loss is given by

$$\mathcal{L}_{\text{gen}} = \sigma^2_{\text{noise}} \text{Tr}\left(\mathbf{W}^{-1}\right) \approx \sigma^2_{\text{noise}} \frac{1}{N/d - 1} \propto \sigma^2_{\text{noise}} \frac{d}{N}, \tag{38}$$

where the key equality is due to taking an expectation over the eigenvalues of an Inverse Wishart matrix, which obey an inverse Marchenko-Pastur distribution distribution Couillet and Liao (2022).

Transition ($N \approx d$): Note that near the interpolation threshold, $\mathcal{L}_{\text{gen}}$ in Eq. (38) exhibits a peak which is one side of "double descent" Belkin et al. (2019).

B.2.2 Overparameterized Regime ($N \ll d$)

In the limit of ridgeless regression ($\lambda \to 0$) for the overparamerized regime, the solution interpolates the training data. The error is mainly determined by the implicit bias of the minimum L_2-norm solution, with the noise contribution appearing only near the interpolation threshold. In this case, the matrix $\hat{\boldsymbol{\Sigma}}$ is rank deficient, with N nonzero eigenvalues and $d - N$ null eigenvalues. This requires a more careful consideration of the ridgeless limit, namely, one must introduce scaling to the ridge parameter, for instance as $\lambda = \tilde{\lambda}/N$ (Bahri et al., 2021). Under this scaling, the test loss becomes

$$\mathcal{L}_{\text{gen}} = \frac{\tilde{\lambda}^2}{N^2 d} \text{Tr}\left[\left(\hat{\boldsymbol{\Sigma}} + \frac{\tilde{\lambda}}{N}\mathbf{I}\right)^{-2} \boldsymbol{\Sigma}\right] + \text{Tr}\left(\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}} + \frac{\tilde{\lambda}}{N}\mathbf{I}\right)^{-1} \sigma^2_{\text{noise}} \hat{\boldsymbol{\Sigma}} \left(\hat{\boldsymbol{\Sigma}} + \frac{\tilde{\lambda}}{N}\mathbf{I}\right)^{-1}\right). \tag{39}$$

The first term can be exactly derived using replica methods, as in Bordelon et al. (2020), but can also be evaluated asymptotically by solving the self consistent equations

$$\text{Bias}^2 \approx \frac{\tilde{\lambda}^2}{N^2 d} \sum_{i=1}^{N} \frac{\sigma_i^2}{(\sigma_i^2 + \frac{\tilde{\lambda}}{N})^2}, \qquad \tilde{\lambda} = \sum_{i=1}^{N} \frac{\frac{\tilde{\lambda}}{N}\sigma_i^2}{(\sigma_i^2 + \frac{\tilde{\lambda}}{N})}. \tag{40}$$

Solving Eq. (40) leads to the power law scaling $\mathcal{L}_{\text{gen}} \propto N^{-\alpha}$ (Bartlett et al., 2020; Hastie et al., 2020). The noise term in the overparameterized regime can be resolved with the same arguments as in the underparameterized case, with the replacement $\{d, N\} \to \{N, d\}$, examplifying double descent (Maloney et al., 2022).

In the context of the LID model, σ^2_{noise} is replaced by $\sigma_\eta^2 \mathbb{E}[1/\tau_{\mathbf{x}}]$. Thus, Assumption 3.2 ($\beta > 2$) is crucial for these scalings to hold with finite prefactors. If $\beta \leq 2$, $\mathbb{E}[1/\tau_{\mathbf{x}}]$ diverges, and standard Ridge regression analysis is compromised.

B.3 Inference Scaling at Fixed N

We provide the detailed steps for the asymptotic evaluation of the pass@k failure probability integral given in Eq. (13) for large k

$$\mathcal{L}_{\text{inf}}(k) = \mathbb{E}_{\tau_{\mathbf{x}} \sim \text{Gamma}(\beta/2,1)} \left[[p(\mathbf{x}, \tau_{\mathbf{x}})]^k \right] \approx \mathbb{E}_{\tau_{\mathbf{x}} \sim \text{Gamma}(\beta/2,1)} [e^{-kc_\delta \sqrt{\tau_{\mathbf{x}}}}]. \tag{41}$$

Here, $c_\delta = 2\delta/(\sqrt{2\pi}\sigma_\eta)$ and $p(\mathbf{x}, \tau_{\mathbf{x}}) \approx 1 - c_\delta \sqrt{\tau_{\mathbf{x}}}$ for small $\tau_{\mathbf{x}}$. The PDF of $\tau_{\mathbf{x}} \sim \text{Gamma}(\beta/2, 1)$ is $f(\tau) = \frac{1}{\Gamma(\beta/2)} \tau^{\beta/2-1} e^{-\tau}$. The expectation integral is

$$\mathcal{L}_{\text{inf}}(k) = \int_0^\infty e^{-kc_\delta \sqrt{\tau}} \frac{1}{\Gamma(\beta/2)} \tau^{\beta/2-1} e^{-\tau} d\tau. \tag{42}$$

For large k, the factor $e^{-kc_\delta \sqrt{\tau}}$ decays extremely rapidly for any $\tau > 0$, forcing the dominant contribution to the integral to come from the region near $\tau = 0$. In this region, the term $e^{-\tau}$ in the Gamma PDF is approximately 1. Thus, we approximate the integral as

$$\mathcal{L}_{\text{inf}}(k) \approx \frac{1}{\Gamma(\beta/2)} \int_0^\infty e^{-kc_\delta \sqrt{\tau}} \tau^{\beta/2-1} d\tau. \tag{43}$$

We perform a change of variable: Let $u = kc_\delta \sqrt{\tau}$. Then $\sqrt{\tau} = u/(kc_\delta)$, which implies $\tau = (u/(kc_\delta))^2 = u^2/(kc_\delta)^2$. The differential is $d\tau = \frac{2u}{(kc_\delta)^2} du$. Substituting these into the integral equation 43

$$\begin{aligned}
\int_0^\infty e^{-u} \left(\frac{u^2}{(kc_\delta)^2} \right)^{\beta/2-1} \frac{2u}{(kc_\delta)^2} du &= \int_0^\infty e^{-u} \frac{u^{\beta-2}}{(kc_\delta)^{\beta-2}} \frac{2u}{(kc_\delta)^2} du \\
&= \frac{2}{(kc_\delta)^{\beta-2}(kc_\delta)^2} \int_0^\infty e^{-u} u^{\beta-1} du \\
&= \frac{2}{(kc_\delta)^\beta} \int_0^\infty u^{\beta-1} e^{-u} du.
\end{aligned}$$

The remaining integral is the definition of the Gamma function, $\Gamma(\beta)$, which converges for $\beta > 0$.

$$\int_0^\infty u^{\beta-1} e^{-u} du = \Gamma(\beta). \tag{44}$$

Substituting this back into the expression for $\mathcal{L}_{\text{inf}}(k)$

$$\mathcal{L}_{\text{inf}}(k) \approx \frac{1}{\Gamma(\beta/2)} \frac{2\Gamma(\beta)}{(kc_\delta)^\beta} = \left(\frac{2\Gamma(\beta)}{\Gamma(\beta/2)(c_\delta)^\beta} \right) k^{-\beta}. \tag{45}$$

This confirms the asymptotic scaling $\mathcal{L}_{\text{inf}}(k) \sim k^{-\beta}$ for large k.

Assumptions for the Tauberian step We use standard Laplace–Stieltjes Tauberian arguments under the following mild conditions:

1. **Regularly varying difficulty near zero.** The latent precision satisfies $\Pr(\tau_x \le t) = t^{\beta/2} L(t)$ as $t \downarrow 0$, with L slowly varying and bounded on $(0, t_0]$.
2. **Uniform small-window expansion.** For some $c_\delta = \sqrt{2/\pi}\, \delta/\sigma_\eta$ and any compact $[\tau_{\text{lo}}, \tau_{\text{hi}}] \subset (0, \infty)$,
$$s(B, \tau) := \Pr\left(|B - \eta| \le \delta \mid \tau \right) = c_\delta \sqrt{\tau}\, \exp\Big(- \tfrac{B^2 \tau}{2\sigma_\eta^2} \Big) (1 + o(1))$$
uniformly in $\tau \in [\tau_{\text{lo}}, \tau_{\text{hi}}]$ as $\delta \downarrow 0$.
3. **Model error.** $B_N(x) = x^\top(\hat{\theta}_\lambda - \theta)$ is centered sub-Gaussian with $\text{Var}[B_N] \asymp \mathcal{L}_{\text{gen}}(N)$ and is independent of τ_x (or weakly dependent so that conditioning on τ_x preserves sub-Gaussian tails).
4. **Independent trials and perfect verification.** Conditional on (x, τ_x) and the training set, the k comparisons are i.i.d., and success is declared if any trial lies within the tolerance.

Under (1)–(4), Karamata-type Tauberian theorems yield the mixture law in equation 17 and the $k^{-\beta}$ tail in equation 15, with constants as stated.

B.4 EQUIVALENCE TO MODEL-DRAW PASS@k.

Consider the alternative protocol that draws k i.i.d. model proposals $\tilde{y}_j = \hat{y} + \xi_j$ with proposal noise ξ having a bounded, smooth density f_ξ independent of τ_x, and verifies against a fixed target $y^\star$. Writing $B_N(x) = \hat{y} - m(x)$, the per-trial success probability is

$$\int_{|u|\le\delta} f_\xi\big(B_N(x) - u\big)\, du \;=\; 2\delta\, f_\xi\big(B_N(x)\big)\,(1 + o(1)) \quad (\delta \downarrow 0). \tag{46}$$

In our target-draw protocol the corresponding quantity is $2\delta\, f_\eta\big(B_N(x); \tau_x\big)\,(1 + o(1))$ with $f_\eta(\cdot; \tau_x) = \mathcal{N}(0, \sigma_\eta^2/\tau_x)$. Thus both protocols reduce to $(1-p)^k$ with $p \propto 2\delta$ times a *local density at* $B_N(x)$; since the only heavy tail in LID comes from $f_\eta(\cdot; \tau_x) \propto \sqrt{\tau_x}$ as $\tau_x \downarrow 0$, the small-success behavior, and hence the $k^{-\beta}$ tail, is unchanged by swapping model vs. target sampling (up to prefactors). This equivalence holds provided f_ξ is bounded and does not itself introduce a τ_x-dependent tail, and trials are conditionally independent.

C TRAINING-DEPENDENT INFERENCE SCALING IN THE LID LINEAR MODEL

Setup. Let $\mathbf{x} \in \mathbb{R}^d$ be features with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and linear teacher $m(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}^*$. Each instance has latent precision $\tau_\mathbf{x} \sim \Gamma(\beta/2, 1)$, independent of $\mathbf{x}$. A single realization of the stochastic target is

$$Y^*_\mathbf{x} \;\sim\; \mathcal{N}\big(m(\mathbf{x}),\; \sigma_\eta^2/\tau_\mathbf{x}\big). \tag{47}$$

Training observes i.i.d. pairs $(\mathbf{x}_i, y_i)$ with $y_i \sim Y^*_{\mathbf{x}_i}$ and fits ridge/OLS to obtain $\hat{\boldsymbol{\theta}}_\lambda(N)$ from N samples. At test time we evaluate (i) the *training loss* $\mathcal{L}_{\text{gen}}(N) = \mathbb{E}_\mathbf{x}[(\mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - \mathbf{x}^\top \boldsymbol{\theta}^*)^2]$, and (ii) the *inference loss* $\mathcal{L}_{\text{inf}}(k; N)$ under a perfect verifier with tolerance $\delta > 0$: for each test $\mathbf{x}$ we draw k i.i.d. $y_j \sim Y^*_\mathbf{x}$ and declare success if $\min_{j\le k} |\mathbf{x}^\top \hat{\boldsymbol{\theta}}_\lambda - y_j| \le \delta$. The quantity $\mathcal{L}_{\text{inf}}(k; N)$ is the population failure probability.

Verification symmetry. Throughout we evaluate pass@k under a perfect verifier by drawing k fresh realizations from Y_x and checking proximity to the deterministic head $x^\top \hat{\theta}_\lambda$. As argued in Sec. 4.2, in the small-window regime the single-try success is proportional to the local density of whichever operand is random near the gap $B_N(x)$. Replacing "draws from Y_x" with "draws from a model sampling distribution m independent of τ_x" therefore leaves the k–dependence and the τ_x-tail exponent unchanged (the prefactor may shift).

Asymptotic reference (bias-free). When the mean is learned well (e.g., $\mathcal{L}_{\text{gen}}(N) \to 0$), a standard Tauberian argument over the small-$\tau_\mathbf{x}$ tail yields

$$\mathcal{L}_{\text{inf}}(k; N) \;\sim\; C_1(\beta, \delta, \sigma_\eta)\, k^{-\beta}, \qquad k \to \infty, \tag{48}$$

with $C_1 = 2\,\Gamma(\beta)/\big[\Gamma(\beta/2)\, c_\delta^\beta\big]$ and $c_\delta = \sqrt{2/\pi}\,\delta/\sigma_\eta$, matching equation 15.

Assumption C.1 (Regular-variation and small-window conditions for Tauberian steps)**.** We assume:

1. **Near-zero tail of difficulty.** The latent precision has a regularly varying CDF $\Pr(\tau_x \le t) = t^{\beta/2} L(t)$ as $t \downarrow 0$, with L slowly varying and $\beta > 0$. (For $\Gamma(\beta/2, 1)$ we have $L(t) \to 1/\Gamma(\beta/2 + 1)$.)

2. **Small-window success form (uniform on compacts).** For tolerance $\delta > 0$,

$$s(B, \tau_x) := \Pr(|B - \eta| \le \delta \mid \tau_x) = c_\delta \sqrt{\tau_x}\, e^{-\frac{B^2 \tau_x}{2\sigma_\eta^2}}\,(1 + o(1)) \quad \text{as } \delta \downarrow 0, \tag{49}$$

   uniformly for τ_x in any fixed compact subset of $(0, \infty)$, with $c_\delta = \sqrt{2/\pi}\,\delta/\sigma_\eta$ and $\eta \sim \mathcal{N}(0, \sigma_\eta^2/\tau_x)$.

3. **Training-induced bias.** $B_N(x) := \mathcal{E}_{\text{gen}}(x)$ is sub-Gaussian with $\mathrm{Var}[B_N] = \Theta(\mathcal{L}_{\text{gen}}(N))$, and (B_N, τ_x) are independent (the latter simplifies exposition and affects only constants/crossover scales).

Finite-N mean error and per-trial success. Define the prediction bias on a fresh test point by

$$B_N(\mathbf{x}) \;:=\; \mathcal{E}_{\text{gen}}(\mathbf{x}) \;=\; \mathbf{x}^\top(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\,. \tag{50}$$

Under standard linear generalization (e.g., benign spectrum), $B_N(\mathbf{x})$ is approximately Gaussian with zero mean and variance $\mathrm{Var}[B_N] = \Theta(\mathcal{L}_{\text{gen}}(N))$; in the under-parameterized regime, $\mathcal{L}_{\text{gen}}(N) = \Theta(1/N)$. For fixed $(\mathbf{x}, \tau_{\mathbf{x}})$ and small δ, a CDF expansion of the Gaussian likelihood gives the small-window approximation that we used in the main text (cf. equation 16):

$$s\big(B_N(\mathbf{x}), \tau_{\mathbf{x}}\big) \;:=\; \mathbb{P}\big(|B_N(\mathbf{x}) - \eta| \le \delta \;\big|\; \tau_{\mathbf{x}}\big) \;\approx\; c_\delta\,\sqrt{\tau_{\mathbf{x}}}\,\exp\Big(-\frac{B_N(\mathbf{x})^2\,\tau_{\mathbf{x}}}{2\sigma_\eta^2}\Big), \tag{51}$$

where $\eta \sim \mathcal{N}\big(0, \sigma_\eta^2/\tau_{\mathbf{x}}\big)$.

Therefore, very small success probability arises either from $\tau_{\mathbf{x}} \downarrow 0$ (intrinsic difficulty) *or* from large $|B_N(\mathbf{x})|$ at typical $\tau_{\mathbf{x}} = \Theta(1)$.

Proposition C.2 (Two-tail law for single-try success probabilities)**.** *Let S_N denote the random single-trial success probability $S_N = s(B_N(\mathbf{x}), \tau_{\mathbf{x}})$ across test $\mathbf{x}$ and $\tau_{\mathbf{x}}$. Assume equation 51 and $B_N(\mathbf{x}) \sim \mathcal{N}(0, \mathrm{Var}[B_N])$ with $\mathrm{Var}[B_N] = \Theta(\mathcal{L}_{gen}(N))$. Then, as $s \downarrow 0$,*

$$\mathbb{P}(S_N \le s) \;=\; A\,s^\beta \;+\; B(N)\,s^{\gamma(N)}\,(1 + o(1)), \tag{52}$$

where $A = c_\delta^{-\beta}/\Gamma(\beta/2+1) > 0$, $B(N) > 0$, and

$$\gamma(N) \;=\; \Theta\Big(\frac{1}{\mathrm{Var}[B_N]}\Big) \;=\; \Theta\Big(\frac{1}{\mathcal{L}_{gen}(N)}\Big). \tag{53}$$

Sketch. Condition on $\tau_{\mathbf{x}}$ and use equation 51. For moderate $\tau_{\mathbf{x}} = \Theta(1)$, the event $S_N \le s$ corresponds to $|B_N(\mathbf{x})| \gtrsim \sqrt{(2\sigma_\eta^2/\tau_{\mathbf{x}})\log(c_\delta\sqrt{\tau_{\mathbf{x}}}/s)}$. Since $B_N(\mathbf{x})$ is (approximately) Gaussian with variance $\mathrm{Var}[B_N]$, Gaussian tail bounds imply $\mathbb{P}(S_N \le s \mid \tau_{\mathbf{x}}) \approx s^{\,\sigma_\eta^2/(\tau_{\mathbf{x}}\mathrm{Var}[B_N])}$ up to slowly varying factors; integrating $\tau_{\mathbf{x}}$ over the $\Gamma(\beta/2, 1)$ density concentrated at $\Theta(1)$ yields the $s^{\gamma(N)}$ contribution with $\gamma(N) = \Theta(1/\mathrm{Var}[B_N])$. Separately, for small $\tau_{\mathbf{x}}$, $s \approx c_\delta\sqrt{\tau_{\mathbf{x}}}$, so $\{S_N \le s\}$ contains $\{\tau_{\mathbf{x}} \le (s/c_\delta)^2\}$ and $\mathbb{P}(\tau_{\mathbf{x}} \le t) \sim t^{\beta/2}/\Gamma(\beta/2+1)$, giving $A\,s^\beta$. Summing the contributions gives the two-tail form. □

Corollary C.3 (Mixture law for pass@k)**.** *Let $\mathcal{L}_{inf}(k; N) = \mathbb{E}\big[(1 - S_N)^k\big]$. By Tauberian theory for Laplace–Stieltjes transforms of regularly varying tails,*

$$\mathcal{L}_{inf}(k; N) \;=\; C_1\,k^{-\beta} \;+\; C_2(N)\,k^{-\gamma(N)}\,(1 + o(1)) \qquad (k \to \infty), \tag{54}$$

with $C_1 = \frac{2\,\Gamma(\beta)}{\Gamma(\beta/2)}\,c_\delta^{-\beta}, c_\delta = \sqrt{\frac{2}{\pi}}\,\frac{\delta}{\sigma_\eta}$ (identical to the constant in equation 15) and $C_2(N) = B(N)\,\Gamma(\gamma(N)+1) > 0$.

Training-dependent effective exponent. For a fixed practical k-window $[k_1, k_2]$, define the local slope

$$\beta_{\text{eff}}(N; [k_1, k_2]) \;:=\; -\,\frac{d\log \mathcal{L}_{\text{inf}}(k; N)}{d\log k}\Big|_{k\in[k_1,k_2]}. \tag{55}$$

By Cor. C.3, the dominant term in the window dictates the slope; hence the summary law reported in the main text:

$$\beta_{\text{eff}}(N) \;\approx\; \min\big\{\beta,\; \gamma(N)\big\}, \qquad \gamma(N) = \Theta\Big(\frac{1}{\mathcal{L}_{\text{gen}}(N)}\Big). \tag{56}$$

Therefore $\beta_{\text{eff}}(N)$ is monotone in N and saturates at β; if $\mathcal{L}_{\text{gen}}(N) \asymp N^{-\nu_{\text{tr}}}$, then $\gamma(N) \asymp N^{\nu_{\text{tr}}}$, and a convenient saturating fit is

$$\beta_{\text{eff}}(N) \;\approx\; \beta \;-\; \frac{\Delta}{1 + c_\beta\,N^\nu} \qquad (\Delta, c_\beta, \nu > 0), \tag{57}$$

as used to summarize empirical curves (cf. center/right panels of Fig. 1).

Summary: finite-N correction and $\beta_{\text{eff}}(N)$

Single-trial success (small window). For tolerance $\delta > 0$,

$$1 - p(\mathbf{x}, \tau_{\mathbf{x}}) \;\approx\; c_\delta \sqrt{\tau_{\mathbf{x}}}\, \exp\Big(- \frac{\mathcal{E}_{\text{gen}}(\mathbf{x})^2\, \tau_{\mathbf{x}}}{2\sigma_\eta^2} \Big), \qquad c_\delta = \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\delta}{\sigma_\eta}. \tag{58}$$

Two-tail mixture. As $k \to \infty$,

$$\mathcal{L}_{\text{inf}}(k; N) = C_1\, k^{-\beta} + C_2(N)\, k^{-\gamma(N)}(1 + o(1)),\ \gamma(N) = \Theta\Big(\tfrac{1}{\mathcal{L}_{\text{gen}}(N)}\Big),\ C_1 = \frac{2\Gamma(\beta)}{\Gamma(\beta/2)}\, c_\delta^{-\beta}. \tag{59}$$

Effective slope. In any fixed k-window,

$$\beta_{\text{eff}}(N) \;\approx\; \min\{\beta,\ \gamma(N)\} \;\nearrow\; \beta \quad \text{as } N \uparrow, \tag{60}$$

with the empirical fit equation 57 used for plots and for the compute-allocation condition in Sec. 4.3.

D ADDITIONAL EXAMPLES FROM CIFAR-10H

Here, we provide some additional examples from the CIFAR-10H dataset, to illustrate the type of stochastic labels inherent in the data.

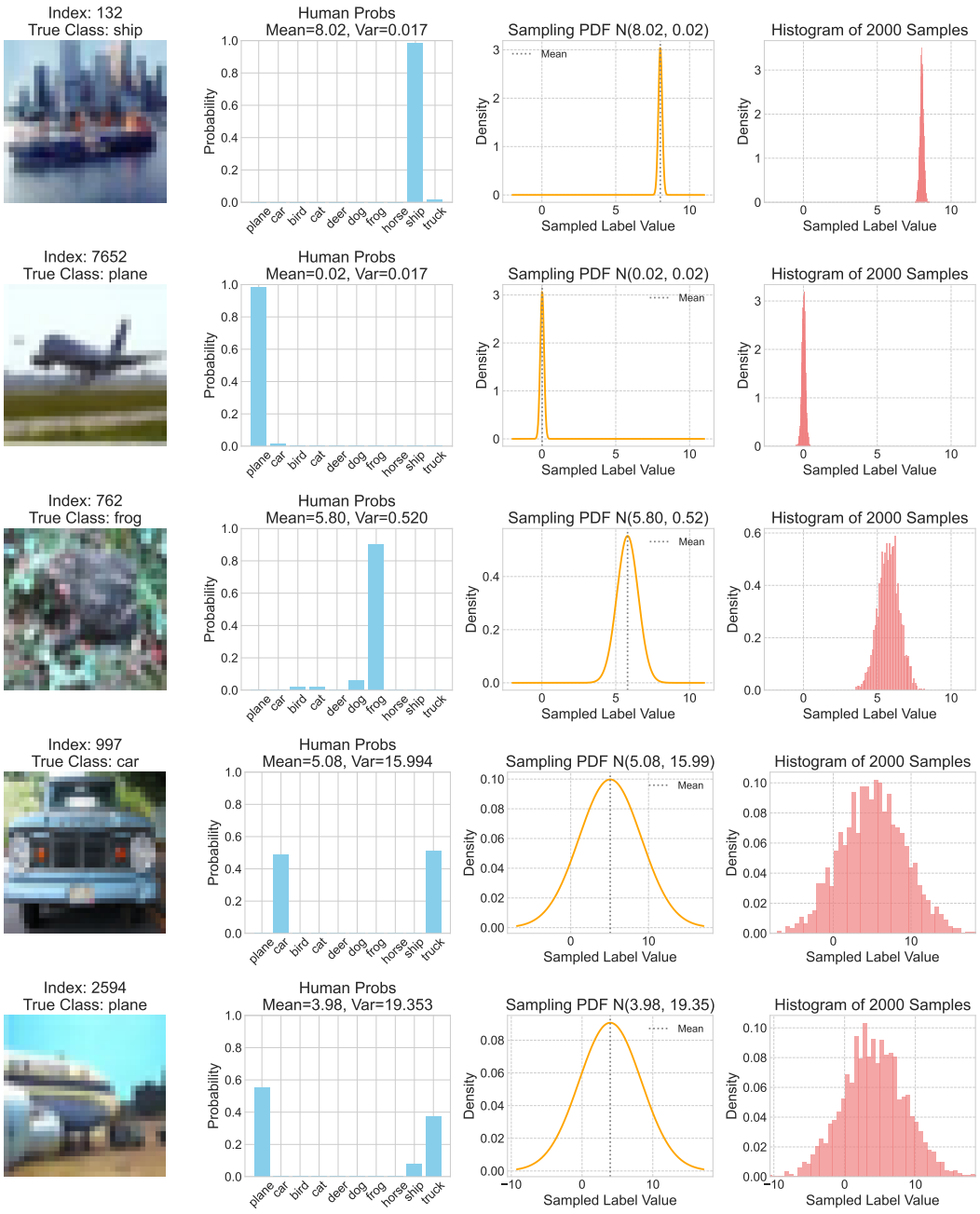

Figure 4: **Examples for low and high variance images from CIFAR-10H.** *Top to bottom:* low variance samples are easier to predict (more localized near the average prediction) while high variance samples are difficult.