Tokens-per-Parameter Coverage Is Critical for Robust LLM Scaling Law Extrapolation

Joshua Shay Kricheli[‖], **Alexander Lawrence Reid**[‖]
Syracuse University
{jkrichel,areid04}@syr.edu

Soumajyoti Sarkar[*]
Amazon AGI Foundations

Venkata Gandikota & Paulo Shakarian
Syracuse University
{vsgandik,pashakar}@syr.edu

Abstract

Neural scaling laws approximate a language model's loss as a power-law function of parameter count N and token count D. Following Chinchilla-style compute-optimal training, many studies fit scaling laws from runs at one or a few fixed tokens-per-parameter (TPP) ratios k, with $D = kN$. We show that this collinear design, combined with the empirically common near-equality of the exponents governing N and D, induces an inherent ill-conditioning in the Gauss-Newton least-squares problem: the condition number of the design grows as the inverse square of the gap between the N- and D-exponents. The scale coefficients become practically unidentifiable, with confidence intervals inflating by orders of magnitude, yielding a "sloppy" model whose extrapolations degrade sharply off the training ray. We prove this for four scaling-law formalisms and derive a closed-form TPP-diversity threshold that characterizes well-conditioned estimation to leading order. Empirically, in controlled experiments with 5M-77M-parameter models, non-collinear designs outperform collinear ones on held-out data with a 97.3% win rate across four laws and five corpora, and 98.7% in bfloat16. We further show the degeneracy is rooted in Jacobian geometry and is not an artifact of the loss function: any smooth objective whose Gauss-Newton weights are bounded above and away from zero inherits the same ill-conditioning.

1 Introduction

Scaling laws have become a cornerstone of large language model (LLM) research, providing a predictive framework for estimating model performance as a function of compute, dataset size, and parameter count (Hestness et al., 2017; Rosenfeld et al., 2020; Kaplan et al., 2020; Alabdulmohsin et al., 2022; Caballero et al., 2023; Bahri et al., 2024). The seminal Chinchilla study (Hoffmann et al., 2022) proposed that model size N and training tokens D scale linearly under a fixed compute budget with $D \approx 20N$. A recent survey of 50+ studies finds the fitted law shifts substantially under changes to the included D/N range, with the optimal ratio a recurring source of disagreement (Li et al., 2025b). These works established the fixed tokens-per-parameter (TPP) heuristic in the compute-optimal setting.

The same fixed-TPP grids are now widely used to fit scaling laws as well. Cerebras-GPT (Dey et al., 2023) trains 111M-13B models at $D = 20N$ and fits only a loss-versus-compute law; Gadre et al. (2025) parameterize over-trained scaling along constant-M lines, each slice collinear, though their

[‖] Equal contribution.
[*] This work does not relate to the author's position at Amazon.

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

pooled grid passes our threshold with equal weight per ratio (Appendix C); the OLMo ladder fits Chinchilla at fixed Chinchilla multiples (Bhagia et al., 2025); and DataDecide reports that eight scaling-law variants fitted at $D = 100N$ fail to beat a naïve single-scale baseline (Magnusson et al., 2025). The practical stakes are high: Choshen et al. (2025) fit over $1{,}000$ laws to 485 published models and find estimates sensitive to the choice and coverage of training configurations. Yet a fixed TPP ratio prescribes how to train one model, not how to estimate a scaling law. When all training runs lie on a single ray $D = kN$ in the (N, D) plane, the predictor variables become collinear, and combined with the empirically common near-equality of the N and D exponents this renders the scaling-law parameters *practically unidentifiable*. We prove this for four scaling-law formalisms - Chinchilla, unsaturated repeated-data (Muennighoff et al., 2025), Kaplan (Kaplan et al., 2020), and Droppo-Elibol (Droppo and Elibol, 2021) - via Gauss-Newton analysis. Because the degeneracy is in the Jacobian geometry rather than in the loss, any smooth objective whose Gauss-Newton weights are bounded above and away from zero inherits the same ill-conditioning (Appendix B.3). The law becomes a "sloppy" model (Section 2): a near-continuum of parameters fits training data indistinguishably yet diverges off the training ray. This directly corrupts isoFLOP prediction: isoFLOP curves cut across the training ray and thereby probe the sloppy direction that collinear fitting leaves unconstrained. Constructively, we give an identified reduced model recovering the combined effect along a single ray (Definition 3), and show empirically that non-collinear designs beat collinear ones when evaluated on the same held-out models (Section 4).

Key contributions. (i) a closed-form TPP-diversity threshold, checkable before training, that characterizes well-conditioned scale terms to leading order (Proposition 1); (ii) a Gauss-Newton analysis of one ray: it is inherently ill-conditioned when the exponents governing model size and data size are close, with the scale-pair condition number growing as the inverse square of the exponent gap (the full one at least as fast) - established for four scaling-law formalisms, with rates that do not depend on model size and a corresponding confidence-interval inflation result (Proposition 2, Corollary 1); (iii) a holdout-prediction ordering separating one-ray from non-collinear designs (Theorem 1; Lemma 3: a margin for $K \geq 2$); beyond both, non-collinear (NC) designs also beat our 12-ray collinear (CO) fan in 97.3% of comparisons in controlled experiments at 5M-77M parameters; (iv) an identified reduced parametrization that recovers what a single-ray design can constrain (Definition 3); and (v) a fully reproducible experimental suite - $\sim$1,900 trained LLMs, all checkpoints, training metrics, and training and analysis code - all publicly released.[2]

Paper road-map. The rest of this paper is as follows. Section 3 develops the Gauss-Newton framework, derives the TPP-diversity threshold, proves condition-number bounds for four formalisms on one ray and confidence-interval inflation, and defines induced isoFLOP curves; proofs are in Appendix B. Section 4 validates the analysis with controlled experiments across five corpora, four laws, and two floating-point precisions. Section 5 synthesizes the empirical patterns, cautions against interpreting separate data versus model contributions from single-ray fits, and states the limits of the evidence: model scale, one architecture, English text, one training recipe, and pre-training loss.

2 Related work

Predictable power-law scaling of generalization error was first reported by Hestness et al. (2017) and formalized for autoregressive language models by Kaplan et al. (2020) using bivariate fits over grids crossing model sizes with dataset subsets. Bahri et al. (2024) linked the observed exponents to data-manifold structure via variance- and resolution-limited regimes. Follow-ups to Chinchilla examined data availability (Villalobos et al., 2024), overtraining (Touvron et al., 2023), and unbounded compute with fixed data (Kim et al., 2026); we extend our analysis to repeated-data regimes (Muennighoff et al., 2025) in Appendix B.7. Sardana et al. (2024) sweep ratios up to $10{,}000$ TPP, producing inherently non-collinear designs that our diagnosis predicts should be well-identified (Section 3) - yet several widely used ladders were fit in the ill-conditioned regime we formalize. Besiroglu et al. (2024) found Chinchilla's reported parametric estimates inconsistent with its other two estimation methods, and Porian et al. (2024) attribute the Kaplan-Chinchilla discrepancy to minor recipe differences (last-layer cost, warm-up, scale-dependent tuning); on a single ray, our result explains why such perturbations can move coefficients far. Bergsma et al. (2025) show even optimizer hyperparameters (weight decay, batch size) obey power laws in the TPP ratio D/N, so fixing a single ray couples

[2]Code: https://github.com/lab-v2/tpp-coverage. Checkpoints and metrics: https://huggingface.co/datasets/leibnitz-lab/colinear_scaling_models.

Table 1: Six of the 24 published designs (all: Table 11). $k_\ell = D/N$: the training ratios; K: their number; $R = k_K/k_1$; V_K/τ_K: margin over the threshold of Proposition 1 at $\kappa_{\text{target}} = 100$ (≥ 1 passes). * No exponent; $\beta_{\text{eff}} = 0.28$ used.

Design	K	k_ℓ	R	V_K/τ_K
Hoffmann et al. (2022)	151	0.0361–341	9.5×10^3	5.55
Muennighoff et al. (2025)	95	0.0912–4.26×10^4	4.7×10^5	5.76
Choshen et al. (2025)*	243	0.874–4284	4.9×10^3	3.74
Hu et al. (2024)	6	10–60	6	0.30
Bhagia et al. (2025)	4	20, 40, 100, 200	10	0.47
Magnusson et al. (2025)*	1	100	1	0.00

recipe with scale. Concurrently, Volkova et al. (2026) report that separate Chinchilla-style fits per optimizer are ill-conditioned, with correlated pairs (A, α) and (B, β) that they attribute to the law form and the narrow N and D ranges; their ratios $\{30, 50, 100, 200\}$ also fail Eq. (2) (Appendix C). Schaeffer et al. (2026) propose sidestepping multicollinearity via compute-envelope or gold-reference parameterizations - both abandon the (N, D) decomposition. Zhang et al. (2026) pursue a related goal at the downstream-evaluation level, estimating compute-to-capability boundaries via quantile regression on post-trained models. Our work is complementary: we keep the (N, D) decomposition and give designs that *restore* its identifiability (Proposition 1). The Relative-Based Scaling Law of Yue et al. (2025) is largely a single-axis sweep in model size, which our analysis flags as insufficient for multivariate identifiability. Farseer (Li et al., 2025a) adds an N-D coupling term and trains $\sim$1,000 models on a diverse (N, D) grid, reporting a 433% reduction in extrapolation error over Chinchilla; even such interaction terms remain ill-conditioned under $D = kN$ (Appendix B.16). Shukor et al. (2025) similarly extend scaling laws to data-mixture weights, enlarging the multivariate design space where collinearity matters. Hu et al. (2026) sidestep parametric fitting altogether by training a neural extrapolator over checkpoint trajectories - orthogonal to our diagnosis of *why* the parametric fit is fragile. Statistically, the pathology we analyze is the nonlinear-regression analog of multicollinearity (Kim, 2019; Vatcheva et al., 2016; Dormann et al., 2013; Montgomery et al., 2012; O'Brien, 2007) and of the "sloppy" model phenomenon (Transtrum et al., 2010, 2015), in which narrow model-manifold cross-sections yield fragile estimates - here induced by fixed-TPP sampling. Like work that detects or reduces the errors of learned models (Kricheli et al., 2024; Leiva et al., 2026; Taparia et al., 2026), the diagnosis flags when a fitted law's off-ray predictions cannot be trusted.

Eq. (2) screens a design from its training ratios and one exponent. Of 206 candidates, seeded by Li et al. (2025b), 24 fit a bivariate law on a recoverable design; at target condition number 100, 11 pass and 13 fail, 3 of them at one ratio (Table 1; Appendix C). The test is proved for collinear designs; none of the 6 collinear ones passes, and no pass fixes the coefficients. Most failures, like our CO fan at $\beta_{\text{eff}} = 0.28$ ($\kappa \approx 250$), miss the target by under fourfold: a weakly determined N-D split, not an unusable fit.

Relative to classical identifiability analyses. The pathology is classical; this paper adds:

(i) a rate $\kappa_{A,B} = \Theta(\varepsilon^{-2})$ in a measured gap (Proposition 2, Eq. (37));
(ii) a threshold checkable before training from one exponent prior that guarantees conditioning to leading order (Proposition 1), where variance inflation factors need a fit and locally optimal design all parameters (Atkinson et al., 2007);
(iii) severity by law: Chinchilla's full $\kappa(J^TJ)$ is $>10^5\times$ Kaplan's (App. B.8), saturated repeated data likely escapes it, and an interaction term adds a second sloppy direction (Appendices B.7, B.16);
(iv) what one ray identifies (Proposition 4);
(v) a holdout-error ordering (Theorem 1).

3 Formal analysis

In plain terms: a collinear design trains at K TPP ratios ($D = k_\ell N$). On one ray the two power-law terms rise and fall together, so only their exponent gap ε (Definition 2) separates them; the spread of the ratios separates them at any gap. Section 3.1 gives the spread that keeps the scale coefficients well conditioned; Section 3.2 treats one ray ($K = 1$), where they are individually free but jointly constrained ("sloppy") and off-ray predictions inherit that uncertainty. Table 8 collects the notation.

Table 2: Jacobian columns at the i-th data point under $D = kN$ (with i subscripts dropped for brevity). Rows correspond, in order, to Chinchilla (Hoffmann et al., 2022), Repeated-data (Muennighoff et al., 2025), Kaplan (Kaplan et al., 2020), and Droppo-Elibol (Droppo and Elibol, 2021). For each law, $\mathbf{j}_2$ is a scalar multiple of $\mathbf{j}_1$ times $N^{\pm\varepsilon}$ ($N^{\tilde{\varepsilon}}$, $\tilde{\varepsilon} = \alpha - \beta$; $s = \alpha_N/\alpha_D$), exposing their shared collinearity.

Law ($\hat{L}$)	**Collinear pair**	$\mathbf{j}_1$	$\mathbf{j}_2$	ε
$AN^{-\alpha} + BD^{-\beta} + E$	(A, B)	$N^{-\alpha}$	$k^{-\beta}N^{-\alpha}N^{\tilde{\varepsilon}}$	$\lvert\alpha - \beta\rvert$
$AN'^{-\alpha} + BD'^{-\beta} + E$	(A, B)	$N'^{-\alpha}$	$k^{-\beta}(D'/kN')^{-\beta}N'^{-\alpha}N'^{\tilde{\varepsilon}}$	$\lvert\alpha - \beta\rvert$
$\left[(N_c/N)^s + D_c/D\right]^{\alpha_D}$	(N_c, D_c)	$\propto N^{-s}$	$\propto N^{-1}$	$\lvert 1 - s\rvert$
$[(N_C/N)^{\gamma_N} + (D_C/D)^{\gamma_D} + E^{1/\alpha}]^{\alpha}$	(N_C, D_C)	$\propto N^{-\gamma_N}$	$\propto N^{-\gamma_N}N^{\varepsilon}$	$\lvert\gamma_N - \gamma_D\rvert$

The standard Chinchilla scaling law (Hoffmann et al., 2022, Eq. 2) approximates the loss as a sum of two power-law terms - capturing limited model capacity and data size - plus an irreducible loss E, a form observed across deep learning domains (Hestness et al., 2017; Kaplan et al., 2020, Sec. 3):

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

where $A, B, \alpha, \beta, E \in \mathbb{R}_{>0}$ and $N, D \in \mathbb{R}_{>0}$.

Definition 1 (Scaling-law experimental dataset)**.** Let $\mathcal{A}$ be a training scheme (architecture, optimizer, algorithm, etc.). Each experiment records loss L_i for a model of size N_i on D_i tokens, yielding $\mathcal{D} = \{(N_i, D_i, L_i)\}_{i=1}^m$. Let $N_1 < \cdots < N_n \in \mathbb{N}_{>0}$ be the model sizes and $k_1 < \cdots < k_K \in \mathbb{R}_{>0}$ the TPP ratios. A *collinear* design pairs each (N_i, k_ℓ) as $(N_i,\ k_\ell N_i,\ L_{i\ell})$, so $m = nK$; the special case $K = 1$ is fully collinear. A parametric law $\hat{L}(N, D; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \mathbb{R}^p$ is fit to $\mathcal{D}$ by minimizing the discrepancy between predicted and observed losses; observations are split into a training set $\mathcal{D}_{\text{train}}$ and a holdout $\mathcal{H}$ of $n_{\mathcal{H}}$ points (possibly at TPP ratios off the training manifold).

Definition 2 (Exponent gap by scaling law)**.** Define ε as a *law-specific exponent gap* in each law's native notation (see Table 2): **Chinchilla** (and repeated-data on effective (N', D')): $\varepsilon \coloneqq |\alpha - \beta|$; **Kaplan** (exponents α_N, α_D): $\varepsilon \coloneqq |1 - \alpha_N/\alpha_D|$; **Droppo-Elibol** (exponents $\alpha_N, \alpha_D, \alpha$ with $\gamma_N \coloneqq \alpha_N/\alpha,\ \gamma_D \coloneqq \alpha_D/\alpha,\ \alpha > 0$): $\varepsilon \coloneqq |\gamma_N - \gamma_D|$. For example, Chinchilla reports $\alpha \approx 0.34$ and $\beta \approx 0.28$; the single-variable fits of Kaplan et al. (2020) give $\alpha_N \approx 0.076$ and $\alpha_D \approx 0.095$.

3.1 Gauss-Newton (GN) framework and multi-ray designs

We fit by nonlinear least squares: minimize $S(\boldsymbol{\theta}) \coloneqq \frac{1}{2}\|\mathbf{r}(\boldsymbol{\theta})\|^2$ over $\boldsymbol{\theta} \in \mathbb{R}^p$, where the residuals $r_i \coloneqq L_i - \hat{L}(N_i, D_i; \boldsymbol{\theta})$ stack into $\mathbf{r}(\boldsymbol{\theta})$, with *Jacobian* $J \in \mathbb{R}^{m\times p}$ given by $J_{ij} = \partial r_i/\partial\theta_j$. We write $\kappa(M) \coloneqq \lambda_{\max}(M)/\lambda_{\min}(M)$ for the condition number of a symmetric positive-definite matrix M. GN linearizes $\mathbf{r}$, giving the normal equations $(J^TJ)\,\Delta\boldsymbol{\theta} = -J^T\mathbf{r}$ (Appendix B.1). For each law's scale coefficients (A, B), let $\kappa_{A,B} \coloneqq \kappa\big((J^TJ)_{A,B}\big)$ on the (A, B) principal submatrix of J^TJ. Training at $K \geq 2$ rays, the spread of the ratios restores well-conditioned estimation of the scale pair:

Proposition 1 (TPP diversity and scale-pair conditioning)**.** *For the Chinchilla and unsaturated repeated-data laws (Kaplan, Droppo-Elibol: a heuristic, Appendix B.12), let J be the Jacobian at the true parameters (well-specified law). For any $\kappa_{\text{target}} \geq 1$ and a collinear design (Definition 1) with ratios $k_1 < \cdots < k_K$ ($K \geq 1$), to leading order as $\varepsilon \to 0$, we have*

$$\kappa_{A,B} \leq \kappa_{\text{target}} \Longleftarrow \underbrace{\frac{1}{K}\sum_{\ell=1}^{K} k_\ell^{-2\beta_{\text{eff}}} - \left(\frac{1}{K}\sum_{\ell=1}^{K} k_\ell^{-\beta_{\text{eff}}}\right)^2}_{V_K} \geq \underbrace{\frac{\left(K + \sum_{\ell=1}^{K} k_\ell^{-2\beta_{\text{eff}}}\right)^2}{K^2\,\kappa_{\text{target}}}}_{\tau_K}, \tag{2}$$

conversely, $\kappa_{A,B} \leq \kappa_{\text{target}}$ implies $V_K \geq \tau_K/(1 + 1/\kappa_{\text{target}})^2$; V_K is the variance of $\{k_\ell^{-\beta_{\text{eff}}}\}_{\ell=1}^K$, and β_{eff} is the law-specific data-size exponent: β for Chinchilla/repeated-data (Kaplan, Droppo-Elibol: Table 9). For $K = 1$, $V_1 = 0 < \tau_1$, so (2) never holds; for $K = 2$ it holds iff $R \coloneqq k_2/k_1 \geq R_{\min}(k_1)$ of (71), which increases with $k_1 \geq 1$ (a B-column scale effect) and is infinite once $\sqrt{\kappa_{\text{target}}}\,k_1^{-\beta_{\text{eff}}} \leq 2 + k_1^{-2\beta_{\text{eff}}}$; for $K > 2$ interior rays reduce *V_K at fixed endpoints, and no K-ray design exceeds the V_K/τ_K of its endpoints by more than the factor $(1 + k_1^{-2\beta_{\text{eff}}}/2)^2$ of (72).*

Proof. Appendix B.12; Appendix B.15 builds a design recipe on it.

Since (2) can be evaluated *before* training using a literature estimate of β_{eff} (e.g. $\beta \approx 0.28$ for Chinchilla), it is an a-priori screen: the exact $\kappa_{A,B}$ is 0.06-3.4 times its leading-order value over 432 two-ratio designs, so the exact value for the planned design decides (Appendix B.12). For $K \geq 2$ the same block orders holdout errors: NC predicts better than CO in expectation when $\operatorname{tr}(G_{A,B}^{-1} M)$, with $G_{A,B} \coloneqq (J^T J)_{A,B}$ and M the holdout's second moment of the pair columns, exceeds NC's predictive variance (Lemma 3).

3.2 The one-ray limit ($K = 1$)

On a single ray ($K = 1$) - a common practical case ($k \approx 20$ for Chinchilla (Hoffmann et al., 2022; Villalobos et al., 2024)) - substituting $D = kN$ into (1) gives $L(N, kN) = AN^{-\alpha} + Bk^{-\beta}N^{-\beta} + E$. Setting $\beta = \alpha$ collapses the two power-law terms, yielding a *reduced model*:

Definition 3 (Reduced Chinchilla model under fixed TPP)**.**

$$L(N;\, \psi, \alpha, E) \;=\; \left(A + Bk^{-\alpha}\right) N^{-\alpha} + E \;=\; \psi N^{-\alpha} + E, \tag{3}$$

where $\psi \coloneqq A + B\, k^{-\alpha}$ combines the Chinchilla-scale coefficients from (1) and $k \in \mathbb{R}_{>0}$ is the fixed TPP ratio on that ray.

With free exponents one ray has a second near-null direction: numerically, $\kappa(J^T J)$ grows as ε^{-6} for Chinchilla (Appendix B.11); rates below are for $\kappa_{A,B}$, bounds for $\kappa(J^T J)$.

Proposition 2 (Full-matrix conditioning)**.** *Let $J \in \mathbb{R}^{m \times p}$ have columns $\mathbf{j}_1, \ldots, \mathbf{j}_p$ of norm $O(1)$, and suppose two columns $a \neq b$ satisfy $\mathbf{j}_b = c\,\mathbf{j}_a + \boldsymbol{\delta}$ with a fixed $c \neq 0$, $\|\mathbf{j}_a\| = \Theta(1)$ and $\|\boldsymbol{\delta}\| = O(\varepsilon)$, where ε is the law-specific exponent gap (Definition 2). Let $\lambda_{\max}(J^T J)$ and $\lambda_{\min}(J^T J)$ denote the largest and smallest eigenvalues of $J^T J$. Then as $\varepsilon \to 0$, $\lambda_{\max}(J^T J) = \Theta(1)$, $\lambda_{\min}(J^T J) = O(\varepsilon^2)$, and*

$$\kappa(J^T J) = \Omega(\varepsilon^{-2}); \tag{4}$$

if the part of $\boldsymbol{\delta}$ orthogonal to $\mathbf{j}_a$ has norm $\Theta(\varepsilon)$, $\kappa((J^T J)_{a,b}) = \Theta(\varepsilon^{-2})$.

Proof. Appendix B.3. The lower bound applies to all four laws of Table 2 (repeated-data unsaturated): on a ray $D = kN$ each scale pair is near-proportional (Appendices B.5-B.9).

On a single ray, the reduced model (3) is identified (Proposition 4) while the full model is ill-conditioned (Proposition 2). The least-squares estimator's uncertainty comes from inverting $J^T J$ (Björck, 1996, §2.8.3, eq. (2.8.7), p. 118): when the data cannot distinguish A from B, only their combination ψ is pinned down, leaving each coefficient individually highly uncertain. The confidence interval on A therefore widens by a factor $\Omega(\varepsilon^{-1})$ relative to the one on ψ:

Corollary 1 (Confidence interval inflation)**.** *Under i.i.d. Gaussian noise with variance σ^2, on a single ray $D = kN$ with $n \geq 5$ model sizes, the nominal 95% least-squares confidence intervals $\mathrm{CI}_{0.95}$ for A in the full Chinchilla model (1) and for the combined coefficient ψ in the reduced model (3) satisfy, as $\varepsilon \to 0$,*

$$\frac{\mathrm{CI}_{0.95}(A)}{\mathrm{CI}_{0.95}(\psi)} \;=\; \Omega\big(\varepsilon^{-1}\big)\,. \tag{5}$$

Proof. Appendix B.11.

The asymptotic factors for three of the four laws are collected in Table 3.

Scope of the rates. None of these bounds assumes that the scale pair is the only weak direction of $J^T J$. The bound $\kappa(J^T J) = \Omega(\varepsilon^{-2})$ needs only near-proportional columns (Appendix B.3), $\kappa(J^T J) \geq \kappa_{A,B}$ by Cauchy interlacing (Appendix B.12), and the Schur-complement floor on the interval of A holds for any full Jacobian (Appendix B.11).

Theorem 1 (Holdout prediction regimes)**.** *For any law of Table 2 (repeated-data unsaturated) with i.i.d. Gaussian residuals of variance σ^2, let CO denote a collinear design ($K \geq 1$), NC a non-collinear design with dense two-dimensional (N, D) coverage, both with full-rank Jacobians (NC's also at $\varepsilon = 0$), and let $\mathcal{H}$ denote a holdout with a nonzero fraction of off-ratio points ($k' = D'/N'$ differs from every training ratio). For $\mathrm{X} \in \{\mathrm{CO}, \mathrm{NC}\}$, $\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{X}}$ is the root-mean-square error of $\hat{L}$ against $\mathbb{E}[L]$ on $\mathcal{H}$ under that training design, and $R_{\mathcal{H}}^{2\,\mathrm{X}}$ the corresponding coefficient of determination on $\mathcal{H}$. The regimes below hold at leading order in σ (linearizing both fits), then as $\varepsilon \to 0$.*

Table 3: Statistical identifiability of scale coefficients under single-TPP designs, using representative published exponent values. ε^{-1}, ε^{-2}: rates of $\mathrm{CI}(A)/\mathrm{CI}(\psi) = \Omega(\varepsilon^{-1})$ (Corollary 1; Kaplan, Droppo-Elibol, without a reduced model: the scale coefficient's own interval, by the Schur floor of Appendix B.11; the design constant is about 10^{-2}, so the proven floor is 0.16, while the computed five-parameter ratio grows, numerically, as ε^{-3}) and $\kappa_{A,B} = \Theta(\varepsilon^{-2})$ (Appendix B.5). Kaplan: the single-variable fits of Kaplan et al. (2020). Droppo-Elibol: gaps of our fits. Repeated-data inherits Chinchilla's exponents. On one ray at $k = 1$-20 and our 14 sizes, $\kappa_{A,B}$ is 10^3-10^4 for Chinchilla and, with unit-norm columns, 1.6-1.7$\times 10^2$ for Kaplan (Appendices B.5 and B.8).

Law	**Exponents**	ε	ε^{-1}	ε^{-2}
Chinchilla	$\alpha \approx 0.34$, $\beta \approx 0.28$	0.06	17	278
Kaplan	$\alpha_N \approx 0.076$, $\alpha_D \approx 0.095$	0.20	5	25
Droppo-Elibol	fitted (Table 48)	0.18-0.80	1.3-5.6	1.6-32

Regime A (CO a single ray, $K = 1$, so $V_1 = 0 < \tau_1$):

$$\mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{NC}}] < \mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{CO}}], \tag{6}$$

and $\mathbb{E}[R^{2\,\mathrm{NC}}_{\mathcal{H}}] > \mathbb{E}[R^{2\,\mathrm{CO}}_{\mathcal{H}}]$ *(misspecification extension for the expected mean squared error: Appendix B.13).*

Regime B (several rays, $K \geq 2$, CO's Jacobian of full rank at $\varepsilon = 0$): $\mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{CO}}]/\mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{NC}}] = \Theta(1)$*, without Regime A's ε-driven separation, and Proposition 1 bounds $\kappa_{A,B}$ when $V_K \geq \tau_K$.*

Proof. Appendix B.13 (Gauss-Newton linearization, a Cauchy-Schwarz bound, RMSE ordering).

3.3 Applying to respective scaling laws

Table 2 gives, for each law, the collinear pair, its Jacobian columns on a single ray $D = k_\ell N$ and its exponent gap, so Proposition 2 yields $\kappa(J^T J) = \Omega(\varepsilon^{-2})$ in all four cases (Appendices B.5, B.7-B.9; statistical interval inflation: Corollary 1, Table 3). Kaplan's gap is 0.20 at the single-variable exponents of Kaplan et al. (2020) (0.26 at their joint fit's) and at least 0.105 in our single-precision (FP32) seed-0 fits, and Kaplan is the best-conditioned law on our designs (with unit-norm columns, the smallest $\kappa(J^T J)$ of three laws at each seed-0 fit); its worst off-ray R^2 (0.50 against 0.78) comes from bias (Appendix B.8). Chinchilla and unsaturated repeated-data, with small near-equal exponents, have a CI-ratio floor $\Omega(\varepsilon^{-1})$: at Hoffmann et al.'s $\varepsilon = 0.06$ it is 0.16, and the computed five-parameter ratio is of order 10^4 (Appendix B.11). Droppo-Elibol is fit-dependent: our fits give $|\gamma_N - \gamma_D|$ from 0.18 to 0.80 (Table 48), and the well-identified outer parameters (α, E) do not relieve the (N_C, D_C) degeneracy. The reduced model (Definition 3) is therefore most needed for Chinchilla-class fits.

3.4 IsoFLOP prediction under collinearity

IsoFLOP curves are contours of constant training compute $C = 6ND$ (Hoffmann et al., 2022). They cut across the training ray and probe the sloppy parameter direction that collinear training leaves unconstrained, making them an informative diagnostic.

Definition 4 (Induced isoFLOP curves plot)**.** Let $\hat{L}(N, D; \boldsymbol{\theta})$ be a fitted scaling law and $\mathcal{H} = \{(N_i, D_i, L_i)\}_{i=1}^{n_{\mathcal{H}}}$ a holdout set of size $n_{\mathcal{H}}$ with induced compute budgets $C_i := 6N_i D_i$. The *induced isoFLOP curves* are the family

$$\{N \mapsto \hat{L}(N, C_i/(6N); \boldsymbol{\theta}) \,:\, N \in [N_{\min}, N_{\max}]\}_{i=1}^{n_{\mathcal{H}}}, \tag{7}$$

with $N_{\min} := \min_j N_j$ and $N_{\max} := \max_j N_j$. The *induced isoFLOP curves plot* draws this family on the (N, L)-plane with each holdout point (N_i, L_i) overlaid on its curve.

Proposition 3 (RMSE invariance under isoFLOP reparameterization)**.** *With $\hat{L}, \mathcal{H}, \{C_i\}$ as in Definition 4, the RMSE of $\hat{L}$ on $\mathcal{H}$ is invariant under $(N, D) \mapsto (N, C)$: evaluating the i-th induced isoFLOP curve at $N = N_i$ yields the same error as evaluating $\hat{L}(N_i, D_i)$ directly (Appendix B.14).*

4 Empirical evaluation

Empirically evaluating dense TPP coverage. To validate the formal statements of Section 3, we fit each scaling law under two training designs - collinear (CO) and non-collinear (NC) - and measure holdout error on a shared CO/NC split across five pre-training corpora: three common and two specialized (see Appendix E.5). All models are LLaMA-style transformers trained with AdamW on a cosine schedule; full hyperparameters are in Appendix E. Model sizes span 5.04-76.5M parameters across 14 configurations ($N_{\max}/N_{\min} \approx 15$; see Table 17). Tokenization uses cl100k_base (V=100,277), and N counts all trainable parameters (embeddings included) per the Hoffmann-Chinchilla convention (Hoffmann et al., 2022); we apply this convention uniformly when fitting all four laws - including Kaplan (Kaplan et al., 2020), which originally defines N on non-embedding parameters - and no fit involves a compute budget C. We fit the training loss of each epoch, the mean of its last 100 batches (Appendix E.4). **Experimental designs.** Both designs share identical architectures, optimizers, and schedules; the *only* difference is how (N, D) pairs are sampled, and total run counts are comparable. **Collinear (CO):** Every run sets $D = kN$ for $k \in \mathcal{K}_{\text{train}} \coloneqq \{1,\ 1.5,\ 1.9,\ 2,\ 2.5,\ 2.7,\ 3,\ 3.3,\ 3.5,\ 4,\ 4.5,\ 5\}$. Although 12 ratios are used, each defines a ray through the origin; the configurations form a *fan of rays* rather than filling a two-dimensional region, producing the collinearity diagnosed in Section 3. **Non-Collinear (NC):** The 10 smallest model sizes are crossed with 12 dataset sizes ($\sim$ 10M-$\sim$ 276M tokens), yielding a 10×12 grid spanning a two-dimensional region. The implied ratio $k = D/N$ varies freely, decoupling the two axes and eliminating the rank deficiency. The grid is visualized in Appendix F.1. We fit the four laws with 100 random L-BFGS-B restarts and a differential-evolution fallback under 30 optimizer seeds (Appendix F.6); repeated-data takes each epoch's checkpoint as an observation, and the other three fit each epoch (1, 2, 3) separately. A pure-bfloat16 (**BF16**) round ($k \in \mathcal{K}_{\text{train}}$; Appendix F.4) tests numerical precision; in a preliminary high-TPP experiment (small-scale, $k \in \mathcal{K}_{\text{big}} \coloneqq \{10,\ 11,\ 12,\ 13,\ 14,\ 15\}$, holdouts up to 20) with Wikipedia and Chinchilla, NC wins 19 of 30 comparisons (63.3%, CI $[45.5\%, 78.1\%]$, which includes 50%; see Appendices F.7 and E.8).

Table 4: R^2 and RMSE summary by training design over 30 optimizer seeds. 95% CI: optimizer-noise interval (Appendix E.8); Std: across (corpus, law, epoch) combinations; best per split in **bold**.

		R^2				RMSE			
Split	**Design**	Mean	95% CI	Median	Std	Mean	95% CI	Median	Std
Train ($\mathcal{D}$)	CO	**0.9848**	**[0.985, 0.985]**	**0.9850**	**0.0047**	**0.1737**	**[0.173, 0.174]**	**0.1616**	**0.0430**
	NC	0.9526	[0.952, 0.953]	0.9542	0.0172	0.2023	[0.202, 0.203]	0.1869	0.0443
Holdout ($\mathcal{H}$)	CO	0.8370	[0.832, 0.842]	0.8811	0.1129	0.2373	[0.233, 0.241]	0.2163	0.1229
	NC	**0.9319**	**[0.930, 0.934]**	**0.9392**	**0.0269**	**0.1561**	**[0.154, 0.158]**	**0.1242**	**0.0599**

Evaluation splits. The collinear holdout $\mathcal{H}_{\text{col}}$ lies just beyond the training fan: the next model size (55.5M) at every training ratio and at $k \in \mathcal{K}_{\text{test}} \coloneqq \{6,\ 6.2,\ 6.5,\ 6.7,\ 7\}$, every training size at $k = 6$, and the three largest sizes (63.9-76.5M) at $\mathcal{K}_{\text{test}}$. The non-collinear holdout $\mathcal{H}_{\text{nc}}$, built the same way in D ($D = 300$M for $k = 6$, $D \in [300, 401]$ M for $\mathcal{K}_{\text{test}}$), lies off the fan in the (N, D) plane. All metrics use the *unified holdout* $\mathcal{H} = \mathcal{H}_{\text{col}} \cup \mathcal{H}_{\text{nc}}$, so every comparison is head-to-head; per split (Appendix F), NC wins 98.7% of the comparisons on $\mathcal{H}_{\text{nc}}$ and 72.1% on $\mathcal{H}_{\text{col}}$, along CO's own fan, where Wikipedia and RedPajama are near parity (48.3% and 50.3%).

TPP coverage tracking. We widen each design's data axis step by step, adding TPP ratios (CO) or dataset sizes (NC), and refit each law at each step; the coverage fraction is the share of levels included (1.0: all of $\mathcal{K}_{\text{train}}$ or all 12 D sizes). The two curves take independent paths through the (N, D) plane (setup and plots: Appendices F.5 and F.8).

Aggregate metrics. Table 4 gives the mean, median and spread of R^2 and RMSE over law-dataset combinations per design and split, and Table 5 tallies seed-paired wins by dataset, law and epoch. Under pure-BF16 training on Wikipedia (Table 6) the effect persists: NC reaches a unified holdout R^2 of 0.966 against CO's 0.941 and wins 98.7% (CI $[96.6\%, 99.5\%]$; Table 23), on a holdout without 14 of the 42 collinear holdout runs. Per-dataset breakdowns are in Appendices F and G.

Table 5: Win rate (NC vs. CO) on $\mathcal{H}$, seed-paired; the epoch column omits repeated-data, fit on all epochs at once. Overall: 97.3% (1,460/1,500), 95% Wilson CI $[96.4\%, 98.0\%]$ over comparisons that share one training run per (N, D), so fitting variability only. NC has the lower mean holdout RMSE in 49 of the 50 (corpus, law, epoch) cells.

Dataset	
C4	93.0%
Cosmopedia	94.7%
peS2o	99.3%
RedPajama	100.0%
Wikipedia	99.7%

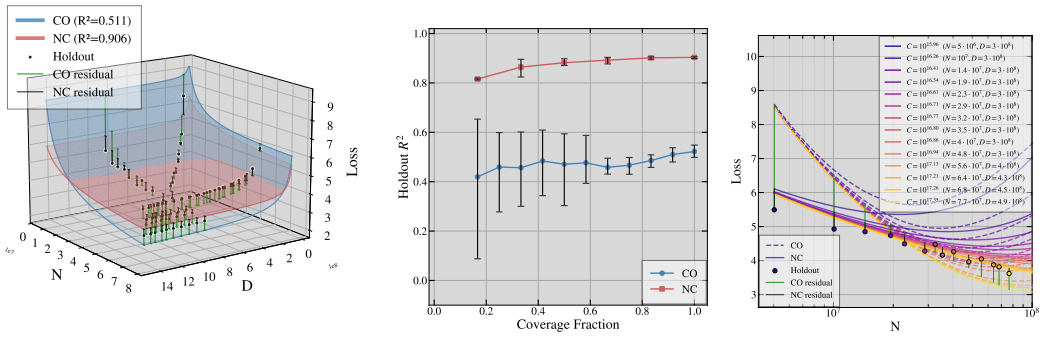
Scaling law	
Chinchilla	97.6%
Droppo-Elibol	98.0%
Kaplan	95.6%
Repeated-data	100.0%

Epoch	
first	99.8%
second	97.6%
final	93.8%

Table 6: BF16 R^2 and RMSE by training design over 30 optimizer seeds. 95% CI: optimizer-noise interval (Appendix E.8); Std: across (law, epoch) combinations; best per split in **bold**.

		R^2				**RMSE**			
Split	**Design**	Mean	95% CI	Median	Std	Mean	95% CI	Median	Std
Train ($\mathcal{D}$)	CO	0.9896	[0.990, 0.990]	0.9916	0.0029	0.1854	[0.185, 0.185]	0.1606	0.0353
	NC	**0.9907**	**[0.991, 0.991]**	**0.9915**	**0.0020**	**0.1457**	**[0.145, 0.146]**	**0.1428**	**0.0171**
Holdout ($\mathcal{H}$)	CO	0.9412	[0.940, 0.942]	0.9473	0.0172	0.2546	[0.252, 0.257]	0.2402	0.0396
	NC	**0.9657**	**[0.965, 0.967]**	**0.9663**	**0.0068**	**0.1949**	**[0.192, 0.198]**	**0.1921**	**0.0203**

Example plots. Figure 1 shows the cell with the largest CO/NC gap of the 50 (more in Appendix F.8): the Kaplan CO fit on RedPajama (first epoch) reaches a holdout R^2 of only 0.51 at seed 0 (0.52 over 30 seeds), NC 0.91 (0.90). CO residuals grow toward small N (Figure 1a). In Figure 1b, **(i)** CO sits below NC at every TPP coverage level; **(ii)** CO's seed bands are 2-80× wider than NC's (7× at full coverage) and narrow with coverage without closing the gap; **(iii)** CO's mean R^2 oscillates while NC's rises smoothly; **(iv)** in Figure 1c, CO's dashed isoFLOP curves splay apart at high N while NC's stay tight. For Kaplan this error is bias, not conditioning (Appendix B.8).



(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 1: Kaplan law on RedPajama, first epoch (seed-0 CO holdout $R^2 = 0.51$, NC $R^2 = 0.91$; CO blue, NC red in (a)-(b), dashed and solid in (c)). (a), (c): seed 0, whose NC (N_c, D_c) is a start draw (Appendix E.7); (b): mean over 30 seeds with 5th-95th percentile bars, from two rays on, as CO's one-ray mean is $R^2 = -1.35$. **(a)** Fitted surface $\hat{L}(N, D; \boldsymbol{\theta})$ with holdout points and residual bars. **(b)** Holdout R^2 vs. TPP coverage. **(c)** Induced isoFLOP curves (Definition 4), one per kept holdout point (highest D per unique N), with residual bars at each N_i; curves and markers share a colormap ordered by $C_i = 6N_iD_i$.

Budget-matched subset enumeration. Table 7 asks empirically whether, below the threshold ($V_K < \tau_K$), NC also beats a budget-matched CO design (the same number of runs) of 1 to 12 rays. We enumerate all $2^{12} - 1 = 4{,}095$ non-empty subsets $S \subseteq \{1, \ldots, 12\}$ of available TPP ratios. Each $\mathrm{CO}(S)$ (Definition 10) is paired with a bounding-box NC design of identical cardinality

(Appendix F.3) and classified a priori via $V_K < \tau_K(\kappa_{\text{target}})$ using literature β_{eff}, a Kaplan subset by its exact $\kappa_{A,B} > \kappa_{\text{target}}$ at the parameters of Kaplan et al. (2020) (Appendix F.3). Because the threshold depends on κ_{target}, we average the NC-win rate over 2,000 values drawn log-uniformly from $[1, 10^9]$. Each κ_{target} is repeated under 22 seeds that vary the bounding-box construction and the random restarts of the fit optimizer (300 restarts per fit; random 70% subsample of $\mathcal{H}$). When no subset is below it for a given κ_{target}, that seed is skipped. NC beats CO on average below the threshold for all three laws: 68.5% (Chinchilla), 57.4% (Droppo-Elibol), and 80.6% (Kaplan) averaged across corpora, all with CIs excluding 50%. The effect is strongest on the web and reference corpora C4, RedPajama, and Wikipedia (NC wins 77-88% for Chinchilla); the two specialized corpora (Cosmopedia and peS2o) show win rates near or below 50% for Chinchilla and Droppo-Elibol and 71-72% for Kaplan, as discussed in Section 5.

Public training-loss releases. Ten public releases from nine studies give per-run losses on a two-dimensional (N, D) core, five as 984 fully trained runs (Appendix D). On a shared off-range holdout, equal-size NC subsets win 47% of pairs on the five and 50% on all ten at $\kappa_{\text{target}} = 100$, where every CO subset is below the threshold at $\beta_{\text{eff}} = 0.28$. Where only NC meets the threshold (mostly $\kappa_{\text{target}} \geq 10^4$) NC wins 53-76%; where both do (Theorem 1 orders none of these multi-ratio subsets), CO is ahead on the five and, on all ten, in windows centred at 1B and above. On the five, CO's relative off-ray error rises 40% ([24%, 52%]) per tenfold rise of NC's V_K/τ_K over CO's, and 28% ([−9%, 41%]) in windows centred at 1B and above. On all ten it is 4% ([−10%, 21%]); its size interaction, −39% ([−57%, −18%]), is negative, but the band estimates do not fall monotonically.

Table 7: Below-threshold NC-win rate (%) by law and dataset, first-epoch fits. Each cell: fraction of paired comparisons where $\text{RMSE}_{\mathcal{H}}^{\text{NC}} < \text{RMSE}_{\mathcal{H}}^{\text{CO}}$ below the predicted threshold ($V_K < \tau_K$; for Kaplan, the exact $\kappa_{A,B} > \kappa_{\text{target}}$). **Dataset Avg.**: mean across corpora with 95% bootstrap CI.

	Web and reference			Specialized		
Law	C4	RedPajama	Wikipedia	Cosmopedia	peS2o	**Dataset Avg.**
Chinchilla	77.1	78.6	88.0	51.8	47.1	**68.5** [67.3, 69.7]
Droppo-Elibol	62.2	67.2	65.9	45.5	46.2	**57.4** [56.1, 58.7]
Kaplan	92.4	87.2	79.8	72.4	71.0	**80.6** [79.0, 82.0]

5 Discussion and conclusions

Collinear designs overfit, non-collinear designs generalize: on training data CO achieves higher R^2 than NC in nearly every experiment (Table 4: aggregate 0.985 vs. 0.953), as a fit along a narrow fan of rays near-interpolates, but on held-out data the ranking reverses - NC raises unified-holdout R^2 from 0.837 to 0.932 and cuts RMSE from 0.237 to 0.156. **Scaling 12 TPP ratios does not rescue collinearity**: our CO design spans $\mathcal{K}_{\text{train}}$ (12 ratios) yet NC still wins 97.3% of 1,500 seed-paired comparisons (Table 5); at $\kappa_{\text{target}} = 100$ its ratios give $V_K/\tau_K = 0.40$ at Chinchilla's $\beta_{\text{eff}} = 0.28$ (3.18, a pass, at Kaplan's $\beta_{\text{eff}} = 1$), and Theorem 1 covers one ray. **The effect persists under pure BF16**: on Wikipedia NC wins 98.7% (296/300; Table 23), as in FP32 on Wikipedia (99.7%), with 14 of the 42 collinear holdout runs missing. **The pattern is consistent across corpora, laws, and epochs**: on $\mathcal{H}$, NC wins at least 93% on every corpus, law and epoch (Table 5). The combined **takeaway** is that single-TPP fits do not support coefficient-level interpretation - along $D = kN$ rays only the combined effect of N and D is well constrained, so claims such as "model size contributes X% and data Y% of the reducible loss" require training to span a two-dimensional region of (N, D); when the design has one ray, the reduced parameterization (Definition 3) is the appropriate reporting target. The stakes are practical: on Wikipedia the seed-0 final-epoch CO and NC Chinchilla fits imply compute-optimal sizes $N^* \propto C^{0.36}$ and $N^* \propto C^{0.68}$ (Table 46; the NC fit's E is at its bound 0), so an allocation derived from collinear-design coefficients reflects the design, not the data.

What each claim rests on. *Scale-independent:* for a well-specified law, Propositions 1-4 (the first at leading order), Corollary 1 and Theorem 1 have size-independent rates. *Validated here:* NC wins 97.3% at 5M-77M (Table 5); NC is ahead of budget-matched CO subsets of 1 to 12 rays on web and reference corpora (Table 7); and the margin effect holds pooled over five public releases (Appendix D). *Next steps:* the gap at frontier scale, on other architectures and languages, and in downstream and cross-domain extrapolation; and, on public releases, at 1B and above with $\kappa_{\text{target}} = 100$.

A design recipe (Appendix B.15), with β_{eff} from published exponents (Tables 3 and 9):

1. pick the lower ratio k_1 and read the minimum spread R at $\kappa_{\text{target}} = 100$ from Table 10;
2. train on $K = 2$ ratios at the two ends of that spread;
3. split the runs evenly between them, log-uniform in N;
4. check the exact $\kappa_{A,B}$ of the planned (N, D) before training.

For example, 20 runs split evenly between $k \in \{20, 400\}$ pass Eq. (2) at $\beta_{\text{eff}} = 0.28$ ($V_2/\tau_2 = 1.22$); $\{20, 100\}$ fail (0.48), though their exact $\kappa_{A,B}$ on our ten training sizes is 80, so the exact check decides. Appendix B.15 works this 20-run design through step by step.

Limitations. Theorem 1 proves the ordering for one ray; for $K \geq 2$, our 12-ray fan included, we establish it empirically (Lemma 3 gives a sufficient margin). Our models span 5.04-76.5M parameters. The rates and the leading-order threshold have no size term, but the exact $\kappa_{A,B}$ that decides a design depends on absolute size (Appendix B.12), and so does the measured gap: a bias shared by both designs can shrink the CO/NC error ratio (86) toward 1. At 1B and above, the public releases we analyze (Section 4) have subsets that are not single rays and, on the five fully trained releases, none passes the threshold at $\kappa_{\text{target}} = 100$, so a budget-matched comparison at that scale is the natural next test. Corollary 1 assumes i.i.d. Gaussian residuals; correlated noise with covariance Σ moves its ratio by at most the square root of $\kappa(\Sigma)$, not its rate (Appendix B.11). We test one LLaMA-style architecture, English text, one recipe applied identically to CO and NC, and pre-training loss, none of which the theorems assume (Appendix E); downstream-task and cross-domain extensions (Bhagia et al., 2025) inherit the Jacobian argument; validating them is future work. On Cosmopedia and peS2o the full-fan comparison holds (94.7% and 99.3%; Table 5) and Kaplan keeps 71-72% in the budget-matched one, while Chinchilla and Droppo-Elibol sit near parity there (Table 7). The recipe follows from Proposition 1; training an isoFLOP layout at its spread is the direct next experiment.

Safeguards. All corpora (C4, RedPajama, Wikipedia, Cosmopedia, peS2o) are public and pre-filtered; no new data collection. Released checkpoints (5.04-76.5M; Appendix E) are diagnostic data points, not deployable models, and the code targets scaling-law fitting and small-scale training runs, not serving production models. The paper's caution against single-TPP coefficient claims offsets any marginal acceleration from better-identified scaling laws.

Broader impact. Scaling laws guide multi-billion-dollar pre-training decisions and inform claims about the relative importance of capacity vs. data volume; under single-TPP designs those claims are practically unidentifiable (Section 3, Corollary 1). Our diagnostics (Proposition 1, Eq. (2)) flag such fits, and the framework shows when a replication gap can follow from ill-conditioning alone: on one Chinchilla ray, A's computed CI is of order 10^4 times ψ's on our grid (Appendix B.11). Because the degeneracy is purely Jacobian-geometric, the analysis transfers to generative vision (Henighan et al., 2020; Alabdulmohsin et al., 2022), time-series (Edwards et al., 2024; Yao et al., 2025), and scientific machine learning; isoFLOP grids vary the ratio within each budget and can pass the threshold (Hoffmann et al.'s does; Appendix C).

LLM usage. All research questions, results, design, and analyses originated with the authors. The primary non-trivial use of LLMs was as a sounding board for candidate proof steps: the authors drafted the structure of each derivation, asked an LLM to propose intermediate algebraic manipulations or identities, and then independently verified, rewrote, or discarded each suggestion before incorporation. Auxiliary uses included exposition, LaTeX typesetting, boilerplate code, and prose editing, all reviewed by the authors. LLMs did not generate any claims, theorems, or experimental conclusions. We used Anthropic's Claude Opus 4.7, 5 and 5.5 and the Cursor coding agent.

Acknowledgments and Disclosure of Funding

This research was funded by ARO grant W911NF-24-1-0007. Separately, this work used DeltaAI at NCSA through allocation CIS230318 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program (Boerner et al., 2023), which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Three AMD EPYC CPU chips, donated by AMD, were used for the experimental section of this paper. This research was supported in part through computational resources provided by Syracuse University, including OrangeGrid, which is supported by NSF award ACI-1341006. The authors declare no competing interests.

References

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 265–279, 2023. URL https://proceedings.mlr.press/v202/aghajanyan23a.html.

Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In *Advances in Neural Information Processing Systems*, volume 35, pages 22300–22312, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/8c22e5e918198702765ecff4b20d0a90-Abstract-Conference.html.

A. C. Atkinson, A. N. Donev, and R. D. Tobias. *Optimum Experimental Designs, With SAS*. Oxford Statistical Science Series. Oxford University Press, 2007. ISBN 9780191537943. URL https://books.google.com/books?id=4hJREAAAQBAJ.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27), 6 2024. ISSN 1091-6490. doi: 10.1073/pnas.2311878121. URL http://dx.doi.org/10.1073/pnas.2311878121.

Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons, 1988.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Cosmopedia, February 2024. URL https://huggingface.co/datasets/HuggingFaceTB/cosmopedia. Hugging Face dataset.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in llm pre-training. In *Advances in Neural Information Processing Systems*, volume 38, pages 125153–125188, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/b5f78a17a94da3e34c935515d1b6adae-Abstract-Conference.html.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024. URL https://arxiv.org/abs/2404.10102.

Akshita Bhagia, Jiacheng Liu, Alexander Wettig, David Heineman, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, and Hannaneh Hajishirzi. Establishing task scaling laws via compute-efficient model ladders. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=FeAM2RVO8l.

Åke Björck. *Numerical Methods for Least Squares Problems*. Society for Industrial and Applied Mathematics (SIAM), 1996. ISBN 9780898713602.

Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. ACCESS: Advancing innovation: NSF's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing (PEARC '23)*, pages 173–176, Portland, OR, USA, 2023. ACM. doi: 10.1145/3569951.3597559.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sckjveqlCZ.

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 10683–10699, 2025. URL https://proceedings.mlr.press/v267/choshen25a.html.

Nolan Dey, Gurpreet Gosal, Zhiming Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023. URL https://arxiv.org/abs/2304.03208.

Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell, and Sven Lautenbach. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013. doi: https://doi.org/10.1111/j.1600-0587.2012.07348.x. URL https://nsojournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0587.2012.07348.x.

Jasha Droppo and Oguz Elibol. Scaling laws for acoustic models. In *Proc. Interspeech 2021*, pages 2576–2580, 2021. doi: 10.21437/Interspeech.2021-1644. URL https://www.isca-archive.org/interspeech_2021/droppo21_interspeech.html.

Thomas D. P. Edwards, James Alvey, Justin Alsing, Nam H. Nguyen, and Benjamin D. Wandelt. Scaling-laws for large time-series models. *arXiv preprint arXiv:2405.13867*, 2024. URL https://arxiv.org/abs/2405.13867.

Elias Frantar, Carlos Riquelme Ruiz, Neil Houlsby, Dan Alistarh, and Utku Evci. Scaling laws for sparsely-connected foundation models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=i9K2ZWkYIP.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Luca Soldaini, Jenia Jitsev, Alexandros G. Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. In *The Thirteenth International Conference on Learning Representations*, pages 67661–67682, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a91869936a63d814971b6423990ecf6e-Abstract-Conference.html.

Mitchell A. Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, 2021. doi: 10.18653/v1/2021.emnlp-main.478. URL https://aclanthology.org/2021.emnlp-main.478/.

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. *arXiv preprint arXiv:2010.14701*, 2020. URL https://arxiv.org/abs/2010.14701.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

Michael Y. Hu, Jane Pan, Ayush Rajesh Jhaveri, Nicholas Lourie, and Kyunghyun Cho. Neural neural scaling laws. In *Advances in Neural Information Processing Systems*, 2026. URL https://neurips.cc/virtual/2026/poster/154166.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3X2L2TFr0f.

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. URL https://arxiv.org/abs/2001.08361.

Jong Hae Kim. Multicollinearity and Misleading Statistical Results. *Korean Journal of Anesthesiology*, 72(6):558–569, 2019. doi: 10.4097/kja.19087.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. In *The Fourteenth International Conference on Learning Representations*, pages 74596–74636, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79682ef54677a2cb3dfb8383b372037f-Abstract-Conference.html.

Joshua Shay Kricheli, Khoa Vo, Aniruddha Datta, Spencer Ozgur, and Paulo Shakarian. Error detection and constraint recovery in hierarchical multi-label classification without prior knowledge. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3842–3846. ACM, 2024. doi: 10.1145/3627673.3679918.

Tanishq Kumar, Zachary Ankner, Benjamin F. Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wg1PCg3CUP.

Mario Leiva, Noel Ngu, Joshua Shay Kricheli, Aditya Taparia, Ransalu Senanayake, Paulo Shakarian, Nathaniel D. Bastian, John Corcoran, and Gerardo Simari. Consistency-based abductive reasoning over perceptual errors of multiple pre-trained models in novel environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19216–19223, 2026. doi: 10.1609/aaai.v40i23.38996.

Houyi Li, Wenzhen Zheng, Qiufeng Wang, Zhenyu Ding, Haoying Wang, Zili Wang, Shijie Xuyang, Ning Ding, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. Predictable scale (part ii) - farseer: A refined scaling law in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pages 120612–120654, 2025a. URL https://papers.nips.cc/paper_files/paper/2025/hash/ae9f1d3b89a81fcb1c45b4fc1afbd4fe-Abstract-Conference.html.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (Mis)Fitting scaling laws: A survey of scaling law fitting techniques in deep learning. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7f77492bb8070a5c825a87c0c5181da2-Abstract-Conference.html.

Jan Ludziejewski, Jakub Krajewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, Marek Cygan, and Sebastian Jaszczur. Scaling laws for fine-grained mixture of experts. In *Proceedings of the 41st International Conference on Machine Learning*, pages 33270–33288, 2024. URL https://proceedings.mlr.press/v235/ludziejewski24a.html.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. Datadecide: How to predict best pretraining data with small experiments. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 42487–42502, 2025. URL https://proceedings.mlr.press/v267/magnusson25a.html.

Sean McLeish, John Kirchenbauer, David Yu Miller, Siddharth Singh, Abhinav Bhatele, Micah Goldblum, Ashwinee Panda, and Tom Goldstein. Gemstones: A model suite for multi-faceted scaling laws. In *Advances in Neural Information Processing Systems*, volume 38, pages 123419–123454, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/b2b781badeeb49896c4b324c466ec442-Abstract-Conference.html.

Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 5th edition, 2012.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *Journal of Machine Learning Research*, 26(53):1–66, 2025. URL https://jmlr.org/papers/v26/24-1000.html.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition, 2006. ISBN 9780387400655.

Robert M. O'Brien. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, 41(5):673–690, 10 2007. doi: 10.1007/s11135-006-9018-6.

OpenAI. tiktoken: a fast BPE tokeniser for use with OpenAI's models. https://github.com/openai/tiktoken, 2022.

Andrei Panferov, Alexandra Volkova, Ionut-Vlad Modoranu, Vage Egiazarian, Mher Safaryan, and Dan Alistarh. Unified scaling laws for compressed representations. In *Advances in Neural Information Processing Systems*, volume 38, pages 151086–151112, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/ddd7eae51dcd8bbfc936e5b19433d339-Abstract-Conference.html.

Michael Poli, Armin W. Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, Ce Zhang, and Stefano Massaroli. Mechanistic design and scaling of hybrid architectures. In *Proceedings of the 41st International Conference on Machine Learning*, pages 40908–40950, 2024. URL https://proceedings.mlr.press/v235/poli24a.html.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 100535–100570, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/b6341525cd84f3be0ef203e4d7cd8556-Abstract-Conference.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *The Eighth International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryenvpEKDr.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43445–43460, 2024. URL https://proceedings.mlr.press/v235/sardana24a.html.

Rylan Schaeffer, Noam Levi, Brando Miranda, and Sanmi Koyejo. Pretraining scaling laws for generative evaluations of language models. In *The Fourteenth International Conference on Learning Representations*, pages 52380–52422, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/55d7fbeddfd9bb6dace146cc4fa02ff0-Abstract-Conference.html.

Anastasiia Sedova, Skyler Seto, Natalie Schluter, and Pierre Ablin. Scaling laws for mixture pretraining under data constraints. *arXiv preprint arXiv:2605.12715*, 2026. URL https://arxiv.org/abs/2605.12715.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. In *Advances in Neural Information Processing Systems*, volume 38, pages 129554–129579, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/bc1d640f841f752c689aae20b31198c1-Abstract-Conference.html.

Luca Soldaini and Kyle Lo. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI, 2023. ODC-By, https://github.com/allenai/pes2o.

Xingwu Sun, Shuaipeng Li, Ruobing Xie, Weidong Han, Kan Wu, Zhen Yang, Yixing Li, An Wang, Shuai Li, Jinbao Xue, Yu Cheng, Yangyu Tao, Zhanhui Kang, Cheng-Zhong Xu, Di Wang, and Jie Jiang. Scaling laws for floating-point quantization training. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 57544–57570, 2025. URL https://proceedings.mlr.press/v267/sun25j.html.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *Advances in Neural Information Processing Systems*, volume 37, pages 114147–114179, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/cf5a019ae9c11b4be88213ce3f85d85c-Abstract-Conference.html.

Aditya Taparia, Noel Ngu, Mario Leiva, Joshua Shay Kricheli, John Corcoran, Nathaniel D. Bastian, Gerardo Simari, Paulo Shakarian, and Ransalu Senanayake. VLC fusion: Vision-language conditioned sensor fusion for robust object detection. In *Machine Learning and Knowledge Discovery in Databases. Research Track and Applied Data Science Track (ECML PKDD 2026)*, Lecture Notes in Computer Science, pages 51–68. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37679-4_4.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Why are nonlinear fits to data so challenging? *Phys. Rev. Lett.*, 104:060201, Feb 2010. doi: 10.1103/PhysRevLett.104.060201. URL https://link.aps.org/doi/10.1103/PhysRevLett.104.060201.

Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna. Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond. *The Journal of Chemical Physics*, 143:010901, 2015. doi: 10.1063/1.4923066.

Kristina P. Vatcheva, MinJae Lee, Joseph B. McCormick, and Mohammad H. Rahbar. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology: Open Access*, 6 (2):227, 2016. doi: 10.4172/2161-1165.1000227.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of LLM scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 49523–49544, 2024. URL https://proceedings.mlr.press/v235/villalobos24a.html.

Alexandra Volkova, Mher Safaryan, Christoph H. Lampert, and Dan Alistarh. Towards robust scaling laws for optimizers. *arXiv preprint arXiv:2602.07712*, 2026. URL https://arxiv.org/abs/2602.07712.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 116462–116492, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/d34497330b1fd6530f7afd86d0df9f76-Abstract-Datasets_and_Benchmarks_Track.html.

Wikimedia Foundation. Wikimedia downloads, 2023. URL https://dumps.wikimedia.org. English Wikipedia dump of 2023-11-01, used through the Hugging Face dataset wikimedia/wikipedia, subset 20231101.en.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.767. URL https://aclanthology.org/2023.acl-long.767/.

Zhiwei Xu, Shihao Wu, Hanseul Cho, Wei Hu, and Yixin Wang. Data-constrained language model pre-training: Improved regularization and scaling laws. In *Advances in Neural Information Processing Systems*, 2026. URL https://neurips.cc/virtual/2026/poster/149999.

Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. Towards neural scaling laws for time series foundation models. In *The Thirteenth International Conference on Learning Representations*, pages 83793–83820, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d04e47d0fdca09e898885c66b67b1e95-Abstract-Conference.html.

Baoqing Yue, Jinyuan Zhou, Zixi Wei, Jingtao Zhan, Qingyao Ai, and Yiqun Liu. Relative-based scaling law for neural language models. *arXiv preprint arXiv:2510.20387*, 2025. URL https://arxiv.org/abs/2510.20387.

Hanlin Zhang, Jikai Jin, Vasilis Syrgkanis, and Sham Kakade. Prescriptive scaling reveals the evolution of language model capabilities. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 156084–156114, 2026. URL https://proceedings.mlr.press/v306/zhang26by.html.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/abs/2205.01068.

A Appendix - notation reference

Table 8 collects all symbols used throughout the paper and appendices.

Table 8: Summary of notation.

Symbol	Meaning	Defined in
Scaling-law variables		
N	Model parameter count	Eq. (1)
D	Training tokens (dataset size)	Eq. (1)
$L(N, D)$	Observed loss	Eq. (1)
$\hat{L}(N, D; \boldsymbol{\theta})$	Predicted loss (scaling law)	Def. 1
$k_1 < \cdots < k_K$	Ordered TPP ratios in a K-ray design	Def. 1
k	Tokens-per-parameter (TPP) ratio ($K = 1$ case, $D = kN$)	Sec. 3.2
C	Training compute, $C = 6ND$ (Chinchilla FLOPs)	Sec. 3.4
C_i	Induced compute budget $6N_iD_i$ for holdout point i	Def. 4
$N_{\min}$, $N_{\max}$	Min/max model sizes used for induced isoFLOP curves	Def. 4
V	Vocabulary size (cl100k_base; empirical count 100,277)	Sec. 4; App. E
Chinchilla parameters		
A, B	Scale coefficients (model, data)	Eq. (1)
α, β	Power-law exponents	Eq. (1)
E	Irreducible loss	Eq. (1)
ψ	Combined coefficient $A + Bk^{-\alpha}$	Eq. (3)
ε	Exponent gap ($\lvert\alpha-\beta\rvert$ for Chinchilla)	Def. 2
Observation model and optimization		
$\mathcal{A}$	Training scheme (architecture, optimizer, algorithm, etc.)	Def. 1
$\mathcal{D}$	Scaling-law dataset $\{(N_i, D_i, L_i)\}_{i=1}^m$	Def. 1
$\mathcal{D}_{\text{train}}$	Training split of $\mathcal{D}$	Def. 1
$\mathcal{D}_{\text{train}}^{\text{CO}}$, $\mathcal{D}_{\text{train}}^{\text{NC}}$	Full collinear and non-collinear training pools before subset enumeration	Def. 10, Def. 11
$\mathcal{N}$	Shared ordered model sizes $\{N_1, \ldots, N_n\}$	Def. 1; App. F.3
m	Number of observations in $\mathcal{D}$	Def. 1
$\boldsymbol{\theta}$	Scaling-law parameter vector	Def. 1
p	Number of parameters in $\boldsymbol{\theta}$	Def. 1
n	Number of distinct model sizes	Def. 1
$N_1 < \cdots < N_n$	Ordered model sizes	Def. 1
K	Number of TPP ratios in collinear design	Def. 1
$\kappa(\cdot)$	Condition number ($\lambda_{\max}/\lambda_{\min}$)	Sec. 3.1
r_i	Residual: $L_i - \hat{L}(N_i, D_i; \boldsymbol{\theta})$	Sec. 3.1
$\mathbf{r}(\boldsymbol{\theta})$	Stacked residual vector $(r_i)_{i=1}^m$	Sec. 3.1
$S(\boldsymbol{\theta})$	GN objective $\frac{1}{2}\lVert\mathbf{r}(\boldsymbol{\theta})\rVert^2$	Sec. 3.1
J	Jacobian of residuals	Sec. 3.1
$\Delta\boldsymbol{\theta}$	Gauss-Newton update step	Sec. 3.1
$\kappa_{A,B}$	Condition number of (A, B) sub-block of J^TJ; a lower bound for $\kappa(J^TJ)$ (Cauchy interlacing)	Sec. 3.1
$\mathbf{j}_a$, $\mathbf{j}_b$	Columns of J (near-proportional pair)	Proposition 2
c	Proportionality scalar, $\mathbf{j}_b = c\,\mathbf{j}_a + \boldsymbol{\delta}$	Proposition 2
$\boldsymbol{\delta}$	Near-proportionality perturbation, $\lVert\boldsymbol{\delta}\rVert = O(\varepsilon)$	Proposition 2
$\mathbf{j}_1$, $\mathbf{j}_2$	First two Jacobian columns in the column-ordering used in Table 2 (alias of the illustrative (a, b) pair in Prop. 2)	Tab. 2
$\hat{\boldsymbol{\theta}}$	Least-squares estimator	Lemma 1
$\mathbb{E}[\cdot]$	Expectation (Regime A/B inequalities and RMSE ratios)	Thm. 1
σ^2	Idealized homoscedastic variance of observational noise in least squares (distinct from appendix-table σ)	Lemma 1; Cor. 1
σ (appendix tables)	Seed-to-seed std. dev. of Monte Carlo summaries across optimizer seeds	App. F; App. G
$\text{CI}_{0.95}(\cdot)$	Nominal 95% confidence interval half-width	Corollary 1
Multi-ray designs and holdout prediction (Prop. 1; Theorem 1)		
$\mathcal{H}$	General holdout set	Def. 1
$n_{\mathcal{H}}$	Holdout cardinality	Def. 1

Continued on next page

(Table 8 continued)

Symbol	Meaning	Defined in
k_K, k_1	Largest/smallest ordered TPP ratio in a design	Prop. 1
R	Two-ray TPP spread k_2/k_1 ($K = 2$)	Prop. 1
β_{eff}	Law-specific data-size exponent (Chinchilla/repeated-data: β; Kaplan: 1; Droppo-Elibol: γ_D)	Prop. 1; Tab. 9
V_K	Second central moment (variance) of $\{k_\ell^{-\beta_{\text{eff}}}\}_{\ell=1}^K$ across TPP rays; $V_1 = 0$	Prop. 1
τ_K	Diversity threshold in (2); at leading order $V_K \geq \tau_K$ implies $\kappa_{A,B} \leq \kappa_{\text{target}}$, and the converse holds up to the factor $(1 + 1/\kappa_{\text{target}})^{-2}$	Prop. 1
κ_K	Shorthand for $\kappa_{A,B}$ on a collinear K-ray design ($V_K \geq \tau_K$ gives $\kappa_K \leq \kappa_{\text{target}}$ at leading order)	App. B.15
κ_{target}	Target condition number; τ_K in (2) depends on it	Prop. 1; Thm. 1
$\text{RMSE}_{\mathcal{H}}^{\text{CO/NC}}$	Holdout RMSE of $\hat{L}$ against $\mathbb{E}[L]$ (superscript: training design)	Thm. 1
$R^{2\,\text{CO/NC}}_{\mathcal{H}}$	Holdout R^2 under design CO or NC	Thm. 1
$G_{A,B}$, M	Gram block of the CO scale-coefficient columns; the holdout's second moment of the same two Jacobian entries	Sec. 3.1; Lemma 3
Experimental design (Sec. 4)		
$\mathcal{K}_{\text{train}}$	Full set of 12 collinear training TPP ratios used in the CO design	Sec. 4
$\mathcal{K}_{\text{test}}$	Set of 5 collinear holdout TPP ratios beyond the training fan	Sec. 4
$\mathcal{K}_{\text{big}}$	Set of 6 collinear training TPP ratios in the preliminary high-TPP CO design	Sec. 4; App. F.7
$\mathcal{H}_{\text{col}}$	Collinear holdout (empirical label)	Sec. 4
$\mathcal{H}_{\text{nc}}$	Non-collinear holdout (empirical label)	Sec. 4
$R^2_{\mathcal{H}}$	Unified-holdout R^2 on $\mathcal{H}$	Sec. 4; Tab. 4
$R^2_{\mathcal{H}_{\text{col}}}$, $R^2_{\mathcal{H}_{\text{nc}}}$	R^2 on CO-only vs. NC-only holdout splits	App. F
R^2_{train}	Coefficient of determination on the training split	App. H
d_{model}, d_{ff}, d_{head}, n_{layers}, n_{heads}	Transformer shape hyperparameters (Table 17)	App. E
RMSE	Root mean squared error on a holdout split	Sec. 4
Alternative scaling-law notation		
N', D'	Effective quantities (repeated-data)	Def. 2
N_c, D_c	Scale parameters (Kaplan)	Tab. 2; App. B.8
α_N, α_D	Separate exponents (Kaplan / Droppo-Elibol)	Def. 2
N_C, D_C	Scale parameters (Droppo-Elibol)	Tab. 2; App. B.9
α (Droppo-Elibol)	Canonical composite exponent in that law	Def. 2; Tab. 2
γ_N, γ_D	Exponent ratios α_N/α, α_D/α (Droppo-Elibol)	Def. 2
L_∞	Irreducible loss (Droppo-Elibol)	App. B.9
R_D^*, R_N^*	Decay constants (repeated-data)	App. B.7
Appendix-only: design-dependent quantities		
Φ_q	Power-sum family $\sum_i N_i^{-q\alpha}$ (index q)	App. B.5; Def. 5
$\Phi_2^{(2\varepsilon)}$	$\sum_i N_i^{-2\beta}$ (Chinchilla; $= \sum_i N_i^{-2\alpha+2\tilde{\varepsilon}}$, $\tilde{\varepsilon} = \alpha - \beta$)	App. B.11
T_q	Log-weighted power sum (e.g. $T_2 = \sum_i N_i^{-2\alpha} \log N_i$)	App. B.11
U_q	Log^2-weighted power sum (e.g. $U_2 = \sum_i N_i^{-2\alpha} (\log N_i)^2$)	App. B.11
w_i	Power-law weights $N_i^{-2\alpha}/\Phi_2$	App. B.10
$\sigma_w^2(\log N)$	Weighted log-variance of model sizes	App. B.10
Appendix-only: experimental design		
$R_{\min}(k_1)$	Smallest two-ray spread $R = k_2/k_1$ with $V_2 \geq \tau_2$ at lower ratio k_1, in closed form (71); ∞ when no spread suffices at leading order	App. B.12; Tab. 10
κ_1	Single-ray baseline κ_K at $K = 1$, $\Theta(\varepsilon^{-2})$ for every k_1	App. B.15
$\Lambda(\mathcal{S})$	Sloppy leverage of test set $\mathcal{S}$	App. B.13, Def. 8
$V_{\mathcal{H}}^{\text{X}}$	Mean predictive variance of design X on $\mathcal{H}$, $\mathbb{E}[\text{MSE}_{\mathcal{H}}^{\text{X}}] = \sigma^2 V_{\mathcal{H}}^{\text{X}}$ (well-specified)	Lemma 3
ρ_K, $\rho^\star(\mathcal{H})$	Conditioning ratio and $K \geq 2$ margin (a ratio above 1 suffices for the MSE ordering, above $\pi/2$ for the RMSE ordering)	Lemma 3
G	Training-design scalar $G = \Phi_2 / \sum_i N_i^{-2\alpha} (\ln N_i)^2 > 0$ in the CO MSE lower bound (79)	Eq. (79)

Continued on next page

(Table 8 continued)

Symbol	Meaning	Defined in
η	Learning rate in optimizer setup	App. E
Appendix-only: subset enumeration (App. F.3)		
$\mathcal{L}$	Scaling-law family (Chinchilla, Kaplan, Droppo-Elibol)	App. F.3
μ	Epoch/mode index in paired-comparison enumeration	Def. 12
$\mathbf{k}^{\mathrm{all}}$	Full set of 12 collinear TPP ratios per dataset	Def. 9
S	Subset of TPP-ratio indices, $S \subseteq \{1, \ldots, 12\}$	Def. 10
$\mathrm{CO}(S)$	Collinear subset design induced by S	Def. 10
n_S	Cardinality of $\mathrm{CO}(S)$	Def. 10
K_S	Effective TPP-ray count of subset design, $K_S = \lvert S \rvert$	Def. 10
$\mathcal{N}_{\mathrm{NC}}$, $\mathcal{D}_{\mathrm{NC}}$	Row/column axes of the NC grid in (N, D) space	Def. 11
M_N, M_D	NC-grid row/column counts, $\lvert\mathcal{N}_{\mathrm{NC}}\rvert$, $\lvert\mathcal{D}_{\mathrm{NC}}\rvert$	Def. 11
$n^\star$, $\mathbf{k}_S$	Target NC cardinality and target TPP ratios from S	Def. 11
$\mathrm{NC}_{\square}(n^\star, \mathbf{k}_S)$	Bounding-box NC design of cardinality $n^\star$	Def. 11
$\mathcal{C}(\mathcal{L}, \mathcal{D}, \mu, S)$	Paired comparison tuple	Def. 12
$\mathcal{P}$	Set of all paired comparisons	Def. 12
$\kappa^\star$	Target condition number for empirical regime classification	Def. 13
$\mathcal{P}^A_{\mathcal{L},\mu}(\kappa^\star)$	Predicted below-threshold set at target $\kappa^\star$ ($V_K < \tau_K$; Kaplan: exact $\kappa_{A,B} > \kappa^\star$)	Def. 13
$\mathrm{WR}(\mathcal{P}')$	NC win rate on a collection $\mathcal{P}' \subseteq \mathcal{P}$	Def. 13
$\mathrm{WR}_A(\mathcal{L}, \mu; \kappa^\star)$	Below-threshold NC win rate at target $\kappa^\star$	Def. 13
$\mathbf{1}\{\cdot\}$	Indicator used in NC win-rate definition	Eq. (107)

B Appendix - proofs and detailed derivations

The following proofs and derivations support the results in Section 3. For Kaplan, the proofs use the compositional form that the experiments fit (Appendix B.8).

B.1 Appendix - Derivation of the Gauss-Newton normal equations

This subsection provides a self-contained derivation of the Gauss-Newton normal equations from the nonlinear least-squares objective (Section 3.1), following the treatment in (Nocedal and Wright, 2006, Chap. 10).

Step 1: Objective and gradient. Recall the objective:

$$S(\boldsymbol{\theta}) = \tfrac{1}{2}\,\|\mathbf{r}(\boldsymbol{\theta})\|^2 = \tfrac{1}{2}\sum_{i=1}^{m} r_i(\boldsymbol{\theta})^2, \tag{8}$$

where $r_i(\boldsymbol{\theta}) = L_i - \hat{L}(N_i, D_i; \boldsymbol{\theta})$. The gradient with respect to $\boldsymbol{\theta}$ is obtained by the chain rule:

$$\frac{\partial S}{\partial \theta_j} = \sum_{i=1}^{m} r_i(\boldsymbol{\theta})\,\frac{\partial r_i}{\partial \theta_j} = -\sum_{i=1}^{m} r_i(\boldsymbol{\theta})\,\frac{\partial \hat{L}_i}{\partial \theta_j}, \tag{9}$$

since $\partial r_i/\partial\theta_j = -\partial\hat{L}_i/\partial\theta_j$. Define the residual Jacobian:

$$J \in \mathbb{R}^{m\times p}, \qquad J_{ij} \coloneqq -\frac{\partial \hat{L}(N_i, D_i; \boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial r_i}{\partial \theta_j}. \tag{10}$$

Thus $J_{ij} = \partial r_i/\partial\theta_j$, matching Section 3.1. With this convention, the gradient (9) can be written in matrix form as:

$$\nabla_{\boldsymbol{\theta}} S = J^T \mathbf{r}(\boldsymbol{\theta}). \tag{11}$$

Step 2: Linearization of the residuals. At a current iterate $\boldsymbol{\theta}_k$, expand each residual to first order in the step $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}_k$:

$$r_i(\boldsymbol{\theta}_k + \Delta\boldsymbol{\theta}) \approx r_i(\boldsymbol{\theta}_k) + \sum_{j=1}^{p} \left.\frac{\partial r_i}{\partial \theta_j}\right|_{\boldsymbol{\theta}_k} \Delta\theta_j = r_{i,k} + (J_k\,\Delta\boldsymbol{\theta})_i, \tag{12}$$

where $r_{i,k} \coloneqq r_i(\boldsymbol{\theta}_k)$ and $J_k \coloneqq J(\boldsymbol{\theta}_k)$ is the Jacobian evaluated at $\boldsymbol{\theta}_k$. In vector notation:

$$\mathbf{r}(\boldsymbol{\theta}_k + \Delta\boldsymbol{\theta}) \approx \mathbf{r}_k + J_k\,\Delta\boldsymbol{\theta}. \tag{13}$$

Step 3: Substitute into the objective. Substituting (13) into (8):

$$\begin{aligned} S(\boldsymbol{\theta}_k + \Delta\boldsymbol{\theta}) &\approx \tfrac{1}{2}\,\|\mathbf{r}_k + J_k\,\Delta\boldsymbol{\theta}\|^2 \\ &= \tfrac{1}{2}\,\big(\mathbf{r}_k + J_k\,\Delta\boldsymbol{\theta}\big)^T\big(\mathbf{r}_k + J_k\,\Delta\boldsymbol{\theta}\big) \\ &= \tfrac{1}{2}\,\mathbf{r}_k^T\mathbf{r}_k + \mathbf{r}_k^T J_k\,\Delta\boldsymbol{\theta} + \tfrac{1}{2}\,\Delta\boldsymbol{\theta}^T J_k^T J_k\,\Delta\boldsymbol{\theta}. \end{aligned}$$

This is an *exact* quadratic in $\Delta\boldsymbol{\theta}$ (no higher-order terms remain because the linearization has already been applied). Denote this quadratic model as $m_k(\Delta\boldsymbol{\theta})$.

Step 4: Stationarity condition. At the minimizer of m_k, the gradient with respect to $\Delta\boldsymbol{\theta}$ vanishes:

$$\nabla_{\Delta\boldsymbol{\theta}}\, m_k = J_k^T\mathbf{r}_k + J_k^T J_k\,\Delta\boldsymbol{\theta} = \mathbf{0}, \tag{14}$$

$$(J_k^T J_k)\,\Delta\boldsymbol{\theta} = -J_k^T\mathbf{r}_k. \tag{15}$$

This is the Gauss-Newton normal equation. When $J_k^T J_k$ is non-singular, the unique solution is $\Delta\boldsymbol{\theta} = -(J_k^T J_k)^{-1} J_k^T\mathbf{r}_k$, and the parameter update is $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \Delta\boldsymbol{\theta}$.

Remark 1 (Relationship to the full Hessian)**.** *The exact Hessian of S is:*

$$H(\boldsymbol{\theta}) = \nabla^2 S = J^T J + Q, \qquad Q \coloneqq \sum_{i=1}^{m} r_i \, \nabla^2 r_i. \tag{16}$$

The Gauss-Newton method drops the second-order correction Q, replacing H with the Gauss-Newton approximation *$J^T J$. This is justified when $\|\mathbf{r}\|$ is small (so $Q \approx 0$) or when the model is approximately linear in $\boldsymbol{\theta}$ (so $\nabla^2 r_i \approx 0$). For the Chinchilla and repeated-data scaling laws, $\hat{L}$ is* linear *in the scale coefficients A and B, so $\nabla^2_{A,B} r_i = 0$ exactly and $Q_{A,B} = 0$: the Gauss-Newton and full Newton curvatures* coincide *in the (A, B) sub-block regardless of residual magnitudes.*

Remark 2 (Why $J^T J$ controls identifiability)**.** *When $J^T J$ is well conditioned, the normal equations (15) have a unique, numerically stable solution. When two columns of J are nearly proportional, $J^T J$ becomes nearly singular:*

$$\lambda_{\min}(J^T J) \to 0 \quad \Longrightarrow \quad \|(J^T J)^{-1}\| \to \infty, \tag{17}$$

and the step $\Delta\boldsymbol{\theta}$ becomes unbounded in the direction of the near-null eigenvector. Under i.i.d. noise with variance σ^2, one has

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 \, (J^T J)^{-1} \tag{18}$$

in the linear least-squares case (Lemma 1), and the same expression approximates the large-sample covariance for nonlinear least squares (Remark 3); thus the parameter variances along the near-null direction scale as $\sigma^2/\lambda_{\min}(J^T J)$. This is the statistical manifestation of the ill-conditioning: any parameter combination aligned with the sloppy direction of $J^T J$ is effectively unconstrained by the data. Section 3 shows that, on each single collinear ray $D = k_\ell N$, the scale-coefficient columns of J are nearly proportional whenever $\varepsilon \ll 1$ (the law-specific exponent gap of Definition 2; e.g., $|\alpha - \beta|$ for Chinchilla), driving $\lambda_{\min}(J^T J) = O(\varepsilon^2)$, hence $\kappa(J^T J) = \Omega(\varepsilon^{-2})$ and $\kappa_{A,B} = \Theta(\varepsilon^{-2})$. With $K > 1$ rays, the spread of the ratios separates the pair: $V_K \geq \tau_K$ in (2) bounds $\kappa_{A,B}$ by κ_{target} at leading order (Proposition 1). For $K = 2$, $V_2 \geq \tau_2$ holds iff $R = k_2/k_1 \geq R_{\min}(k_1)$, which increases with $k_1 \geq 1$ (Eq. (71), Table 10), a scale effect of the B column only.

B.2 Appendix - Covariance under homoscedastic i.i.d. noise

Lemma 1 (Parameter covariance in linear least squares)**.** *For any $\mathbf{M} \in \mathbb{R}^{m \times p}$ of full column rank and $\boldsymbol{\theta}^\star \in \mathbb{R}^p$, suppose*

$$L_i = (\mathbf{M}\boldsymbol{\theta}^\star)_i + \epsilon_i, \qquad i = 1, \ldots, m,$$

where $\{\epsilon_i\}_{i=1}^m$ are i.i.d. with mean zero and variance σ^2. Let $\hat{\boldsymbol{\theta}}$ minimize $\sum_{i=1}^m \big(L_i - (\mathbf{M}\boldsymbol{\theta})_i\big)^2$, and let $J \in \mathbb{R}^{m \times p}$ be the Jacobian of the residuals $r_i(\boldsymbol{\theta}) \coloneqq L_i - (\mathbf{M}\boldsymbol{\theta})_i$ with respect to $\boldsymbol{\theta}$ (holding $\mathbf{M}$ fixed). Then $J = -\mathbf{M}$ and

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 \, (J^T J)^{-1}. \tag{19}$$

Proof. Differentiating $r_i(\boldsymbol{\theta}) = L_i - \sum_{j=1}^p M_{ij}\, \theta_j$ gives $\partial r_i / \partial \theta_j = -M_{ij}$, hence $J = -\mathbf{M}$ and $J^T J = \mathbf{M}^T \mathbf{M}$. The normal equations yield $\mathbf{M}^T \mathbf{M}\, \hat{\boldsymbol{\theta}} = \mathbf{M}^T \mathbf{L}$ with $\mathbf{L} \coloneqq (L_1, \ldots, L_m)^T$, so

$$\hat{\boldsymbol{\theta}} = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{L}.$$

Substituting $\mathbf{L} = \mathbf{M}\boldsymbol{\theta}^\star + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \coloneqq (\epsilon_1, \ldots, \epsilon_m)^T$:

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \boldsymbol{\epsilon}.$$

Because $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, also $\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}^\star$. Because $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 I_m$,

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\theta}}) &= (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T (\sigma^2 I_m)\, \mathbf{M}\, (\mathbf{M}^T \mathbf{M})^{-1} \\ &= \sigma^2 \, (\mathbf{M}^T \mathbf{M})^{-1} = \sigma^2 \, (J^T J)^{-1}. \end{aligned}$$

□

Remark 3 (Nonlinear least squares)**.** *Lemma 1 applies directly when the predictor $\hat{L}(N_i, D_i; \boldsymbol{\theta})$ is affine in $\boldsymbol{\theta}$ on the design (e.g., fitting (A, B, E) in (1) with (α, β) held fixed). For general nonlinear $\hat{L}$, the same matrix $\sigma^2 (J^T J)^{-1}$ with $J_{ij} = \partial r_i / \partial \theta_j$ evaluated at the true parameter is the standard large-sample approximation to $\mathrm{Cov}(\hat{\boldsymbol{\theta}})$ under regularity conditions for nonlinear least squares (Bates and Watts, 1988).*

B.3 Appendix - Proof of Proposition 2 (Full-matrix conditioning)

Proposition (Proposition 2, restated)**.** *Let $J \in \mathbb{R}^{m\times p}$ have columns $\mathbf{j}_1, \dots, \mathbf{j}_p$ of norm $O(1)$, and suppose two columns $a \neq b$ satisfy $\mathbf{j}_b = c\,\mathbf{j}_a + \boldsymbol{\delta}$ with a fixed $c \neq 0$, $\|\mathbf{j}_a\| = \Theta(1)$ and $\|\boldsymbol{\delta}\| = O(\varepsilon)$, where ε is the law-specific exponent gap (Definition 2). Then as $\varepsilon \to 0$, $\lambda_{\max}(J^TJ) = \Theta(1)$, $\lambda_{\min}(J^TJ) = O(\varepsilon^2)$, and $\kappa(J^TJ) = \Omega(\varepsilon^{-2})$ (4); if the part of $\boldsymbol{\delta}$ orthogonal to $\mathbf{j}_a$ has norm $\Theta(\varepsilon)$, $\kappa((J^TJ)_{a,b}) = \Theta(\varepsilon^{-2})$, which is $\kappa_{A,B}$ when (a, b) is the scale pair.*

We proceed in three steps: (i) show that Proposition 2's assumption $\mathbf{j}_b = c\,\mathbf{j}_a + \boldsymbol{\delta}$ forces $\lambda_{\min}(J_1^TJ_1) = O(\varepsilon^2)$; (ii) use the Rayleigh quotient to transfer this to the full $p \times p$ Gram matrix; (iii) conclude the condition-number bound.

Step 1: $\lambda_{\min}(J_1^TJ_1) = O(\varepsilon^2)$.

By the near proportionality assumption of Proposition 2, $\mathbf{j}_b = c\,\mathbf{j}_a + \boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\| = O(\varepsilon)$. Let

$$J_1 \coloneqq [\,\mathbf{j}_a \mid \mathbf{j}_b\,] \in \mathbb{R}^{m\times 2} \tag{20}$$

hold the two columns. Its 2×2 Gram matrix is:

$$G \coloneqq J_1^TJ_1 = \begin{pmatrix} \|\mathbf{j}_a\|^2 & \mathbf{j}_a^T\mathbf{j}_b \\ \mathbf{j}_b^T\mathbf{j}_a & \|\mathbf{j}_b\|^2 \end{pmatrix}. \tag{21}$$

Substituting $\mathbf{j}_b = c\,\mathbf{j}_a + \boldsymbol{\delta}$:

$$\|\mathbf{j}_b\|^2 = c^2\,\|\mathbf{j}_a\|^2 + 2c\,\mathbf{j}_a^T\boldsymbol{\delta} + \|\boldsymbol{\delta}\|^2, \tag{22}$$

$$\mathbf{j}_a^T\mathbf{j}_b = c\,\|\mathbf{j}_a\|^2 + \mathbf{j}_a^T\boldsymbol{\delta}. \tag{23}$$

Write $s \coloneqq \|\mathbf{j}_a\|^2 > 0$ and $\eta \coloneqq \mathbf{j}_a^T\boldsymbol{\delta}$ (so $|\eta| \le \|\mathbf{j}_a\|\,\|\boldsymbol{\delta}\| = O(\varepsilon)$ by Cauchy-Schwarz). Then:

$$G = \begin{pmatrix} s & cs+\eta \\ cs+\eta & c^2s + 2c\eta + \|\boldsymbol{\delta}\|^2 \end{pmatrix}.$$

The determinant is:

$$\begin{aligned} \det(G) &= s\,(c^2s + 2c\eta + \|\boldsymbol{\delta}\|^2) - (cs+\eta)^2 \\ &= s\,c^2s + 2sc\eta + s\|\boldsymbol{\delta}\|^2 - c^2s^2 - 2cs\eta - \eta^2 \\ &= s\,\|\boldsymbol{\delta}\|^2 - \eta^2. \end{aligned} \tag{24}$$

By the Cauchy-Schwarz inequality applied to $\mathbf{j}_a$ and $\boldsymbol{\delta}$: $\eta^2 = (\mathbf{j}_a^T\boldsymbol{\delta})^2 \le \|\mathbf{j}_a\|^2\,\|\boldsymbol{\delta}\|^2 = s\,\|\boldsymbol{\delta}\|^2$, with equality if and only if $\boldsymbol{\delta}$ is proportional to $\mathbf{j}_a$ (which would make $\mathbf{j}_b$ exactly proportional to $\mathbf{j}_a$ and hence $\det(G) = 0$). In general:

$$0 \;\le\; \det(G) \;=\; s\,\|\boldsymbol{\delta}\|^2 - \eta^2 \;\le\; s\,\|\boldsymbol{\delta}\|^2 \;=\; O(\varepsilon^2). \tag{25}$$

Since G is a 2×2 positive semidefinite matrix with $\mathrm{tr}(G) = s + c^2s + O(\varepsilon) = \Theta(1)$, its eigenvalues $\lambda_+ \ge \lambda_- \ge 0$ satisfy $\lambda_+ + \lambda_- = \mathrm{tr}(G) = \Theta(1)$ and $\lambda_+ \cdot \lambda_- = \det(G) = O(\varepsilon^2)$. Therefore:

$$\lambda_-(G) = \frac{\det(G)}{\lambda_+(G)} = \frac{O(\varepsilon^2)}{\Theta(1)} = O(\varepsilon^2). \tag{26}$$

Let $\mathbf{w} \in \mathbb{R}^2$ be a unit eigenvector of G corresponding to $\lambda_- = \lambda_{\min}(G)$. Then:

$$\mathbf{w}^T G\,\mathbf{w} = \lambda_{\min}(G) = O(\varepsilon^2). \tag{27}$$

Step 2: Rayleigh-quotient transfer to the full system.

Partition $J = [\,J_1 \mid J_2\,]$ where $J_1 \in \mathbb{R}^{m\times 2}$ consists of columns a and b, and $J_2 \in \mathbb{R}^{m\times(p-2)}$ contains all remaining columns. Construct the padded test vector $\tilde{\mathbf{w}} \in \mathbb{R}^p$ by placing the entries of $\mathbf{w}$ in positions a and b:

$$\tilde{w}_a = w_1, \qquad \tilde{w}_b = w_2, \qquad \tilde{w}_j = 0 \text{ otherwise}, \tag{28}$$

so that $\|\tilde{\mathbf{w}}\| = \|\mathbf{w}\| = 1$.

The full Gram matrix has block structure:

$$J^T J = \begin{pmatrix} J_1^T J_1 & J_1^T J_2 \\ J_2^T J_1 & J_2^T J_2 \end{pmatrix}. \tag{29}$$

Evaluating the quadratic form at $\tilde{\mathbf{w}}$:

$$\begin{aligned} \tilde{\mathbf{w}}^T (J^T J)\, \tilde{\mathbf{w}} &= \begin{pmatrix} \mathbf{w} \\ \mathbf{0} \end{pmatrix}^T \begin{pmatrix} J_1^T J_1 & J_1^T J_2 \\ J_2^T J_1 & J_2^T J_2 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{w}^T (J_1^T J_1)\, \mathbf{w} + \underbrace{\mathbf{0}^T (J_2^T J_1)\, \mathbf{w}}_{=0} + \underbrace{\mathbf{w}^T (J_1^T J_2)\, \mathbf{0}}_{=0} + \underbrace{\mathbf{0}^T (J_2^T J_2)\, \mathbf{0}}_{=0} \\ &= \mathbf{w}^T (J_1^T J_1)\, \mathbf{w} = \lambda_{\min}(J_1^T J_1) = O(\varepsilon^2). \end{aligned} \tag{30}$$

By the Rayleigh-Ritz principle, $\lambda_{\min}(M) = \min_{\|\mathbf{v}\|=1} \mathbf{v}^T M \mathbf{v}$, so every unit $\mathbf{v}$ satisfies $\lambda_{\min}(M) \le \mathbf{v}^T M \mathbf{v}$:

$$\lambda_{\min}(J^T J) \le \tilde{\mathbf{w}}^T (J^T J)\, \tilde{\mathbf{w}} = O(\varepsilon^2). \tag{31}$$

Step 3: Condition-number bound.

For $\lambda_{\max}$: it is at least $\lambda_{\max}(J_1^T J_1) \ge \operatorname{tr}(G)/2 = \Theta(1)$ (since G has two eigenvalues summing to $\operatorname{tr}(G) = \Theta(1)$, the larger is at least half the trace), and $\lambda_{\max}(J^T J) \le \operatorname{tr}(J^T J) = \|J\|_F^2 = \Theta(1)$ for any non-degenerate dataset (all columns of J have bounded norms). Therefore $\lambda_{\max}(J^T J) = \Theta(1)$, and:

$$\kappa(J^T J) = \frac{\lambda_{\max}(J^T J)}{\lambda_{\min}(J^T J)} \ge \frac{\Theta(1)}{O(\varepsilon^2)} = \Omega(\varepsilon^{-2}). \tag{32}$$

For the pair alone, $\det(J_1^T J_1) = \|\mathbf{j}_a\|^2 \|(I - P_{\mathbf{j}_a})\mathbf{j}_b\|^2 = \Theta(\varepsilon^2)$ under the additional hypothesis, where $P_{\mathbf{j}_a}$ projects onto $\mathbf{j}_a$, while $\operatorname{tr}(J_1^T J_1) = \Theta(1)$; so $\lambda_{\min}(J_1^T J_1) = \det / \lambda_{\max} = \Theta(\varepsilon^2)$ and $\kappa(J_1^T J_1) = \Theta(\varepsilon^{-2})$. No matching upper bound holds for $\kappa(J^T J)$ in general: with free exponents on one ray, the exponent columns add a second near-null direction, and numerically $\kappa(J^T J)$ of the five-parameter Chinchilla law grows as ε^{-6} (Appendix B.11). □

Other smooth objectives. The argument uses only the column relation $\mathbf{j}_b = c\,\mathbf{j}_a + \boldsymbol{\delta}$. For an objective $\sum_i \rho(r_i)$ with $\rho'' \ge 0$ (weighted least squares, log-space fitting, or a Huber loss on residuals in its quadratic zone), the Gauss-Newton curvature is $J^T W J$ with $W = \operatorname{diag}(\rho''(r_1), \ldots, \rho''(r_m))$. Since $W^{1/2}\mathbf{j}_b = c\, W^{1/2}\mathbf{j}_a + W^{1/2}\boldsymbol{\delta}$, the proof above applies to $W^{1/2}J$ whenever the weights are bounded above and away from zero: $\kappa(J^T W J) = \Omega(\varepsilon^{-2})$.

B.4 Appendix - Cauchy-Schwarz gap identity

Lemma 2 (Cauchy-Schwarz gap identity)**.** *Let $\mathbf{f}, \mathbf{g} \in \mathbb{R}^m$ with $g_i = c\, f_i\, h_i$ for a scalar $c \ne 0$ and arbitrary $h_i > 0$. Define the power-law weights $w_i := f_i^2 / \sum_j f_j^2$ and the weighted variance $\sigma_w^2(h) := \sum_i w_i\, h_i^2 - \left(\sum_i w_i\, h_i\right)^2$. Then:*

$$\det \begin{pmatrix} \mathbf{f}^T\mathbf{f} & \mathbf{f}^T\mathbf{g} \\ \mathbf{g}^T\mathbf{f} & \mathbf{g}^T\mathbf{g} \end{pmatrix} = c^2 \left(\sum_i f_i^2\right)^2 \sigma_w^2(h). \tag{33}$$

Proof. Expand: $\mathbf{f}^T\mathbf{g} = c\sum_i f_i^2 h_i$, $\mathbf{g}^T\mathbf{g} = c^2 \sum_i f_i^2 h_i^2$, so $\det = (\sum_i f_i^2)(c^2 \sum_i f_i^2 h_i^2) - c^2(\sum_i f_i^2 h_i)^2 = c^2(\sum_i f_i^2)^2[\sum_i w_i h_i^2 - (\sum_i w_i h_i)^2]$. □

For $h_i = N_i^\eta$ with $|\eta|$ small, a Taylor expansion gives:

$$\sigma_w^2(N_i^\eta) = \eta^2\, \sigma_w^2(\log N_i) + O(|\eta|^3). \tag{34}$$

For Chinchilla, take $\eta = \tilde{\varepsilon} := \alpha - \beta$ and $\varepsilon := |\alpha - \beta|$ (Definition 2), so $\eta^2 = \varepsilon^2$. Thus Lemma 2 applies to $h_i = N_i^{\alpha-\beta}$, giving the exact determinant (37); since $\kappa_{A,B} \le \operatorname{tr}^2 / \det$ for a 2×2 block, this yields a non-asymptotic bound on $\kappa_{A,B}$ depending on ε, the sample size n, and the spread of $\{N_i\}$.

B.5 Appendix - Chinchilla scaling law: detailed analysis

The Chinchilla parameter vector is $\boldsymbol{\theta} = [A, B, E, \alpha, \beta]^T$ ($p = 5$). The following analysis holds on a single collinear ray $D_i = k_\ell N_i$ (i.e., $K = 1$ or a single fixed ratio k_ℓ from the design); the scale-coefficient columns then satisfy:

$$\frac{\partial \hat{L}_i}{\partial B} = \underbrace{k_\ell^{-\beta}}_{c} N_i^{-\alpha} N_i^{\alpha-\beta}, \tag{35}$$

with $\varepsilon := |\alpha - \beta|$. The corresponding residual Jacobian entries are $J_{iA} = -\partial \hat{L}_i/\partial A$ and $J_{iB} = -\partial \hat{L}_i/\partial B$, but $(J^T J)_{A,B}$ equals the Gram matrix of $(\partial \hat{L}/\partial A,\ \partial \hat{L}/\partial B)$ because both columns pick up the same sign. The determinant of the (A, B) sub-block of $J^T J$ satisfies:

$$\det\big((J^T J)_{A,B}\big) = O(\varepsilon^2), \tag{36}$$

and by (37) it is $\Theta(\varepsilon^2)$, so $\kappa((J^T J)_{A,B}) = \Theta(\varepsilon^{-2})$. Proposition 2 gives only $\kappa(J^T J) = \Omega(\varepsilon^{-2})$ for the full 5×5 system; numerically it grows as ε^{-6}, since the exponent columns add a second near-null direction (Appendix B.11).

Explicit finite-sample bound.

Applying Lemma 2 with $f_i = N_i^{-\alpha}$, $c = k_\ell^{-\beta}$, and $h_i = N_i^{\alpha-\beta}$, the sub-block determinant admits the exact factorization:

$$\det\big((J^T J)_{A,B}\big) = k_\ell^{-2\beta}\, \Phi_2^2\, \sigma_w^2\big(N_i^{\alpha-\beta}\big), \tag{37}$$

where $\Phi_2 := \sum_{i=1}^n N_i^{-2\alpha}$ and the weights are $w_i = N_i^{-2\alpha}/\Phi_2$. Using the Taylor expansion (34):

$$\det\big((J^T J)_{A,B}\big) \approx k_\ell^{-2\beta}\, \Phi_2^2\, \varepsilon^2\, \sigma_w^2(\log N), \tag{38}$$

with $\sigma_w^2(\log N) = \sum_i w_i (\log N_i)^2 - \big(\sum_i w_i \log N_i\big)^2$. The condition number of the sub-block is therefore:

$$\kappa\big((J^T J)_{A,B}\big) \approx \frac{\mathrm{tr}\big((J^T J)_{A,B}\big)^2}{k_\ell^{-2\beta}\, \Phi_2^2\, \varepsilon^2\, \sigma_w^2(\log N)}, \tag{39}$$

which makes the three sources of ill-conditioning explicit: (i) small exponent gap $\varepsilon \to 0$; (ii) narrow model-size range, i.e. small $\sigma_w^2(\log N)$ when the N_i are clustered; and (iii) finite sample effects through Φ_2 and the weights w_i, which concentrate on the *smallest* models and reduce the effective sample diversity. For the empirical gap $|\alpha - \beta| \approx 0.06$, the exact determinant (37) gives $\kappa_{A,B}$ on the order of 10^3-10^4 on our grid, rendering the scale pair numerically ill-conditioned. Equation (39) overstates it there, since on our 14 sizes $\sigma_w^2(N_i^{\tilde{\varepsilon}})$ is 7.4 times its Taylor value $\varepsilon^2 \sigma_w^2(\log N)$ at $\varepsilon = 0.06$; it gives the rate, not the value. □

B.6 Appendix - Full Jacobian of the Chinchilla law

Full Jacobian analysis. The preceding argument concerns only the A and B columns. We now list the five model sensitivities $\partial \hat{L}_i/\partial \theta_j$ on a single collinear ray $D_i = k_\ell N_i$; by (10), the residual Jacobian J is obtained by negating each of these columns, so $J^T J$ equals the Gram matrix of the sensitivities below. We claim the (E, α, β) coordinates do not rescue the conditioning. The complete set of partial derivatives is:

$$\begin{aligned}
\frac{\partial \hat{L}_i}{\partial A} &= N_i^{-\alpha}, & \frac{\partial \hat{L}_i}{\partial B} &= k_\ell^{-\beta} N_i^{-\beta}, \\
\frac{\partial \hat{L}_i}{\partial E} &= 1, & \frac{\partial \hat{L}_i}{\partial \alpha} &= -A N_i^{-\alpha} \log N_i, \\
\frac{\partial \hat{L}_i}{\partial \beta} &= -B k_\ell^{-\beta} N_i^{-\beta} (\log k_\ell + \log N_i).
\end{aligned} \tag{40}$$

Observe that (writing $\mathbf{j}_u$ for the m-vector with entries $\partial \hat{L}_i/\partial u$) $\mathbf{j}_\alpha = -A\ \mathrm{diag}(\log N_i)\, \mathbf{j}_A$ and $\mathbf{j}_\beta = -B\, (\log k_\ell)\, \mathbf{j}_B - B\ \mathrm{diag}(\log N_i)\, \mathbf{j}_B$. Using $\mathbf{j}_B \approx c\, \mathbf{j}_A + O(\varepsilon)$ (from (35)), both exponent columns

lie approximately in the two-dimensional subspace $\mathrm{span}\{\mathbf{j}_A, \mathrm{diag}(\log N_i)\,\mathbf{j}_A\}$. In particular, if $\mathbf{j}_B = c\,\mathbf{j}_A$ exactly, then $\mathbf{j}_\beta - (Bc/A)\,\mathbf{j}_\alpha = -Bc(\log k_\ell)\,\mathbf{j}_A$, so $\mathbf{j}_\alpha$ and $\mathbf{j}_\beta$ are *not* scalar multiples whenever $\log k_\ell \neq 0$; they lie near the span of the scale-coefficient pair and its log-modulated variants. The exponent columns nevertheless add a further near-null direction (Appendix B.11); the Rayleigh bound below needs only the pair. The E-column of sensitivities is $\mathbf{1}$, hence the E-column of J is $-\mathbf{1}$; either way it is linearly independent of the power-law columns but likewise does not break the collinearity between the A/B pair.

By Proposition 2, the near-null eigenvector $\mathbf{w}$ of $(J^TJ)_{A,B}$, extended to $\tilde{\mathbf{w}} \in \mathbb{R}^5$ with zeros in the (E, α, β) coordinates matching (28), satisfies $\tilde{\mathbf{w}}^T(J^TJ)\,\tilde{\mathbf{w}} = O(\varepsilon^2)$, hence $\kappa(J^TJ) = \Omega(\varepsilon^{-2})$ for the full 5×5 system; the exponent columns make it grow faster (Appendix B.11). □

B.7 Appendix - Repeated-data scaling law: detailed analysis

This subsection analyses the Jacobian of the repeated-data law on a single collinear ray ($K = 1$, $D_i = k_\ell N_i$). Muennighoff et al. (2025) retain the Chinchilla form but replace N and D with *effective* quantities N' and D' that decay under repetition:

$$L(N, D) = \frac{A}{N'^{\alpha}} + \frac{B}{D'^{\beta}} + E, \tag{41}$$

with $D' = U_D + U_D R_D^*(1 - e^{-R_D/R_D^*})$, $U_D = \min\{D_C, D\}$, $R_D = (D/U_D) - 1$, and analogously $N' = U_N + U_N R_N^*(1 - e^{-R_N/R_N^*})$, $R_N = (N/U_N) - 1$, with U_N Chinchilla-optimal for U_D. When $R_D = R_N = 0$ (single epoch, no excess parameters), $D' = D$ and $N' = N$, recovering Chinchilla exactly. We fit (41) as the released code does, with $U_N = N$, U_D the unique tokens and, in place of the definitions above, $R_N = R_D = e - 1$ at epoch e, all three epochs in one fit.

Under $D_i = k_\ell N_i$, both D_i' and N_i' become deterministic functions of N_i; write $D_i' = \phi(N_i)$ and $N_i' = \nu(N_i)$. The scale-coefficient sensitivities are $\partial\hat{L}_i/\partial A = \nu(N_i)^{-\alpha}$ and $\partial\hat{L}_i/\partial B = \phi(N_i)^{-\beta}$, so Lemma 2 applies to the scale-coefficient pair with $f_i = \nu(N_i)^{-\alpha}$, $c = k_\ell^{-\beta}$, and $h_i = k_\ell^{\beta}\,\phi(N_i)^{-\beta}\,\nu(N_i)^{\alpha}$, yielding $\det\big((J^TJ)_{A,B}\big) = c^2\,\Phi_2^2\,\sigma_w^2(h)$ with $\Phi_2 = \sum_i \nu(N_i)^{-2\alpha}$ and $w_i = \nu(N_i)^{-2\alpha}/\Phi_2$. The conditioning of the (A, B) sub-block therefore depends on the variance of h_i across the ray, which differs sharply between two regimes.

Unsaturated regime ($R_D = R_N = 0$). Here $\phi(N) = k_\ell N$ and $\nu(N) = N$, so $h_i = N_i^{\alpha-\beta}$. The Taylor expansion $\sigma_w^2(N_i^{\alpha-\beta}) \approx \varepsilon^2\,\sigma_w^2(\log N_i)$ (with $\varepsilon = |\alpha - \beta|$) gives $\det((J^TJ)_{A,B}) = \Theta(\varepsilon^2)$, so $\kappa_{A,B} = \Theta(\varepsilon^{-2})$ and, by Proposition 2, $\kappa(J^TJ) = \Omega(\varepsilon^{-2})$ for the full 5×5 system: the law inherits Chinchilla's ill-conditioning.

Saturated regime ($R_D, R_N \neq 0$). Once the ratio $\phi(N_i)/\nu(N_i)$ varies systematically with N_i, h_i ceases to be a pure power of N_i: the saturation curves of ϕ and ν inject an ε-independent contribution into $\log h_i$. We therefore expect $\sigma_w^2(h)$ to stay bounded away from zero as $\varepsilon \to 0$, in which case the Taylor expansion that produced the ε^2 factor no longer applies, the scale-coefficient columns are not near-proportional, and $\det((J^TJ)_{A,B}) = \Omega(1)$. This is a heuristic: we do not bound $\sigma_w^2(h)$ away from zero, nor treat the exponent and saturation columns R_D^*, R_N^*, so we claim no rate for $\kappa(J^TJ)$ in this regime.

The repeated-data law therefore inherits the $\Theta(\varepsilon^{-2})$ collinearity of Chinchilla *only* in the unsaturated regime. The (A, B) sub-block of Q is identically zero (as for Chinchilla) because $\hat{L}$ is linear in A and B. Empirically, in our seed-paired comparisons the repeated-data law tracks the Chinchilla pattern (CO holdout R^2 degrades and NC consistently wins), although its epoch-2 and epoch-3 observations are in the saturated regime: $R_D = 1, 2$ against fitted R_D^* of 0.80-2.35 (Table 50), so at epoch 3 D' is 58-78% of the tokens seen. The formal guarantee does not cover these fits, and the NC advantage on this law is empirical. □

B.8 Appendix - Kaplan scaling law: detailed analysis

This subsection expands on Section 3 for a single collinear ray ($K = 1$, $D_i = k_\ell N_i$). We fit the compositional form of Kaplan et al. (2020),

$$\hat{L}(N, D) = u^{\alpha_D}, \qquad u \coloneqq (N_c/N)^s + D_c/D, \qquad s \coloneqq \alpha_N/\alpha_D, \tag{42}$$

with $p = 4$ parameters $\boldsymbol{\theta} = [N_c, D_c, \alpha_N, \alpha_D]^T$ (the code's `kaplan_loss`; Table 47). It is not a reparameterization of the additive form $(N_c/N)^{\alpha_N} + (D_c/D)^{\alpha_D}$: its mixed derivative $\partial^2 \hat{L}/\partial N \partial D$ is nonzero, while the additive form's is zero. Its columns are

$$\mathbf{j}_{N_c} = \alpha_N \, u^{\alpha_D - 1} \, N_c^{s-1} N^{-s}, \qquad \mathbf{j}_{D_c} = \alpha_D \, u^{\alpha_D - 1} \, D^{-1}, \tag{43}$$

$$\mathbf{j}_{\alpha_N} = u^{\alpha_D - 1} (N_c/N)^s \log(N_c/N), \qquad \mathbf{j}_{\alpha_D} = \hat{L} \log u - s \, \mathbf{j}_{\alpha_N}, \tag{44}$$

with products taken element-wise over the observations.

Exponent gap. On the ray $D = k_\ell N$ the data term $D_c/(k_\ell N)$ falls as N^{-1}, against N^{-s} for the model term, so $\mathbf{j}_{D_c} = c \, \mathbf{j}_{N_c} \odot \mathbf{h}$ with $c = (\alpha_D/\alpha_N) \, N_c^{1-s} k_\ell^{-1}$ and $h_i = N_i^{\tilde{\varepsilon}}$, $\tilde{\varepsilon} := s - 1 = \alpha_N/\alpha_D - 1$. The gap is $\varepsilon = |1 - \alpha_N/\alpha_D|$ (Definition 2), not $|\alpha_D - \alpha_N|$. At $\varepsilon = 0$ the law on the ray is $((N_c + D_c/k_\ell)/N)^{\alpha_D}$, a function of $N_c + D_c/k_\ell$ and α_D alone, so the Jacobian has rank 2. Lemma 2 with $f_i = (\mathbf{j}_{N_c})_i$ gives $\det((J^T J)_{N_c, D_c}) \approx c^2 \Phi_2^2 \, \varepsilon^2 \, \sigma_w^2 (\log N)$ with $\Phi_2 := \sum_i f_i^2$ and weights $w_i \propto f_i^2$; ill-conditioning worsens when the model sizes are clustered or ε is small. The exponent columns (44) add a further near-null direction, as for Chinchilla (Appendix B.11); the Rayleigh bound below needs only the pair.

Size of the gap. At the single-variable exponents of Kaplan et al. (2020) ($\alpha_N = 0.076$ from their $L(N)$, $\alpha_D = 0.095$ from $L(D)$), $\varepsilon = 0.20$ (their joint $L(N, D)$ fit has $\alpha_D = 0.103$, $\varepsilon = 0.26$); with unit-norm columns κ_{N_c, D_c} is 1.6-1.7$\times 10^2$ and $\kappa(J^T J)$ is 4.3-4.8$\times 10^8$ on one ray of our 14 sizes (k = 1-20), against 1.2-1.4$\times 10^{14}$ for Chinchilla at Hoffmann et al.'s fit. The seed-0 fits of Table 47 and its first- and second-epoch companions have $\varepsilon = 0.105$ to 0.882 on CO; over all 450 CO fits (30 seeds), 445 have $\varepsilon \geq 0.105$, the smallest is 0.004, and the BF16 CO fit of Table 52 has 0.008. At every seed-0 fit, on its own CO or NC training design, Kaplan has the smallest unit-norm $\kappa(J^T J)$ of the three (N, D) laws. Its poor off-ray prediction after CO training (mean R^2 on $\mathcal{H}_{\text{nc}}$ of 0.50 against 0.78 for Chinchilla) is therefore not a conditioning effect: its Gauss-Newton predictive standard deviation on $\mathcal{H}_{\text{nc}}$ is 0.049 against 0.048 for Chinchilla, while its holdout RMSE is 0.366 against 0.242 (medians over the seed-0 CO fits); the error is bias, part of it in the CO fits whose (N_c, D_c) never left a start draw. In 183 of the 900 Kaplan fits (37 CO, 146 NC; 9 at seed 0) the optimizer left (N_c, D_c) within a relative 10^{-6} of one of its start draws in native units and moved only the exponents (Appendix E.7); four of them are rows of Table 47, marked there (`tools/kaplan_numbers.py` computes every number of this paragraph).

Applying Proposition 2 with $p = 4$: the extended near-null vector $\tilde{\mathbf{w}} \in \mathbb{R}^4$ padding $\mathbf{w}$ with zeros in the (α_N, α_D) coordinates as in (28) satisfies $\tilde{\mathbf{w}}^T (J^T J) \tilde{\mathbf{w}} = O(\varepsilon^2)$, hence $\kappa(J^T J) = \Omega(\varepsilon^{-2})$ for the full 4×4 system; with free exponents the full rate is faster, as for Chinchilla. □

B.9 Appendix - Droppo-Elibol scaling law: detailed analysis

We use the form fit in our experiments and recorded in Table 2,

$$\hat{L}(N, D) = \left[L_\infty^{1/\alpha} + (N_C/N)^{\gamma_N} + (D_C/D)^{\gamma_D} \right]^\alpha, \qquad \gamma_N := \alpha_N/\alpha, \;\; \gamma_D := \alpha_D/\alpha. \tag{45}$$

Droppo and Elibol (2021) motivate this compositional structure; we fit (45) with $p = 6$ parameters $(L_\infty, N_C, D_C, \alpha_N, \alpha_D, \alpha)$, as in Definition 2 and Table 48 (Table 2 writes E for L_∞). Substituting $D = k_\ell N$ (single collinear ray, $K = 1$),

$$\hat{L}(N, k_\ell N) = \left[L_\infty^{1/\alpha} + N^{-\gamma_N} \left(N_C^{\gamma_N} + D_C^{\gamma_D} \, k_\ell^{-\gamma_D} \, N^{\gamma_N - \gamma_D} \right) \right]^\alpha. \tag{46}$$

In particular, when $\gamma_N = \gamma_D$ this reduces to $[\, L_\infty^{1/\alpha} + C_{\text{total}} \, N^{-\gamma_N} \,]^\alpha$ with $C_{\text{total}} = N_C^{\gamma_N} + D_C^{\gamma_D} \, k_\ell^{-\gamma_D}$. When $\varepsilon = |\gamma_N - \gamma_D|$ (Definition 2) is small, Taylor expanding $N^{\gamma_N - \gamma_D}$ about $\gamma_N = \gamma_D$ recovers the same $O(\varepsilon)$ near-proportionality as in the Chinchilla case. The irreducible loss L_∞ remains independently identifiable.

Explicit finite-sample bound. With $v := L_\infty^{1/\alpha} + (N_C/N)^{\gamma_N} + (D_C/D)^{\gamma_D}$, the scale columns are $\partial \hat{L}/\partial N_C = \alpha \, v^{\alpha - 1} \, \gamma_N \, N_C^{\gamma_N - 1} N^{-\gamma_N}$ and $\partial \hat{L}/\partial D_C = \alpha \, v^{\alpha - 1} \, \gamma_D \, D_C^{\gamma_D - 1} D^{-\gamma_D}$. After absorbing the shared chain-rule factor $\alpha \, v_i^{\alpha - 1}$ into f_i, Lemma 2 applies with $f_i \propto \alpha \, v_i^{\alpha - 1} \, \gamma_N \, N_C^{\gamma_N - 1} \, N_i^{-\gamma_N}$, the scalar $\lambda = (\gamma_D/\gamma_N) \, (D_C^{\gamma_D - 1}/N_C^{\gamma_N - 1}) \, k_\ell^{-\gamma_D}$, and $h_i = N_i^{\tilde{\varepsilon}}$ with $\tilde{\varepsilon} := \gamma_N - \gamma_D$ and $\varepsilon = |\tilde{\varepsilon}|$, yielding

$$\det\left((J^T J)_{N_C, D_C} \right) \approx \lambda^2 \Phi_2^2 \, \varepsilon^2 \, \sigma_w^2 (\log N), \tag{47}$$

with weights $w_i \propto f_i^2$ and $\Phi_2 = \sum_i f_i^2$. As in the Chinchilla and Kaplan cases, ill-conditioning worsens when the training configurations are clustered in $\log N$ or when $|\gamma_N - \gamma_D|$ is small.

Full Jacobian analysis. Write $\mathbf{j}_u$ for the sensitivity vector $(\partial \hat{L}_i/\partial u)_i$. We derive the four remaining columns $\mathbf{j}_{L_\infty}, \mathbf{j}_{\alpha_N}, \mathbf{j}_{\alpha_D}$, and $\mathbf{j}_\alpha$, and show they do not break the near-singularity.

Irreducible loss: the chain rule gives $\mathbf{j}_{L_\infty} = \alpha\, v^{\alpha-1}\, \partial v/\partial L_\infty = L_\infty^{1/\alpha-1}\,(v_i^{\alpha-1})_i$, i.e. the constant scalar $L_\infty^{1/\alpha-1}$ times the observation-dependent vector $(v_i^{\alpha-1})_i$ - so $\mathbf{j}_{L_\infty}$ is *not* a constant vector, and only the scalar $\partial v/\partial L_\infty$ is i-independent. Linear independence with $(\mathbf{j}_{N_C}, \mathbf{j}_{D_C})$ follows because $v_i^{\alpha-1}$ depends on (N_i, D_i) through the sum v_i raised to $\alpha - 1$ and is therefore not a scalar multiple of $N_i^{-\gamma_N}$ or $D_i^{-\gamma_D}$, so $\mathbf{j}_{L_\infty}$ does not share the (N_C, D_C) near-collinearity.

Exponent columns: with $\partial\gamma_N/\partial\alpha_N = 1/\alpha$ and $\partial\gamma_N/\partial\alpha = -\gamma_N/\alpha$,

$$\frac{\partial \hat{L}}{\partial \alpha_N} = v^{\alpha-1}\,(N_C/N)^{\gamma_N}(\log N_C - \log N), \tag{48}$$

$$\frac{\partial \hat{L}}{\partial \alpha_D} = v^{\alpha-1}\,(D_C/D)^{\gamma_D}(\log D_C - \log D), \tag{49}$$

$$\frac{\partial \hat{L}}{\partial \alpha} = v^{\alpha}\log v - v^{\alpha-1}\Big[\tfrac{1}{\alpha}L_\infty^{1/\alpha}\log L_\infty + \gamma_N(N_C/N)^{\gamma_N}\log(N_C/N) + \gamma_D(D_C/D)^{\gamma_D}\log(D_C/D)\Big]. \tag{50}$$

The α_N-column (48) equals $\mathbf{j}_{N_C}$ scaled element-wise by $(N_C/\alpha_N)(\log N_C - \log N_i)$; likewise the α_D-column (49) equals $\mathbf{j}_{D_C}$ scaled element-wise by $(D_C/\alpha_D)(\log D_C - \log D_i)$. Both therefore lie in the subspace spanned by the scale-coefficient columns and their log-modulated variants. The α-column (50) combines $\log v$ with the L_∞ term and log-modulated scale terms. The exponent columns add a further near-null direction (Appendix B.11); the Rayleigh bound below needs only the pair.

Applying Proposition 2 with $p = 6$, extend the near-null $\mathbf{w} \in \mathbb{R}^2$ for the (N_C, D_C) pair to $\tilde{\mathbf{w}} \in \mathbb{R}^6$ by padding with zeros in the other four coordinates as in (28); then $\tilde{\mathbf{w}}^T(J^TJ)\tilde{\mathbf{w}} = O(\varepsilon^2)$, hence $\kappa(J^TJ) = \Omega(\varepsilon^{-2})$ for the full 6×6 system. The bound comes from the model-capacity and data-size terms. □

B.10 Appendix - Identified submodels on a single collinear ray ($K = 1$)

This subsection provides the formal statements supporting Section 3.2. Throughout, we fix $K = 1$, i.e. all training points share a single TPP ratio k_ℓ, so $D_i = k_\ell N_i$. The reduced model $L(N) = \psi\, N^{-\alpha} + E$ with $\psi \coloneqq A + B\, k_\ell^{-\alpha}$ (equation (3)) is defined in the main text. We next define the power sums Φ_q, weights w_i, and $\sigma_w^2(\log N)$ (Definitions 5, 6, and 7). Proposition 4 shows that the (ψ, α) Gram block of the reduced Jacobian has determinant $\psi^2\Phi_2^2\sigma_w^2(\log N)$ (via Lemma 2 in Appendix B.4), positive for any design with at least two distinct model sizes and independent of ε, and Corollary 2 records an $\Omega(\varepsilon^{-2})$ condition-number gain relative to the full model.

Let $N_1 < \cdots < N_n$ be the distinct model sizes (as in the observation model, Section 3).

Definition 5 (Power sums)**.** For $q > 0$, the *power sums* are $\Phi_q \coloneqq \sum_{i=1}^n N_i^{-q\alpha}$. In particular, $\Phi_2 \coloneqq \sum_{i=1}^n N_i^{-2\alpha}$.

Definition 6 (Weights)**.** Given Φ_2 from Definition 5, the *weights* are $w_i \coloneqq N_i^{-2\alpha}/\Phi_2$ for $i = 1, \ldots, n$.

Definition 7 (Weighted log-variance)**.** Given weights (w_i) from Definition 6, the *weighted log-variance* is $\sigma_w^2(\log N) \coloneqq \sum_{i=1}^n w_i(\log N_i)^2 - \big(\sum_{i=1}^n w_i \log N_i\big)^2$.

Proposition 4 (Identification of the reduced model)**.** *The Gram determinant of the (ψ, α) sub-block of $J_{\mathrm{red}}^T J_{\mathrm{red}}$ is*

$$\det\big((J_{\mathrm{red}}^T J_{\mathrm{red}})_{\psi,\alpha}\big) = \psi^2\,\Phi_2^2\,\sigma_w^2(\log N), \tag{51}$$

*where $\sigma_w^2(\log N)$ is as in Definition 7. The determinant is **positive** for any design with at least two distinct model sizes and is **independent of the exponent gap** ε. The full 3×3 matrix $J_{\mathrm{red}}^T J_{\mathrm{red}}$, with*

J_{red} the reduced Jacobian of columns 1, $N^{-\alpha}$ and $-\psi N^{-\alpha}\ln N$ (eqn. (58)), is non-singular iff the design has at least three distinct model sizes.

Corollary 2 (Condition-number improvement)**.** *In the setting of Proposition 4, with at least three distinct model sizes, and with $\kappa_{\mathrm{full}} := \kappa(J^TJ)$ for the full model (1) and $\kappa_{\mathrm{red}} := \kappa(J_{\mathrm{red}}^TJ_{\mathrm{red}})$, $\kappa_{\mathrm{full}}/\kappa_{\mathrm{red}} = \Omega(\varepsilon^{-2})$, since $\kappa_{\mathrm{full}} \geq \kappa_{A,B} = \Theta(\varepsilon^{-2})$ while κ_{red} does not depend on ε.*

B.10.1 Identified submodels: proofs

Proof of Proposition 4. Apply Lemma 2 with $f_i = N_i^{-\alpha}$ (the ψ-sensitivity $\partial\hat{L}_i/\partial\psi$), $g_i = c\,f_i\,h_i$ where $c = -\psi$ and $h_i = \log N_i$ (so that $g_i = \partial\hat{L}_i/\partial\alpha$). The Gram determinant is $c^2\,(\sum f_i^2)^2\,\sigma_w^2(h) = \psi^2\,\Phi_2^2\,\sigma_w^2(\log N)$. No ε appears because the reduced model contains a single power-law term. For the 3×3 matrix: with at most two distinct sizes, the three columns live in a space of dimension at most two, so it is singular. With three or more, suppose $c_0 + N^{-\alpha}(c_1 - c_2\psi\ln N)$ vanishes at every size, and let $t = \ln N$ and $f(t) = c_0 + e^{-\alpha t}(c_1 + c_2' t)$ with $c_2' = -c_2\psi$. Then $f'(t) = e^{-\alpha t}(c_2' - \alpha c_1 - \alpha c_2' t)$ has at most one zero unless $c_2' = c_1 = 0$, so by Rolle's theorem f has at most two zeros unless it is the constant c_0; three distinct zeros therefore force $c_0 = c_1 = c_2 = 0$, and the matrix is non-singular. □

Proof of Corollary 2. By Cauchy interlacing, $\kappa_{\mathrm{full}} \geq \kappa_{A,B}$, and $\kappa_{A,B} = \Theta(\varepsilon^{-2})$ by Proposition 2, whose orthogonal-part condition holds by eqn. (37). The reduced Jacobian, with columns 1, $N^{-\alpha}$ and $-\psi N^{-\alpha}\ln N$, has a limit as $\varepsilon \to 0$ that is non-singular with three or more distinct sizes (Proposition 4), so $\kappa_{\mathrm{red}} = \Theta(1)$. Hence $\kappa_{\mathrm{full}}/\kappa_{\mathrm{red}} = \Omega(\varepsilon^{-2})$. □

B.11 Appendix - Proof of Corollary 1 (Confidence interval inflation)

Corollary (Confidence interval inflation (full statement))**.** *Under i.i.d. Gaussian noise with variance σ^2 and a single collinear ray $D = k_\ell N$ ($K = 1$) with $n \geq 5$ distinct model sizes, the least-squares estimator $\hat{A}$ in the full Chinchilla model (1) satisfies*

$$\mathrm{Var}(\hat{A}) \;\geq\; \frac{\sigma^2\;\Phi_2^{(2\varepsilon)}}{\Phi_2^2\;\sigma_w^2(N_i^{\tilde{\varepsilon}})}, \tag{52}$$

where $\Phi_2^{(2\varepsilon)} := \sum_{i=1}^n N_i^{-2\beta}$ (equivalently $\sum_i N_i^{-2\alpha+2\tilde{\varepsilon}}$ with $\tilde{\varepsilon} := \alpha - \beta$ and $|\tilde{\varepsilon}| = \varepsilon$ from Definition 2), and $\sigma_w^2(N_i^{\tilde{\varepsilon}})$ uses the same weights (w_i) as $\sigma_w^2(\log N)$ (Definitions 6-7), with equality when E, α, and β are known. To leading order in ε:

$$\mathrm{Var}(\hat{A}) \;\geq\; \frac{\sigma^2}{\Phi_2\;\varepsilon^2\;\sigma_w^2(\log N)}\;\big(1 + O(\varepsilon)\big). \tag{53}$$

The corresponding variance of $\hat{\psi}$ in the reduced model (3) satisfies

$$\mathrm{Var}(\hat{\psi}) \;=\; \sigma^2\;\frac{G^{\mathrm{eff}}_{\alpha\alpha}}{\big(\Phi_2 - \Phi_1^2/n\big)\,G^{\mathrm{eff}}_{\alpha\alpha} - \big(G^{\mathrm{eff}}_{\psi\alpha}\big)^2}, \tag{54}$$

where G^{eff} is the (ψ,α) Schur complement after profiling E (see below for the explicit entries); all quantities are bounded and independent of ε. Combining (52)-(54) with nominal Gaussian interval half-widths (the reduced model is misspecified at $\varepsilon > 0$, so the interval on ψ has no exact 95% coverage) yields the simplified asymptotic ratio (5) recorded in Corollary 1, a lower bound.

We derive the variance bounds for $\hat{A}$ (full model) and $\hat{\psi}$ (reduced model) under i.i.d. Gaussian noise $L_i = \hat{L}(N_i, D_i; \boldsymbol{\theta}^*) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ with $\mathbb{E}[\epsilon_i] = 0$.

Variance of $\hat{A}$ in the full model. Under Gaussian noise the linearized covariance of the least-squares estimator (Lemma 1, Remark 3) is $\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2\,(J^TJ)^{-1}$, so $\mathrm{Var}(\hat{A}) = \sigma^2\,\big[(J^TJ)^{-1}\big]_{AA}$.

Partition J^TJ conformably as

$$J^TJ = \begin{pmatrix} G_{AB} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where $G_{AB} \coloneqq (J^T J)_{A,B} \in \mathbb{R}^{2\times 2}$ is the (A, B) sub-block and G_{22} contains the remaining (E, α, β) entries. By the Schur-complement formula for the inverse of a partitioned matrix:

$$\left[(J^T J)^{-1}\right]_{(A,B)} = \left(G_{AB} - G_{12}\, G_{22}^{-1}\, G_{21}\right)^{-1}. \tag{55}$$

Since $G_{12}\, G_{22}^{-1}\, G_{21}$ is positive semidefinite, the Schur complement satisfies $S_{AB} \coloneqq G_{AB} - G_{12}\, G_{22}^{-1}\, G_{21} \preceq G_{AB}$ (Löwner order), and therefore

$$\left[(J^T J)^{-1}\right]_{(A,B)} = S_{AB}^{-1} \succeq G_{AB}^{-1}. \tag{56}$$

In particular, the $(1, 1)$ entry obeys $\left[(J^T J)^{-1}\right]_{AA} \geq \left[G_{AB}^{-1}\right]_{AA}$.

For the 2×2 matrix $G_{AB} = \left(\begin{smallmatrix} a & b \\ b & d \end{smallmatrix}\right)$ with $a = \mathbf{j}_A^T \mathbf{j}_A = \Phi_2$, $d = \mathbf{j}_B^T \mathbf{j}_B = k_\ell^{-2\beta}\, \Phi_2^{(2\varepsilon)}$, and $\det(G_{AB}) = k_\ell^{-2\beta}\, \Phi_2^2\, \sigma_w^2(N_i^{\tilde{\varepsilon}})$ (equation (37)):

$$\left[G_{AB}^{-1}\right]_{AA} = \frac{d}{\det(G_{AB})} = \frac{k_\ell^{-2\beta}\, \Phi_2^{(2\varepsilon)}}{k_\ell^{-2\beta}\, \Phi_2^2\, \sigma_w^2(N_i^{\tilde{\varepsilon}})} = \frac{\Phi_2^{(2\varepsilon)}}{\Phi_2^2\, \sigma_w^2(N_i^{\tilde{\varepsilon}})}. \tag{57}$$

Hence

$$\mathrm{Var}(\hat{A}) = \sigma^2 \left[(J^T J)^{-1}\right]_{AA} \geq \sigma^2 \left[G_{AB}^{-1}\right]_{AA} = \frac{\sigma^2\, \Phi_2^{(2\varepsilon)}}{\Phi_2^2\, \sigma_w^2(N_i^{\tilde{\varepsilon}})},$$

establishing (52).

To obtain the leading-order form (53), note that $\Phi_2^{(2\varepsilon)} = \sum_i N_i^{-2\beta} = \Phi_2 + 2\tilde{\varepsilon}\, T_2 + O(\varepsilon^2)$ where $T_2 = \sum_i N_i^{-2\alpha} \log N_i$, so $\Phi_2^{(2\varepsilon)}/\Phi_2 = 1 + O(\varepsilon)$. Combined with the Taylor expansion $\sigma_w^2(N_i^{\tilde{\varepsilon}}) = \tilde{\varepsilon}^2\, \sigma_w^2(\log N_i) + O(|\tilde{\varepsilon}|^3)$ (equation (34)):

$$\frac{\Phi_2^{(2\varepsilon)}}{\Phi_2^2\, \sigma_w^2(N_i^{\tilde{\varepsilon}})} = \frac{1 + O(\varepsilon)}{\Phi_2\, \varepsilon^2\, \sigma_w^2(\log N)} = \frac{1}{\Phi_2\, \varepsilon^2\, \sigma_w^2(\log N)}\big(1 + O(\varepsilon)\big).$$

Variance of $\hat{\psi}$ in the reduced model. The reduced model (3) has parameter vector $\boldsymbol{\theta}_{\mathrm{red}} = (\psi, \alpha, E)^T$ and Jacobian $J_{\mathrm{red}} \in \mathbb{R}^{n\times 3}$ with columns

$$\mathbf{j}_\psi = \left(N_i^{-\alpha}\right)_{i=1}^n, \quad \mathbf{j}_\alpha = \left(-\psi\, N_i^{-\alpha} \log N_i\right)_{i=1}^n, \quad \mathbf{j}_E = \mathbf{1}_n. \tag{58}$$

Define $\Phi_1 \coloneqq \sum_i N_i^{-\alpha}$, $T_1 \coloneqq \sum_i N_i^{-\alpha} \log N_i$, $U_2 \coloneqq \sum_i N_i^{-2\alpha} (\log N_i)^2$. The 3×3 Gram matrix is

$$G_{\mathrm{red}} = \begin{pmatrix} \Phi_2 & -\psi\, T_2 & \Phi_1 \\ -\psi\, T_2 & \psi^2\, U_2 & -\psi\, T_1 \\ \Phi_1 & -\psi\, T_1 & n \end{pmatrix}.$$

Since $\mathrm{Var}(\hat{\psi}) = \sigma^2\, [G_{\mathrm{red}}^{-1}]_{\psi\psi}$, we compute the (ψ, α) Schur complement after profiling E:

$$\begin{aligned} G_{\psi\psi}^{\mathrm{eff}} &= \Phi_2 - \Phi_1^2/n, \\ G_{\psi\alpha}^{\mathrm{eff}} &= -\psi\, (T_2 - \Phi_1\, T_1/n), \\ G_{\alpha\alpha}^{\mathrm{eff}} &= \psi^2\, (U_2 - T_1^2/n). \end{aligned}$$

The first entry $G_{\psi\psi}^{\mathrm{eff}}$ is the sample variance of $\{N_i^{-\alpha}\}$ (times n); the other entries are analogous centered moments. The profiled 2×2 determinant is

$$\det(G^{\mathrm{eff}}) = G_{\psi\psi}^{\mathrm{eff}}\, G_{\alpha\alpha}^{\mathrm{eff}} - \left(G_{\psi\alpha}^{\mathrm{eff}}\right)^2,$$

and

$$\mathrm{Var}(\hat{\psi}) = \sigma^2\, \frac{G_{\alpha\alpha}^{\mathrm{eff}}}{\det(G^{\mathrm{eff}})}. \tag{59}$$

No power of ε appears anywhere in G^{eff}: the entries depend only on $\{N_i\}$, α, ψ, and n, all of which are $O(1)$ with respect to the exponent gap. Hence $\mathrm{Var}(\hat{\psi})$ is bounded and ε-independent, confirming (54).

Confidence-interval ratio. Under Gaussianity, the nominal 95% interval half-width for any scalar parameter θ_j is $1.96\sqrt{\mathrm{Var}(\hat{\theta}_j)}$. Therefore:

$$\begin{aligned}\frac{\mathrm{CI}_{0.95}(A)}{\mathrm{CI}_{0.95}(\psi)} &\geq \sqrt{\frac{\mathrm{Var}(\hat{A})}{\mathrm{Var}(\hat{\psi})}} \\ &\geq \sqrt{\frac{\Phi_2^{(2\varepsilon)}}{\Phi_2^2\ \sigma_w^2(N_i^{\tilde{\varepsilon}})} \cdot \frac{\det(G^{\mathrm{eff}})}{G^{\mathrm{eff}}_{\alpha\alpha}}} \\ &= \frac{1}{\varepsilon}\sqrt{\frac{\Phi_2^{(2\varepsilon)}}{\Phi_2}} \cdot \underbrace{\sqrt{\frac{\det(G^{\mathrm{eff}})}{\Phi_2\ \sigma_w^2(\log N)\ G^{\mathrm{eff}}_{\alpha\alpha}}}}_{\Theta(1)} \cdot (1+O(\varepsilon)),\end{aligned}$$

where in the last step we used $\sigma_w^2(N_i^{\tilde{\varepsilon}}) = \varepsilon^2\,\sigma_w^2(\log N)(1+O(\varepsilon))$ and $\Phi_2^{(2\varepsilon)}/\Phi_2 = 1+O(\varepsilon)$. The $\Theta(1)$ factor depends on the design $\{N_i\}$ but *not* on ε; it can be far below 1, so at a finite gap the bound may be weak. □

Correlated noise. With a noise covariance Σ in place of $\sigma^2 I$ (checkpoints of one run, a shared corpus), the linearized covariance of the estimator is $(J^TJ)^{-1}J^T\Sigma J(J^TJ)^{-1}$, which lies between $\lambda_{\min}(\Sigma)\,(J^TJ)^{-1}$ and $\lambda_{\max}(\Sigma)\,(J^TJ)^{-1}$ in the Löwner order, and likewise in the reduced model. Each variance above is then its value at $\sigma^2 = 1$ times a factor in $[\lambda_{\min}(\Sigma), \lambda_{\max}(\Sigma)]$, so a fixed Σ moves the ratio (5) by at most $\sqrt{\kappa(\Sigma)}$ in either direction and keeps its ε^{-1} rate.

Numerical rates. With E, α and β free, the exponent columns add a second near-null direction, whose rate we do not prove. In 80-digit arithmetic at the parameters of Hoffmann et al. (2022) on our 14 sizes and the ray $k = 20$ (`tools/single_ray_rates.py`), $\kappa(J^TJ)$ grows as ε^{-6}; numerically the ratio of the five-parameter fit grows as ε^{-3} and the off-ray predictive variance as ε^{-6}, while the floor is 0.16 at $\varepsilon = 0.06$, where the computed ratio is of order 10^4.

B.12 Appendix - Proof of Proposition 1 (TPP diversity and scale-pair conditioning)

We give a self-contained proof. Throughout, β_{eff} denotes the law-specific data-size exponent (Table 9). For any $K \geq 1$ distinct TPP ratios $k_1 < \cdots < k_K$, let each ratio be trained at the same $n \geq 1$ model sizes $\{N_i\}_{i=1}^n$, as in a collinear design (Definition 1). The proof uses this hypothesis: with unequal runs or model sizes per ratio, the rays enter $G_{A,B}$ with unequal weights. The training Jacobian is evaluated at the collinear design $D_{i\ell} = k_\ell N_i$.

Step 1: Jacobian columns for the scale-coefficient pair. For the Chinchilla and unsaturated repeated-data laws of Definition 2, the two scale-coefficient columns of J on observation $(N_i, k_\ell N_i)$ are (cf. Appendix B.5, eqn. (35) for Chinchilla; analogous expressions hold for unsaturated repeated-data with the substitutions of Table 9):

$$(j_A)_{i\ell} = N_i^{-\alpha}, \qquad (j_B)_{i\ell} = k_\ell^{-\beta_{\mathrm{eff}}}\,N_i^{-\alpha}\,N_i^{\tilde{\varepsilon}}, \tag{60}$$

where $\tilde{\varepsilon}$ is the *signed* native gap in each law (e.g. $\alpha-\beta$ for Chinchilla and repeated-data, $\alpha_N/\alpha_D - 1$ for Kaplan, $\gamma_N - \gamma_D$ for Droppo-Elibol), so $|\tilde{\varepsilon}| = \varepsilon$ in Definition 2, and the displayed α is the model-side exponent paired with β_{eff} as in Table 9.

Kaplan and Droppo-Elibol. Their native columns carry constant prefactors (for Kaplan $\alpha_N N_c^{\alpha_N/\alpha_D - 1}$ and α_D, Appendix B.8; for Droppo-Elibol $\gamma_N N_C^{\gamma_N - 1}$ and $\gamma_D D_C^{\gamma_D - 1}$, Appendix B.9) and a chain-rule factor that changes from ray to ray. $\kappa_{A,B}$ changes when a column is rescaled, while V_K and τ_K do not, so for these two laws the proposition holds for the term coefficients (for Droppo-Elibol $N_C^{\gamma_N}$, $D_C^{\gamma_D}$) only where the chain-rule factor is constant, and is otherwise a leading-order heuristic. The Droppo-Elibol law at $\alpha = 1$, which is the Chinchilla function with $\gamma_N = \alpha_N$, $N_C = A^{1/\alpha_N}$ and $D_C = B^{1/\alpha_D}$, at the parameters of Hoffmann et al. (2022) on our 10 training sizes has $\kappa_{A,B} = 27{,}186$ for $\{20, 400\}$ in its native pair (N_C, D_C), against 27 in (A, B), at the same V_K/τ_K.

Step 2: (A, B) Gram sub-block. Define $\Phi_2 := \sum_{i=1}^n N_i^{-2\alpha}$ (Definition 5). The four entries of the 2×2 Gram sub-block $G_{A,B} = J_{A,B}^T J_{A,B}$ are obtained by summing over all $m = nK$ observations.

Working at leading order in ε (i.e., $N_i^{\tilde{\varepsilon}} = 1 + O(\varepsilon)$, $N_i^{2\tilde{\varepsilon}} = 1 + O(\varepsilon)$):

$$(G_{A,B})_{11} = \sum_{\ell=1}^{K} \sum_{i=1}^{n} N_i^{-2\alpha} = K\,\Phi_2, \tag{61}$$

$$(G_{A,B})_{12} = \sum_{\ell=1}^{K} k_\ell^{-\beta_{\text{eff}}} \sum_{i=1}^{n} N_i^{-2\alpha}\, N_i^{\tilde{\varepsilon}} = \Phi_2 \sum_{\ell=1}^{K} k_\ell^{-\beta_{\text{eff}}} + O(\varepsilon), \tag{62}$$

$$(G_{A,B})_{22} = \sum_{\ell=1}^{K} k_\ell^{-2\beta_{\text{eff}}} \sum_{i=1}^{n} N_i^{-2\alpha}\, N_i^{2\tilde{\varepsilon}} = \Phi_2 \sum_{\ell=1}^{K} k_\ell^{-2\beta_{\text{eff}}} + O(\varepsilon). \tag{63}$$

Write $S_1 \coloneqq \sum_{\ell=1}^{K} k_\ell^{-\beta_{\text{eff}}}$ and $S_2 \coloneqq \sum_{\ell=1}^{K} k_\ell^{-2\beta_{\text{eff}}}$. Then at leading order:

$$G_{A,B} \;=\; \Phi_2 \begin{pmatrix} K & S_1 \\ S_1 & S_2 \end{pmatrix} + O(\varepsilon). \tag{64}$$

Step 3: Trace, determinant, and condition number. From (64):

$$\operatorname{tr}(G_{A,B}) = \Phi_2\,(K + S_2) + O(\varepsilon), \tag{65}$$

$$\det(G_{A,B}) = \Phi_2^2\,(K\,S_2 - S_1^2) + O(\varepsilon). \tag{66}$$

Note that $K\,S_2 - S_1^2 = K^2 V_K$, where V_K is the sample variance of $\{k_\ell^{-\beta_{\text{eff}}}\}$ defined in the proposition statement (eqn. (2)):

$$V_K = \frac{1}{K}\sum_{\ell=1}^{K} k_\ell^{-2\beta_{\text{eff}}} - \Big(\frac{1}{K}\sum_{\ell=1}^{K} k_\ell^{-\beta_{\text{eff}}}\Big)^2 = \frac{K\,S_2 - S_1^2}{K^2}. \tag{67}$$

For a 2×2 positive-semidefinite matrix with eigenvalues $\lambda_+ \geq \lambda_- \geq 0$, $\kappa = \lambda_+/\lambda_-$ and the identities $\lambda_+ \cdot \lambda_- = \det$, $\lambda_+ + \lambda_- = \operatorname{tr}$ give $\lambda_\pm = \frac{1}{2}\big(\operatorname{tr} \pm \sqrt{\operatorname{tr}^2 - 4\det}\big)$, hence $\kappa = \frac{\operatorname{tr}^2}{2\det} - 1 + \frac{\operatorname{tr}^2}{2\det}\sqrt{1 - 4\det/\operatorname{tr}^2}$. At leading order, when $\det \ll \operatorname{tr}^2$ (the regime of interest), $\sqrt{1 - 4\det/\operatorname{tr}^2} \approx 1 - 2\det/\operatorname{tr}^2$, so $\kappa \approx \operatorname{tr}^2/\det - 2 \approx \operatorname{tr}^2/\det$. Therefore:

$$\kappa_{A,B} \;\approx\; \frac{\operatorname{tr}(G_{A,B})^2}{\det(G_{A,B})} \;=\; \frac{\Phi_2^2\,(K + S_2)^2}{\Phi_2^2\,K^2\,V_K} \;=\; \frac{(K + S_2)^2}{K^2\,V_K}\,. \tag{68}$$

The factors of Φ_2^2 cancel exactly.

Step 4: The two directions. For a 2×2 positive-definite matrix with condition number κ, $\operatorname{tr}^2/\det = (\lambda_+ + \lambda_-)^2/(\lambda_+\lambda_-) = \kappa + 2 + 1/\kappa$ exactly, and $x \mapsto x + 2 + 1/x$ increases on $x \geq 1$. At leading order in ε, eqns. (65)-(67) give $\operatorname{tr}^2/\det = (K + S_2)^2/(K^2 V_K) = \kappa_{\text{target}}\tau_K/V_K$ with τ_K as in (2). Hence, for $\kappa_{\text{target}} \geq 1$,

$$\kappa_{A,B} \leq \kappa_{\text{target}} \iff \frac{\kappa_{\text{target}}\,\tau_K}{V_K} \leq \kappa_{\text{target}} + 2 + \frac{1}{\kappa_{\text{target}}} \iff V_K \geq \tau_K\,\frac{\kappa_{\text{target}}^2}{(1 + \kappa_{\text{target}})^2}. \tag{69}$$

($\Longleftarrow$ of (2)) If $V_K \geq \tau_K$, then $V_K \geq \tau_K \kappa_{\text{target}}^2/(1 + \kappa_{\text{target}})^2$, so $\kappa_{A,B} \leq \kappa_{\text{target}}$ by (69). **(Converse)** If $\kappa_{A,B} \leq \kappa_{\text{target}}$, then (69) gives the stated necessary condition, which is weaker than $V_K \geq \tau_K$ by the factor $(1 + 1/\kappa_{\text{target}})^{-2}$, about 2% at $\kappa_{\text{target}} = 100$. This establishes Proposition 1.

Step 5: The three cases.

Case $K = 1$. When $K = 1$, there is a single TPP ratio k_1, so $S_1 = k_1^{-\beta_{\text{eff}}}$, $S_2 = k_1^{-2\beta_{\text{eff}}}$, and $V_1 = k_1^{-2\beta_{\text{eff}}} - (k_1^{-\beta_{\text{eff}}})^2 = 0$. The threshold satisfies $\tau_1 = (1 + k_1^{-2\beta_{\text{eff}}})^2/\kappa_{\text{target}} > 0$ for every $\kappa_{\text{target}} > 0$. Therefore $V_1 = 0 < \tau_1$ and the condition (2) can never hold. Equivalently, $\det(G_{A,B}) = O(\varepsilon^2)$ (the next-order term from $N_i^{\tilde{\varepsilon}} \neq 1$; see eqns. (37)-(38)), confirming the ill-conditioning $\kappa_{A,B} = \Theta(\varepsilon^{-2})$ established by Proposition 2.

Case $K = 2$. Let $k_1 < k_2$ and $R \coloneqq k_2/k_1$. Then:

$$\begin{aligned} S_1 &= k_1^{-\beta_{\text{eff}}}(1 + R^{-\beta_{\text{eff}}}), \qquad S_2 = k_1^{-2\beta_{\text{eff}}}(1 + R^{-2\beta_{\text{eff}}}), \\ V_2 &= \tfrac{1}{2}S_2 - \big(\tfrac{1}{2}S_1\big)^2 = \tfrac{1}{4}\,k_1^{-2\beta_{\text{eff}}}\,(1 - R^{-\beta_{\text{eff}}})^2. \end{aligned} \tag{70}$$

Let $s \coloneqq k_1^{-2\beta_{\text{eff}}}$, $r \coloneqq R^{-\beta_{\text{eff}}} \in (0,1)$ and $a \coloneqq \sqrt{\kappa_{\text{target}}\, s}$. By (70), $V_2 = s(1-r)^2/4$, and $\tau_2 = (2+s+sr^2)^2/(4\kappa_{\text{target}})$ by (2). Hence

$$\begin{aligned} V_2 \geq \tau_2 &\iff \kappa_{\text{target}}\, s\,(1-r)^2 \geq (2+s+sr^2)^2 && \text{multiplying by } 4\kappa_{\text{target}} > 0 \\ &\iff a\,(1-r) \geq 2+s+sr^2 && \text{both sides are positive} \\ &\iff q(r) \coloneqq sr^2 + ar + 2 + s - a \leq 0. \end{aligned}$$

Since $s, a > 0$, q is strictly increasing on $[0,1]$, and $q(1) = 2+2s > 0$. If $q(0) = 2+s-a \geq 0$, no $r \in (0,1)$ satisfies $q(r) \leq 0$. Otherwise the $r \in (0,1)$ with $q(r) \leq 0$ form the interval $(0, r^*]$, where $r^* \coloneqq \big(\sqrt{a^2 - 4s(2+s-a)} - a\big)/(2s)$ is the positive root of q, and $r^* < 1$. As $R = r^{-1/\beta_{\text{eff}}}$ decreases in r, and $a > 2+s$ reads $\sqrt{\kappa_{\text{target}}}\, k_1^{-\beta_{\text{eff}}} > 2 + k_1^{-2\beta_{\text{eff}}}$, this proves

$$V_2 \geq \tau_2 \iff R \geq R_{\min}(k_1) \coloneqq \begin{cases} (r^*)^{-1/\beta_{\text{eff}}} & \text{if } \sqrt{\kappa_{\text{target}}}\, k_1^{-\beta_{\text{eff}}} > 2 + k_1^{-2\beta_{\text{eff}}}, \\ \infty & \text{otherwise,} \end{cases} \tag{71}$$

so no spread suffices when $R_{\min}(k_1) = \infty$. $R_{\min}$ increases with $k_1 \geq 1$, a scale effect of the B column only: at fixed R, k_1 multiplies $\mathbf{j}_B$ by a constant. Let $g(s) \coloneqq s/(2+(1+r^2)s)^2$. At fixed R, $V_2/\tau_2 = \kappa_{\text{target}}(1-r)^2\, g(s)$, and $g'(s) = (2-(1+r^2)s)/(2+(1+r^2)s)^3 > 0$, because $(1+r^2)s < 2$ when $s \leq 1$, that is, when $k_1 \geq 1$. So at fixed R, V_2/τ_2 decreases as k_1 grows. Let $1 \leq k_1 < k_1'$ with $R' \coloneqq R_{\min}(k_1') < \infty$. Then $V_2/\tau_2 = 1$ at (k_1', R'), since $q(r^*) = 0$, so $V_2/\tau_2 > 1$ at (k_1, R'). By (71) at k_1, $R_{\min}(k_1) \leq R'$, and $R_{\min}(k_1) \neq R'$, since $V_2/\tau_2 = 1$ at $(k_1, R_{\min}(k_1))$. Hence $R_{\min}(k_1) < R_{\min}(k_1')$; when $R_{\min}(k_1') = \infty$, $R_{\min}(k_1) \leq R_{\min}(k_1')$ holds trivially. Table 10 tabulates (71) at $\kappa_{\text{target}} = 10^2$.

Case $K > 2$. For $K > 2$ there is no closed form like (71); one evaluates $V_K \geq \tau_K$ in (2) for the proposed ratios. Let $x_\ell \coloneqq k_\ell^{-\beta_{\text{eff}}}$, so $x_K < \cdots < x_1$, and let $c \coloneqq (x_1 + x_K)/2$. A variance is at most the mean squared deviation from any point, and $|x_\ell - c| \leq (x_1 - x_K)/2$, so $V_K \leq \frac{1}{K}\sum_{\ell=1}^{K}(x_\ell - c)^2 \leq (x_1 - x_K)^2/4$, the V_2 of the endpoint pair $\{k_1, k_K\}$. The second inequality is strict when some x_ℓ lies strictly between x_K and x_1, so interior rays lower V_K. The threshold depends on the rays only through S_2/K, the mean of the x_ℓ^2, which lies in $[x_K^2, x_1^2]$, so $\tau_K \geq (1+x_K^2)^2/\kappa_{\text{target}}$. Let $\tau_2^{\text{end}} \coloneqq (1+(x_1^2+x_K^2)/2)^2/\kappa_{\text{target}}$ be the threshold of the endpoint pair. Expanding the right side gives $1 + (x_1^2+x_K^2)/2 \leq (1+x_1^2/2)(1+x_K^2)$, and $x_1^2 = k_1^{-2\beta_{\text{eff}}}$, so

$$\begin{aligned} \frac{V_K}{\tau_K} &\leq \frac{(x_1-x_K)^2/4}{(1+x_K^2)^2/\kappa_{\text{target}}} && \text{both bounds above} \\ &= \Big(\frac{1+(x_1^2+x_K^2)/2}{1+x_K^2}\Big)^2 \frac{(x_1-x_K)^2/4}{\tau_2^{\text{end}}} && \text{definition of } \tau_2^{\text{end}} \\ &\leq \Big(1 + \tfrac{1}{2}\, k_1^{-2\beta_{\text{eff}}}\Big)^2 \frac{(x_1-x_K)^2/4}{\tau_2^{\text{end}}} && \text{the inequality above.} \end{aligned} \tag{72}$$

So a K-ray design beats the two-ray design on its own endpoints by at most the factor $(1+k_1^{-2\beta_{\text{eff}}}/2)^2$, which is 1.09 to 1.20 at $k_1 = 20$ for $\beta_{\text{eff}} \in [0.28, 0.40]$; in the worked example of Appendix B.15, a middle ray lowers V_K/τ_K from 1.13 to 0.785.

Step 6: The full matrix. By Cauchy interlacing, $\kappa_{A,B} \leq \kappa(J^TJ)$, so $\kappa(J^TJ) \leq \kappa_{\text{target}}$ implies $\kappa_{A,B} \leq \kappa_{\text{target}}$ and hence, at leading order, $V_K \geq \tau_K/(1+1/\kappa_{\text{target}})^2$ (Step 4): the threshold is nearly necessary for a well-conditioned full matrix. It is not sufficient: the offset and exponent columns enter J^TJ at other scales, and passing designs such as $k \in \{20, 400\}$ keep $\kappa(J^TJ)$ many orders of magnitude above κ_{target}.

Accuracy of the leading order. Beyond the factor $(1+1/\kappa_{\text{target}})^{-2}$ of Step 4, (2) rests on $N_i^{\bar{\varepsilon}} = 1 + O(\varepsilon)$ (the leading-order values below are $\text{tr}^2/\det$), although $\varepsilon \ln N_i$ is near 1 at our sizes. On our 10 training sizes at $\alpha = 0.34$ and $\beta = 0.28$, the exact $\kappa_{A,B}$ of $\{20, 100\}$ is 80 (leading order 208, fails $\kappa_{\text{target}} = 100$), of $\{1, 5\}$ it is 148 (leading order 88, passes), and of $\{20, 400\}$ it is 27 (leading order 82); a thousandfold rise in every size moves the second to 282. Over the 432 two-ratio designs with k_1 from 1 to 400, R from 1.5 to 100, $\beta_{\text{eff}} \in \{0.28, 0.35, 0.40\}$, a gap of 0.06, our 10 sizes and a thousandfold rise of them, the exact $\kappa_{A,B}$ runs from 0.06 to 3.4 times its leading-order value. The leading order can therefore misjudge a design several-fold in either direction, and its

statement that no spread suffices past $k_1^{\max}$ (Table 10) holds only at leading order: $\{400, 4000\}$ at $\beta_{\text{eff}} = 0.28$, past $k_1^{\max} = 292$ with $V_2/\tau_2 = 0.19$, has exact $\kappa_{A,B} = 90 \leq 100$ on our sizes. The exact $\kappa_{A,B}$ of a planned design is computable before training from its planned (N, D) and priors on α and β_{eff}; it, not (2), is the decision rule of the recipe (Appendix B.15), and it depends on absolute size even where its rate does not. □

B.13 Appendix - Proof of Theorem 1 (Holdout prediction regimes)

Theorem (Theorem 1, restated)**.** *For any law of Table 2 (repeated-data unsaturated) with i.i.d. Gaussian residuals of variance σ^2, let CO be a collinear design ($K \geq 1$), NC a non-collinear design with dense two-dimensional (N, D) coverage, both with full-rank Jacobians (NC's also at $\varepsilon = 0$), and $\mathcal{H}$ a holdout with a nonzero fraction of off-ratio points; RMSE is against $\mathbb{E}[L]$. At leading order in σ (Gauss-Newton linearization of both fits), then as $\varepsilon \to 0$: **Regime A** (CO a single ray, $K = 1$, so $V_1 = 0 < \tau_1$): $\mathbb{E}[\text{RMSE}_{\mathcal{H}}^{\text{NC}}] < \mathbb{E}[\text{RMSE}_{\mathcal{H}}^{\text{CO}}]$, and thus $\mathbb{E}[R^{2\,\text{NC}}_{\mathcal{H}}] > \mathbb{E}[R^{2\,\text{CO}}_{\mathcal{H}}]$. **Regime B** ($K \geq 2$, CO's Jacobian of full rank at $\varepsilon = 0$): $\mathbb{E}[\text{RMSE}_{\mathcal{H}}^{\text{CO}}]/\mathbb{E}[\text{RMSE}_{\mathcal{H}}^{\text{NC}}] = \Theta(1)$.*

Scope. The linearization in σ, not a bound on $\kappa(J^T J)$, licenses the Gauss-Newton expansions below. We treat Regime A via the $K = 1$ collinear (CO) ray $D_i = kN_i$ ($V_1 = 0$, so (2) fails) versus a non-collinear (NC) design whose Jacobian has full column rank at $\varepsilon = 0$; NC need not satisfy $V_K \geq \tau_K$. We derive the RMSE ordering on a general holdout $\mathcal{H}$ containing at least one off-ratio point, then establish Regime B for every $K \geq 2$ design whose Jacobian has full rank at $\varepsilon = 0$, whatever V_K/τ_K. Lemma 3 after the proof gives a sufficient condition for the ordering at $K \geq 2$.

Setup and linearization. Both designs fit the same scaling law $\hat{L}(N, D; \boldsymbol{\theta})$ by nonlinear least squares (Sec. 3.1). Well-specified training losses satisfy $L_i = \hat{L}(N_i, D_i; \boldsymbol{\theta}^\star) + \eta_i$ with $\mathbb{E}[\eta_i] = 0$ and $\text{Var}(\eta_i) = \sigma^2$, i.i.d. across i (as in Corollary 1). Under Gauss-Newton linearization, the parameter estimation error from training set $\mathcal{D}_T$ satisfies

$$\delta\boldsymbol{\theta} = (J_T^T J_T)^{-1} J_T^T \boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|^2), \tag{73}$$

where J_T is the training Jacobian and $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$. The prediction error at test point (N', D') is, to first order:

$$\hat{L}(N', D'; \hat{\boldsymbol{\theta}}) - L^*(N', D') \approx \mathbf{j}_{(N', D')}^T \delta\boldsymbol{\theta}. \tag{74}$$

For any test set $\mathcal{S} = \{(N_i', D_i')\}_{i=1}^{|\mathcal{S}|}$, the leading-order unconditional expected squared prediction error is (Björck, 1996, §2.8.3, eq. (2.8.7), p. 118):

$$\mathbb{E}\big[(\hat{L}(N', D'; \hat{\boldsymbol{\theta}}) - L^*(N', D'))^2\big] = \sigma^2\, \mathbf{j}^T (J_T^T J_T)^{-1} \mathbf{j} + O(\sigma^4), \tag{75}$$

$$\mathbb{E}[\text{MSE}_{\mathcal{S}}] = \frac{\sigma^2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \mathbf{j}_i^T (J_T^T J_T)^{-1} \mathbf{j}_i, \tag{76}$$

where $L^*(N', D') = \mathbb{E}[L \mid N', D']$ under the same homoscedastic model. Every expansion below is first in σ (the Gauss-Newton linearization of both fits) and then as $\varepsilon \to 0$; the $O(\sigma^4)$ remainder of the CO fit has constants that grow as $\varepsilon \to 0$, so this order of limits is part of the statement.

A lower bound on the CO predictive variance. On the single ray $D_i = kN_i$ of the Chinchilla law the sloppy space is not one-dimensional: at $\varepsilon = 0$ both $\mathbf{j}_B = k^{-\beta}\mathbf{j}_A$ and $\mathbf{j}_\beta \in \text{span}(\mathbf{j}_A, \mathbf{j}_\alpha)$, so $J_T^T J_T$ has two small eigenvalues, and we do not use its spectrum. For a positive definite Q and any $\mathbf{u} \neq \mathbf{0}$, the Cauchy-Schwarz inequality applied to $Q^{-1/2}\mathbf{j}$ and $Q^{1/2}\mathbf{u}$ gives $\mathbf{j}^T Q^{-1} \mathbf{j} \geq (\mathbf{j}^T \mathbf{u})^2/(\mathbf{u}^T Q \mathbf{u})$. With $Q = J_T^T J_T$ and $\mathbf{u}$ equal to 1 on A, $-k^\beta$ on B and 0 elsewhere, $J_T \mathbf{u}$ has entries $N_i^{-\alpha} - N_i^{-\beta} = N_i^{-\alpha}(1 - N_i^{\tilde{\varepsilon}})$ with $\tilde{\varepsilon} := \alpha - \beta$, so $\mathbf{u}^T Q \mathbf{u} = \varepsilon^2 \sum_i N_i^{-2\alpha} (\ln N_i)^2\, (1 + O(\varepsilon))$, while at a test point (N', D') with $k' = D'/N'$

$$\mathbf{j}^T \mathbf{u} = N'^{-\alpha} \big[1 - (k/k')^\beta\, N'^{\tilde{\varepsilon}}\big]. \tag{77}$$

Definition 8 (Sloppy leverage)**.** For test set $\mathcal{S} = \{(N_i', D_i')\}_{i=1}^{|\mathcal{S}|}$ evaluated against a model trained at TPP k, with β_{eff} as in Proposition 1:

$$\Lambda(\mathcal{S}) := \frac{1}{|\mathcal{S}|\, \Phi_2} \sum_{i=1}^{|\mathcal{S}|} N_i'^{-2\alpha} \left[1 - (k/k_i')^{\beta_{\text{eff}}}\right]^2, \tag{78}$$

where $k_i' = D_i'/N_i'$.

Averaging the bound over $\mathcal{S}$ and using (76):

$$\mathbb{E}[\mathrm{MSE}_{\mathcal{S}}^{\mathrm{CO}}] \;\geq\; \sigma^2\,\varepsilon^{-2}\,\Lambda(\mathcal{S})\,G\,\big(1+O(\varepsilon)\big), \qquad G \coloneqq \frac{\Phi_2}{\sum_i N_i^{-2\alpha}(\ln N_i)^2} > 0, \tag{79}$$

where G depends only on the training design. The bound is not tight: the second near-null direction, from the exponent columns, makes $\mathbb{E}[\mathrm{MSE}_{\mathcal{S}}^{\mathrm{CO}}]$ grow as ε^{-6} numerically (Appendix B.11), which only strengthens the ordering below.

Other laws. The same step needs only Proposition 2's hypothesis on the ray, $\mathbf{j}_b = c\,\mathbf{j}_a + \boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\| = O(\varepsilon)$: with $\mathbf{u}$ equal to $-c$ on a, 1 on b and 0 elsewhere, $J_T\mathbf{u} = \boldsymbol{\delta}$, so $\mathbf{u}^T Q\mathbf{u} = O(\varepsilon^2)$ and $\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{CO}}] = \Omega(\sigma^2\varepsilon^{-2})$ as soon as one holdout row has $(\mathbf{j}')_b - c\,(\mathbf{j}')_a \neq 0$ at $\varepsilon = 0$. For the four laws of Table 2 the ratio of the pair's columns depends on the TPP ratio, so every off-ratio point qualifies. For the saturated repeated-data law the pair is expected, heuristically, not to be near-proportional on one ray (Appendix B.7); the hypothesis may then fail, and Regime A is not claimed there.

NC Gram matrix: all eigenvalues $\Theta(1)$. By the NC hypothesis of Theorem 1, J_{NC} has full column rank at $\varepsilon = 0$, so $\lambda_{\min}(J_{\mathrm{NC}}^T J_{\mathrm{NC}}) = \Theta(1)$ as $\varepsilon \to 0$; NC need not satisfy $V_K \geq \tau_K$. The scale pair shows why dense coverage gives this. The Jacobian columns for the scale coefficients are: $(\mathbf{j}_A)_i = N_i^{-\alpha}$ and $(\mathbf{j}_B)_i = D_i^{-\beta}$. Under the collinear constraint $D_i = kN_i$, these satisfy the near-proportionality $\mathbf{j}_B = k^{-\beta}\mathbf{j}_A + O(\varepsilon)$. Under the NC design, observations sample diverse $k_i' = D_i/N_i$, so

$$\|\mathbf{j}_B - c\,\mathbf{j}_A\|^2 = \sum_i \big(D_i^{-\beta} - c\,N_i^{-\alpha}\big)^2 \tag{80}$$

is $\Theta(1)$ for any scalar c, since the ratio $D_i^{-\beta}/N_i^{-\alpha}$ varies across training points. The near-proportionality condition in Proposition 2 fails, so Proposition 2 does not apply.

The predictive variance at any test point is therefore:

$$\mathbf{j}^T(J_T^T J_T)^{-1}\mathbf{j} = \sum_{j=1}^{p} \frac{(\mathbf{j}^T\mathbf{v}_j^{\mathrm{NC}})^2}{\lambda_j^{\mathrm{NC}}} = \frac{O(1)}{\Theta(1)} = O(1), \tag{81}$$

regardless of test geometry. Hence:

$$\mathbb{E}[\mathrm{MSE}_{\mathcal{S}}^{\mathrm{NC}}] = O(\sigma^2) \tag{82}$$

for any test set $\mathcal{S}$.

General holdout RMSE ordering. Let $\mathcal{H}$ be a holdout set containing at least one point with $k_i' \neq k$ (off-ratio). We first establish the strict ordering for $\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{X}}]$ defined in (76) (errors against $L^* = \mathbb{E}[L]$, as in Theorem 1). Theorem 1 is stated for $\mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{X}}]$. Under the Gauss-Newton linearization, $\mathrm{MSE}_{\mathcal{H}}^{\mathrm{X}}$ is a positive-semidefinite quadratic form in the centred Gaussian noise, so $\mathrm{MSE}_{\mathcal{H}}^{\mathrm{X}} = \sigma^2\sum_j s_j z_j^2$ with z_j i.i.d. $\mathcal{N}(0,1)$, $s_j \geq 0$ and $\sigma^2\sum_j s_j = \mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{X}}]$. Hence $\mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{X}}] = c_{\mathrm{X}}\sqrt{\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{X}}]}$ with $c_{\mathrm{X}} \coloneqq \mathbb{E}\sqrt{\sum_j s_j z_j^2}/\sqrt{\sum_j s_j} \in [\sqrt{2/\pi}, 1]$: the upper end is Jensen's inequality, and the lower end holds because $\mathbf{s} \mapsto \mathbb{E}\sqrt{\sum_j s_j z_j^2}$ is concave on the simplex $\sum_j s_j = 1$, so its minimum is at a vertex, where it equals $\mathbb{E}|z_1| = \sqrt{2/\pi}$. Therefore

$$\frac{\mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{CO}}]}{\mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{NC}}]} \;\geq\; \sqrt{\frac{2}{\pi}\,\frac{\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{CO}}]}{\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{NC}}]}}, \tag{83}$$

and an expected-MSE ratio above $\pi/2$ orders the expected RMSEs. By (79), with $\mathcal{S} = \mathcal{H}$,

$$\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{CO}}] \geq \sigma^2\,\varepsilon^{-2}\,\Lambda(\mathcal{H})\,G\,\big(1+O(\varepsilon)\big). \tag{84}$$

Since at least one holdout point has $k_i' \neq k$, $[1-(k/k_i')^{\beta_{\mathrm{eff}}}]^2 > 0$ for that point, so $\Lambda(\mathcal{H}) > 0$ by Definition 8. Therefore $\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{CO}}] = \Omega(\sigma^2\varepsilon^{-2})$.

Since $\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{NC}}] = O(\sigma^2)$ by (82), and $\varepsilon^{-2} \gg 1$:

$$\mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{NC}}] < \mathbb{E}[\mathrm{MSE}_{\mathcal{H}}^{\mathrm{CO}}], \quad \text{hence} \quad \mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{NC}}] < \mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{CO}}], \tag{85}$$

the second by (83), since the expected-MSE ratio is $\Omega(\varepsilon^{-2})$ and so exceeds $\pi/2$ for small ε. Writing the holdout coefficient of determination as $R^2_{\mathcal{H}} = 1 - \mathrm{MSE}_{\mathcal{H}}/\mathrm{Var}(\mathcal{H})$ with $\mathrm{Var}(\mathcal{H})$ fixed across designs, $\mathbb{E}[\mathrm{MSE}^{\mathrm{NC}}_{\mathcal{H}}] < \mathbb{E}[\mathrm{MSE}^{\mathrm{CO}}_{\mathcal{H}}]$ implies $\mathbb{E}[R^{2\,\mathrm{NC}}_{\mathcal{H}}] > \mathbb{E}[R^{2\,\mathrm{CO}}_{\mathcal{H}}]$ at leading order (ignoring $O(\sigma^4)$ ratio-of-expectations corrections).

Misspecification extension. Under model misspecification with design-independent bias $B^2 > 0$, the strict ordering of the expected MSE behind Theorem 1 is preserved. For $K = 1$, Proposition 1 gives $V_1 = 0 < \tau_1$, so (2) never holds; in the well-specified case $B^2 = 0$, $\mathbb{E}[\mathrm{MSE}^{\mathrm{CO}}_{\mathcal{H}}]/\mathbb{E}[\mathrm{MSE}^{\mathrm{NC}}_{\mathcal{H}}] = \Omega(\varepsilon^{-2})$, and for $B^2 > 0$ the same rate appears in the estimation-variance pieces in (88). The total expected-MSE ratio takes the form

$$\frac{\mathbb{E}[\mathrm{MSE}^{\mathrm{CO}}_{\mathcal{H}}]}{\mathbb{E}[\mathrm{MSE}^{\mathrm{NC}}_{\mathcal{H}}]} \;=\; \frac{B^2 + \Omega(\sigma^2\,\varepsilon^{-2})}{B^2 + O(\sigma^2)}, \tag{86}$$

which is $\Omega(\varepsilon^{-2})$ as $\varepsilon \to 0$ at any fixed B^2, since the CO variance grows without bound; at a fixed gap, the ratio is near 1 when B^2 dominates the CO estimation variance. For $K \geq 2$, the theorem asserts no ordering; under the margin of Lemma 3, $V^{\mathrm{CO}}_{\mathcal{H}} > V^{\mathrm{NC}}_{\mathcal{H}}$ in (88), so the ratio exceeds 1 for every $B^2 \geq 0$, without an ε-rate.

Derivation. Suppose the true data-generating process is $L_i = f(N_i, D_i) + \eta_i$ where f does not lie in the model family $\{\hat{L}(\cdot;\boldsymbol{\theta})\}$. Let $\boldsymbol{\theta}^*$ denote the pseudo-true parameters minimizing the population risk. The total MSE at a test point decomposes as:

$$\mathbb{E}[(\hat{L} - f)^2] = \underbrace{(\hat{L}(\cdot;\boldsymbol{\theta}^*) - f)^2}_{=:\, b^2 \text{ (approximation bias)}} + \underbrace{\sigma^2\,\mathbf{j}^T(J_T^T J_T)^{-1}\mathbf{j}}_{\text{estimation variance}}, \tag{87}$$

where we assume that the cross-term $\mathbb{E}\big[\delta\boldsymbol{\theta}^T\mathbf{j}\,(\hat{L}(\cdot;\boldsymbol{\theta}^*) - f)\big]$ vanishes to leading order under the Gauss-Newton linearization; this, like a design-independent B^2, is an assumption, since the pseudo-true parameters can depend on the training design.

Averaging over a test set and writing $B^2 = |\mathcal{S}|^{-1}\sum_i b_i^2$ for the mean squared approximation bias:

$$\mathbb{E}[\mathrm{MSE}^X_{\mathcal{S}}] = B^2 + \sigma^2\,V^X_{\mathcal{S}}, \tag{88}$$

where $V^X_{\mathcal{S}}$ is the mean predictive variance from (76) for design X.

For a general holdout $\mathcal{H}$ with $\Lambda(\mathcal{H}) > 0$: $V^{\mathrm{CO}}_{\mathcal{H}} = \Omega(\varepsilon^{-2})$ and $V^{\mathrm{NC}}_{\mathcal{H}} = O(1)$ by (84) and (82), which gives (86); its numerator exceeds its denominator for small ε at every $B^2 \geq 0$.

Regime B: several rays. With $K \geq 2$ distinct ratios the CO design satisfies the non-proportionality condition (80), so $\|\mathbf{j}_B - c\,\mathbf{j}_A\| = \Theta(1)$ for all c, whether or not $V_K \geq \tau_K$; the threshold only bounds $\kappa_{A,B}$ by κ_{target} (Proposition 1). By the Regime B hypothesis, J_T has full rank at $\varepsilon = 0$, which also covers the exponent columns, so by continuity all eigenvalues of $J_T^T J_T$ are $\Theta(1)$ as $\varepsilon \to 0$, and for each holdout row $\|\mathbf{j}\|^2/\lambda_{\max} \leq \mathbf{j}^T(J_T^T J_T)^{-1}\mathbf{j} \leq \|\mathbf{j}\|^2/\lambda_{\min}$ with both bounds $\Theta(1)$ (a scale-coefficient entry $N'^{-\alpha}$ of $\mathbf{j}$ is nonzero), and (76) with (83) gives $\mathbb{E}[\mathrm{RMSE}^{\mathrm{CO}}_{\mathcal{H}}] = \Theta(\sigma)$. The same two-sided bound holds for NC, whose eigenvalues are $\Theta(1)$ (paragraph *NC Gram matrix*), so the expected-RMSE ratio is:

$$\frac{\mathbb{E}[\mathrm{RMSE}^{\mathrm{CO}}_{\mathcal{H}}]}{\mathbb{E}[\mathrm{RMSE}^{\mathrm{NC}}_{\mathcal{H}}]} = \Theta(1), \tag{89}$$

with the $\Theta(1)$ constant depending on the stiff-mode prefactors of each design. □

Lemma 3 (Collinear designs with $K \geq 2$)**.** *Assume the well-specified setup of this appendix (i.i.d. Gaussian noise and the Gauss-Newton linearization), and let CO be a collinear design with $K \geq 2$ rays and NC any design, both with full-rank Jacobians. Let $V^{\mathrm{X}}_{\mathcal{H}}$ be the mean predictive variance of design X in (88), so that $\mathbb{E}[\mathrm{MSE}^{\mathrm{X}}_{\mathcal{H}}] = \sigma^2 V^{\mathrm{X}}_{\mathcal{H}}$ in the well-specified case. Let $G_{A,B}$ be the Gram block of the two scale-coefficient columns of J_{CO} (Table 9), let $M := |\mathcal{H}|^{-1}\sum_i (\mathbf{j}_i)_{A,B}(\mathbf{j}_i)^T_{A,B}$ be the holdout's second moment of the same two entries, and define*

$$\rho_K := \frac{\kappa_{\mathrm{target}}\det G_{A,B}}{(\mathrm{tr}\,G_{A,B})^2}, \qquad \rho^\star(\mathcal{H}) := \frac{\kappa_{\mathrm{target}}\big[\mathrm{tr}\,G_{A,B}\,\mathrm{tr}\,M - \mathrm{tr}(G_{A,B}M)\big]}{(\mathrm{tr}\,G_{A,B})^2\,V^{\mathrm{NC}}_{\mathcal{H}}}.$$

If $\rho_K < \rho^\star(\mathcal{H})$, then $\mathbb{E}[\mathrm{MSE}^{\mathrm{NC}}_{\mathcal{H}}] < \mathbb{E}[\mathrm{MSE}^{\mathrm{CO}}_{\mathcal{H}}]$; if $\rho^\star(\mathcal{H}) > (\pi/2)\,\rho_K$, then also $\mathbb{E}[\mathrm{RMSE}^{\mathrm{NC}}_{\mathcal{H}}] < \mathbb{E}[\mathrm{RMSE}^{\mathrm{CO}}_{\mathcal{H}}]$.

Proof. Order the parameters with the scale pair first, so that $J_{\mathrm{CO}}^T J_{\mathrm{CO}} = \left(\begin{smallmatrix} G_{A,B} & Q_{12} \\ Q_{12}^T & Q_{22} \end{smallmatrix}\right)$, which is positive definite. Let $\Gamma := Q_{22} - Q_{12}^T G_{A,B}^{-1} Q_{12}$, positive definite as a Schur complement, and let $W := \left(\begin{smallmatrix} -G_{A,B}^{-1} Q_{12} \\ I \end{smallmatrix}\right)$. Multiplying out verifies $(J_{\mathrm{CO}}^T J_{\mathrm{CO}})^{-1} = \left(\begin{smallmatrix} G_{A,B}^{-1} & 0 \\ 0 & 0 \end{smallmatrix}\right) + W \Gamma^{-1} W^T$, so every holdout row satisfies $\mathbf{j}^T (J_{\mathrm{CO}}^T J_{\mathrm{CO}})^{-1} \mathbf{j} \geq \mathbf{j}_{A,B}^T G_{A,B}^{-1} \mathbf{j}_{A,B}$. Averaging over $\mathcal{H}$ and using $G_{A,B}^{-1} = (\operatorname{tr} G_{A,B}\, I - G_{A,B}) / \det G_{A,B}$ for a 2×2 matrix gives

$$V_{\mathcal{H}}^{\mathrm{CO}} \geq \operatorname{tr}(G_{A,B}^{-1} M) = \frac{\operatorname{tr} G_{A,B} \operatorname{tr} M - \operatorname{tr}(G_{A,B} M)}{\det G_{A,B}} = \frac{\rho^\star(\mathcal{H})}{\rho_K} V_{\mathcal{H}}^{\mathrm{NC}}.$$

If $\rho_K < \rho^\star(\mathcal{H})$, this gives $V_{\mathcal{H}}^{\mathrm{CO}} > V_{\mathcal{H}}^{\mathrm{NC}}$, and multiplying by σ^2 gives the MSE claim. If $\rho^\star(\mathcal{H}) > (\pi/2)\rho_K$, the expected-MSE ratio exceeds $\pi/2$, and (83) gives the RMSE claim. □

The ratio $\rho^\star/\rho_K = \operatorname{tr}(G_{A,B}^{-1} M)/V_{\mathcal{H}}^{\mathrm{NC}}$, whose size above 1 (resp. $\pi/2$) is sufficient for the two claims but not necessary, depends neither on κ_{target} nor on how the two scale-coefficient columns are scaled; ρ_K and $\rho^\star$ separately do. For the Chinchilla columns $N^{-\alpha}$ and $D^{-\beta}$, eqns. (65)-(67) give $\rho_K = V_K/\tau_K$ at leading order, which treats $N_i^{\tilde{\varepsilon}}$ as $1 + O(\varepsilon)$ although $\varepsilon \ln N_i$ is not small at our model sizes. With $\lambda_\pm$, $\mathbf{u}_\pm$ the eigenpairs of $G_{A,B}$, the numerator bracket is $\lambda_+ \mathbf{u}_-^T M \mathbf{u}_- + \lambda_- \mathbf{u}_+^T M \mathbf{u}_+$, so $\rho^\star$ weighs the K-ray sloppy leverage $\mathbf{u}_-^T M \mathbf{u}_-$ against the predictive variance of NC, which need not satisfy $V_K \geq \tau_K$.

Our trained collinear designs fall outside this margin (`tools/regime_a_margin.py`, at the seed-0 NC fit of each corpus, law and epoch mode): on the 12-ray fan the ratio $\rho^\star/\rho_K$ is 0.14 to 0.43 in all 50 cells, and it is below 1 for all 1,347,390 budget-matched CO-NC pairs with $K \geq 2$ behind Table 7 (3 laws, 5 corpora, 22 bounding-box seeds). At the same $\boldsymbol{\theta}^\star$, the Gauss-Newton ordering itself, $V_{\mathcal{H}}^{\mathrm{CO}} > V_{\mathcal{H}}^{\mathrm{NC}}$, holds in all 50 fan cells (ratio 1.09 to 4.82) but, per law, in only 2 to 51% of those pairs whose bounding-box NC design has full rank, and in 42 to 98% of such single-ray pairs; at the fit of Hoffmann et al. (2022) ($\varepsilon = 0.06$), in 626 of the 628 full-rank Chinchilla single-ray pairs. For Chinchilla and Droppo-Elibol, 692 of the 1,320 single-ray pairs each have a rank-deficient NC design (column-equilibrated condition number above 10^{10}), outside the theorem's NC hypothesis, and the single-ray fits of Table 7 include them. Apart from those, the single-ray reversals occur at the fitted parameters, whose exponent gaps are far from the $\varepsilon \to 0$ regime of Theorem 1 (at Hoffmann et al.'s parameters the reversals nearly all vanish, but all five parameters change), while for $K \geq 2$ the variance ordering holds in only 5% of the full-rank Chinchilla pairs even at the fit of Hoffmann et al. (2022), so the NC wins of Table 7 are an empirical finding that the variance mechanism alone does not explain.

B.14 Appendix - Proof of Proposition 3 (RMSE invariance)

By Definition 4, $C_i = 6\, N_i\, D_i$ for each $i \in \{1, \ldots, n_{\mathcal{H}}\}$. Substituting into the isoFLOP evaluation:

$$\frac{C_i}{6\, N_i} \;=\; \frac{6\, N_i\, D_i}{6\, N_i} \;=\; D_i. \tag{90}$$

Therefore, for every holdout index i:

$$\hat{L}\big(N_i,\; C_i/(6N_i);\; \boldsymbol{\theta}\big) \;=\; \hat{L}(N_i,\; D_i;\; \boldsymbol{\theta}). \tag{91}$$

Each summand in the RMSE is identical under both parameterizations; hence the realized RMSE on $\mathcal{H}$ coincides and $\mathbb{E}[\mathrm{RMSE}_{\mathcal{H}}]$ agrees under either evaluation map. □

B.15 Appendix - General recipe for TPP line selection

We now generalize from the two-TPP rule-of-thumb to a recipe for selecting the number K, placement $\{k_1, \ldots, k_K\}$, and per-ratio allocation $\{m_1, \ldots, m_K\}$ of TPP lines given three inputs:

- a target condition number κ_{target} (e.g., 10^2),
- the effective data-size exponent β_{eff} (e.g., 0.28 for Chinchilla),
- a total observation budget m (total number of (N, D) configurations).

Law-specific exponent substitutions. The recipe uses one law-specific quantity, the effective data-size exponent β_{eff}. Table 9 gives it for each of the four scaling laws, together with the exponent gap ε.

Table 9: Exponent substitutions for each scaling law. β_{eff} is the exponent on the token/data-count term of the collinear Jacobian pair, and the exponent in V_K and τ_K of (93). For Kaplan and Droppo-Elibol the threshold is a heuristic (Appendix B.12).

Law	**Collinear pair**	ε	β_{eff}
Chinchilla	$(A,\, B)$	$\lvert\alpha - \beta\rvert$	β
Repeated-data	$(A,\, B)$	$\lvert\alpha - \beta\rvert$	β
Kaplan	$(N_c,\, D_c)$	$\lvert 1 - \alpha_N/\alpha_D\rvert$	1
Droppo-Elibol	$(N_C,\, D_C)$	$\lvert\gamma_N - \gamma_D\rvert$	γ_D

General diversity condition ($K \geq 2$). For $K \geq 2$ rays with ratios $k_1 < \cdots < k_K$, the leading-order Gram sub-block of the two scale-coefficient columns (Table 9; Chinchilla (A, B), Kaplan (N_c, D_c), etc.) satisfies $\operatorname{tr}(G_{A,B}) \approx \Phi_2(K + \sum_\ell k_\ell^{-2\beta_{\text{eff}}})$ and $\det(G_{A,B}) \approx K^2\Phi_2^2\, V_K$, where $\Phi_2 \coloneqq \sum_i N_i^{-2\alpha}$ with α the model-side exponent in the Jacobian column paired with β_{eff} (Table 9; Chinchilla α, Kaplan α_N/α_D, Droppo-Elibol γ_N) and the TPP diversity is

$$V_K \;\coloneqq\; \frac{1}{K}\sum_\ell k_\ell^{-2\beta_{\text{eff}}} - \Big(\frac{1}{K}\sum_\ell k_\ell^{-\beta_{\text{eff}}}\Big)^2. \tag{92}$$

The condition number is therefore $\kappa_{A,B} \approx \operatorname{tr}^2/\det = \Phi_2^2(K + \sum_\ell k_\ell^{-2\beta_{\text{eff}}})^2/(K^2\Phi_2^2 V_K)$; the Φ_2^2 factors cancel. Hence $\kappa_{A,B} \leq \kappa_{\text{target}}$ holds if (and, up to the factor $(1 + 1/\kappa_{\text{target}})^{-2}$ of Step 4 in Appendix B.12, only if)

$$V_K \;\geq\; \tau_K \;\coloneqq\; \frac{(K + \sum_\ell k_\ell^{-2\beta_{\text{eff}}})^2}{K^2\,\kappa_{\text{target}}}. \tag{93}$$

We evaluate (93) with raw counts: N in parameters and D in tokens, so $k = D/N$ is in tokens per parameter (Appendix C keeps three published designs in their own data units). A condition number changes when a column is rescaled, so the values of $\kappa_{A,B}$ and the verdicts of (93) are relative to this convention. For $K = 2$ with $k_2 = Rk_1$, (93) holds iff $R \geq R_{\min}(k_1)$, given in closed form by (71) in the proof of Proposition 1. $R_{\min}(k_1)$ depends on k_1, β_{eff} and κ_{target}, increases with $k_1 \geq 1$ (a B-column scale effect, Step 1), and is infinite once $\sqrt{\kappa_{\text{target}}}\, k_1^{-\beta_{\text{eff}}} \leq 2 + k_1^{-2\beta_{\text{eff}}}$.

Step 1: determine the minimum spread (for $K = 2$). We write κ_K for $\kappa_{A,B}$ on a collinear training design with $K \geq 1$ TPP ratios (Sec. 3.1; Proposition 1). A single ratio does not suffice: $V_1 = 0 < \tau_1$, and $\kappa_1 = \Theta(\varepsilon^{-2})$ for every k_1 (Proposition 2). Choose the lower ratio k_1 and read $R_{\min}(k_1)$ from Table 10 at $\kappa_{\text{target}} = 10^2$, or compute it from (71) for other targets; then $R \geq R_{\min}(k_1)$ gives $\kappa_2 \leq \kappa_{\text{target}}$ at leading order. If $\sqrt{\kappa_{\text{target}}}\, k_1^{-\beta_{\text{eff}}} \leq 2 + k_1^{-2\beta_{\text{eff}}}$ (at $\kappa_{\text{target}} = 10^2$, $k_1 \geq k_1^{\max}$ in Table 10), no spread reaches κ_{target} at this k_1 at leading order: relax κ_{target} or let the exact check of Step 5 decide. At a fixed R, k_1 changes only the scale of the B column, so $\operatorname{Var}(\hat{A})$ and the correlation of $(\hat{A}, \hat{B})$ do not depend on it; lowering k_1 changes $\kappa_{A,B}$, not what the design identifies.

For $K > 2$ at the same endpoints k_1 and k_K, interior rays lower V_K, while τ_K depends on the rays only through the mean of the $k_\ell^{-2\beta_{\text{eff}}}$. By (72), V_K/τ_K exceeds its two-endpoint value by at most the factor $(1 + k_1^{-2\beta_{\text{eff}}}/2)^2$, and in the worked example below a middle ray lowers it from 1.13 to 0.785. Rays beyond two serve lack-of-fit checks and outlier diagnostics (Step 2). Screen any specific K-ray design with (93) and decide by its exact $\kappa_{A,B}$ (Step 5).

Table 10: Smallest two-ratio spread $R_{\min}(k_1)$ of (71), with $R = k_2/k_1$, that satisfies (93) at $\kappa_{\text{target}} = 10^2$, by lower ratio k_1. All entries are rounded up. The k_1 dependence reflects the scale of the B column only. At leading order no spread suffices once $k_1 \geq k_1^{\max}$ (last row). At $\kappa_{\text{target}} = 10$, feasibility needs $k_1 < 1.41$ to 1.62, and $k_1 = 1$ needs $R > 10^3$.

k_1	$\beta_{\text{eff}} = 0.40$	$\beta_{\text{eff}} = 0.35$	$\beta_{\text{eff}} = 0.28$
1	2.9	3.4	4.5
5	4.5	5.0	6.3
20	19.7	14.9	13.4
50	1.3×10^4	120	37.0
$k_1^{\max}$	54	94	292

Sensitivity to exponent values and noise. $R_{\min}(k_1)$ depends on k_1, β_{eff} and κ_{target}, and *not* on ε or on the noise level σ^2. The condition number κ is a property of the Jacobian J^TJ, which is determined by the design and the functional form, not by the residual variance. Noise affects *parameter variance* ($\text{Var}(\hat{\boldsymbol{\theta}}) \propto \sigma^2(J^TJ)^{-1}$) but not the design threshold for a given κ_{target}. To achieve a target *coefficient precision* (e.g., a desired width of the confidence interval for A or B), one must increase the number of observations m when noise is high; the required R stays the same.

The direction of the β_{eff} effect depends on k_1 (Table 10). At $1 \leq k_1 \leq 5$ a smaller β_{eff} needs a larger R (at $k_1 = 1$, 4.5 at $\beta_{\text{eff}} = 0.28$ against 2.9 at 0.40); at $k_1 = 50$ a larger β_{eff} does, and at $k_1 = 20$ it does above $\beta_{\text{eff}} \approx 0.29$, where $R_{\min}$ is smallest (19.7 at 0.40 against 13.4 at 0.28). When the prior on β_{eff} is rough, use the largest entry over its plausible range.

Adaptive adjustment of R. We recommend an adaptive scheme when the design is built sequentially:

1. choose the initial R as the largest $R_{\min}(k_1)$ over the plausible β_{eff} range;
2. after fitting on the first batch, obtain an updated $\hat{\beta}_{\text{eff}}$;
3. recompute $R_{\min}(k_1)$ from (71) at $\hat{\beta}_{\text{eff}}$;
4. if (93) fails at $\hat{\beta}_{\text{eff}}$, add augmentation runs at a second TPP ratio k' satisfying the updated $R_{\min}$ (or increase the spread of existing TPP lines if the design is not yet fixed), and check the final design by its exact $\kappa_{A,B}$ (Step 5).

This refines the design as exponent estimates sharpen, avoiding both over-investment (overly large R when priors were pessimistic) and under-design (insufficient R when priors were optimistic).

Simple rule. Given a rough prior range for β_{eff}, the planned lower ratio k_1 and κ_{target}:

1. read $R_{\min}(k_1)$ from Table 10 at $\kappa_{\text{target}} = 10^2$, taking the largest entry over the plausible β_{eff} range; for another k_1 or target, compute it from (71) instead of interpolating;
2. round up: $R \geq \lceil R_{\min} \rceil$ (e.g., $R_{\min} = 14.9 \Rightarrow R \geq 15$);
3. screen the final ratios with (93) and decide by their exact $\kappa_{A,B}$ (Step 5).

Step 2: decide K (number of TPP lines). Two TPP lines ($K = 2$) already identify the parameters when J has full rank (Theorem 1, Regime B); $R \geq R_{\min}(k_1)$ brings $\kappa_{A,B}$ below κ_{target} at leading order ($\kappa(J^TJ)$ is not bounded). Additional lines $K > 2$ at the same endpoints do not improve conditioning beyond the factor of (72) and can lower V_K/τ_K (Step 1). They offer two practical benefits: (a) they enable *lack-of-fit* tests (degrees of freedom for model diagnostics), and (b) they reduce sensitivity to outlier runs at any single ratio. A useful rule:

- $K = 2$: minimum for identifiability;
- $K = 3$: recommended when the budget permits, to enable residual diagnostics; placed log-uniformly by (94), it needs a wider spread than $K = 2$;
- $K \geq 4$: useful for meta-analytic settings or when the functional form is uncertain.

Step 3: place the TPP ratios. For $K = 2$, place k_1 and k_2 as far apart as feasible (subject to $R \geq R_{\min}(k_1)$, Step 1); endpoint placement maximizes V_K at fixed support. For $K \geq 3$ at the

same endpoints, interior rays lower V_K, and no placement beats the $K = 2$ design by more than the factor of (72) (Step 1). When intermediate ratios are desired for diagnostic value (lack-of-fit, residual checks across TPP), a log-uniform grid over $[k_1, k_K]$ is a natural choice:

$$k_j = k_1 \, R^{(j-1)/(K-1)}, \qquad j = 1, \ldots, K, \tag{94}$$

at the cost of a lower V_K than the endpoint-only design; screen it with (93) and decide by its exact $\kappa_{A,B}$ (Step 5).

Step 4: allocate runs across ratios. Given a total budget of m observations split into K groups of m_j observations each ($\sum_j m_j = m$), the D-optimal allocation for the scale-coefficient sub-block is approximately *balanced*: $m_j \approx m/K$. The intuition is that the sloppy direction receives information from the *contrast* between TPP groups, and this contrast is maximized when each group has equal weight. When the groups have different model-size ranges, slight rebalancing toward the group with smaller spread is beneficial. Eq. (93) assumes equal runs per ratio at the same model sizes (Appendix B.12); with unequal runs, the rays enter $G_{A,B}$ with unequal weights.

Within each TPP group, the model sizes should span as wide a range as possible and be approximately log-uniformly spaced:

$$N_{j,\ell} = N_{\min} \left(N_{\max}/N_{\min}\right)^{(\ell-1)/(m_j-1)}, \qquad \ell = 1, \ldots, m_j. \tag{95}$$

Step 5: verify conditioning. Before committing to expensive large-scale runs, compute the exact $\kappa_{A,B}$ of the planned design: the condition number of the 2×2 Gram block of the columns $N_i^{-\alpha}$ and $D_i^{-\beta_{\text{eff}}}$ over the planned (N_i, D_i), with raw counts (N in parameters, D in tokens) and priors on α and β_{eff}. This is the decision rule: (93), the criterion of Proposition 1 and of Table 1, is its leading-order screen and can misjudge it several-fold either way (Appendix B.12), because the leading order sets $N^{\alpha-\beta_{\text{eff}}}$ to 1, while over $N \in [10^7, 10^9]$ at $\alpha - \beta_{\text{eff}} = 0.06$ it runs from 2.6 to 3.5. If the exact check fails, increase R (Step 1) and repeat. Replicating every run equally does not change $\kappa_{A,B}$ - it scales $J^T J$ uniformly and leaves the condition number invariant (replicating some runs reweights the rays, Step 4); it improves *precision* ($\mathrm{Var}(\hat{\boldsymbol{\theta}}) \propto \sigma^2 (J^T J)^{-1}$) but not identifiability.

Worked example. We work through the 20-run design $k \in \{20, 400\}$ of Section 5, which passes (93) there at $\beta_{\text{eff}} = 0.28$ ($V_2/\tau_2 = 1.22$), at its own priors here. Target $\kappa_{\text{target}} = 10^2$, Chinchilla-like priors ($\beta_{\text{eff}} = 0.35$, $\alpha = \beta_{\text{eff}} + 0.06 = 0.41$, $p = 5$).

1. $k_1 = 20$ (the Chinchilla baseline), so $R_{\min} = 14.9$ (Table 10).
2. $K = 2$ (minimum budget).
3. $k_2 = 400$ ($R = 20$).
4. Budget $m = 20$ runs: $m_1 = m_2 = 10$.
5. Model sizes $N \in [10^7, 10^9]$, log-uniformly spaced within each group.
6. Screen: $V_2/\tau_2 = 1.13 \geq 1$, so the design passes (93); equivalently, the leading-order $\mathrm{tr}^2/\det$ of $G_{A,B}$ is $88 \leq 10^2$.
7. Decide (Step 5): the exact $\kappa_{A,B}$ at these sizes is $21.0 \leq 10^2$, so the design passes.

The ratios $k \in \{20, 100\}$ ($R = 5$) fail (93) ($V_2/\tau_2 = 0.49$). With $K = 3$ and equal runs per ratio (Step 4), a middle ratio $k = 89$ (the geometric mean of 20 and 400) gives a lack-of-fit check, but $V_3/\tau_3 = 0.785 < 1$ fails (93): the log-uniform $K = 3$ design (94) at $k_1 = 20$ needs $R \geq 40.3$ at this β_{eff}, against 14.9 at $K = 2$. These are leading-order verdicts. The exact $\kappa_{A,B}$ at the planned sizes (Step 5), which decides, accepts both: it is 55.0 for $\{20, 100\}$ and 28.7 for the $K = 3$ design. It depends on the prior on α: at $\alpha = \beta_{\text{eff}} = 0.35$ the three designs give 86.2, 203.3 and 125.4, and only $\{20, 400\}$ passes.

B.16 Appendix - Interaction-term scaling laws

The Chinchilla and repeated-data laws are *additive* in the N- and D-contributions (Kaplan and Droppo-Elibol compose them inside a power). Adding an explicit N-D interaction $F\, N^{-\gamma_N} D^{-\gamma_D}$ (as in Farseer (Li et al., 2025a)) does not alleviate the degeneracy: under $D = k_\ell N$ the interaction

column becomes a third near-proportional power of N, so when all exponent gaps are small the (A, B, F) sub-block has *two* near-zero eigenvalues. We derive this formally below.

Representative model. Consider a Chinchilla-style law augmented with a multiplicative interaction (with $p = 8$ parameters $\boldsymbol{\theta} = [A, B, F, E, \alpha, \beta, \gamma_N, \gamma_D]^T$ or fewer if some exponents are tied):

$$\hat{L}(N, D) = A\,N^{-\alpha} + B\,D^{-\beta} + F\,N^{-\gamma_N}\,D^{-\gamma_D} + E, \tag{96}$$

$$\mathbf{j}_{A,i} = -N_i^{-\alpha}, \tag{97}$$

$$\mathbf{j}_{B,i} = -k_\ell^{-\beta}\,N_i^{-\beta} = -k_\ell^{-\beta}\,N_i^{-\alpha}\,N_i^{\varepsilon_1}, \tag{98}$$

$$\mathbf{j}_{F,i} = -k_\ell^{-\gamma_D}\,N_i^{-(\gamma_N+\gamma_D)} = -k_\ell^{-\gamma_D}\,N_i^{-\alpha}\,N_i^{\varepsilon_2}. \tag{99}$$

The interaction coefficient F and its associated exponents (γ_N, γ_D) introduce a third "capacity $\times$ data" channel. We use residual Jacobian columns $J_{i\theta} = -\partial \hat{L}_i/\partial\theta$ as in (10). Substituting $D_i = k_\ell N_i$, the three scale-coefficient columns are the last three rows of the display above, where $\varepsilon_1 = \alpha - \beta$ and $\varepsilon_2 = \alpha - (\gamma_N + \gamma_D)$. All three columns share the base function $N_i^{-\alpha}$, differing only by perturbations $N_i^{\varepsilon_1}$ and $N_i^{\varepsilon_2}$.

Near-proportionality of the Jacobian columns. Writing $N_i^{\varepsilon_j} = 1 + \varepsilon_j \log N_i + O(\varepsilon_j^2)$ for small $|\varepsilon_j|$, we obtain $\mathbf{j}_B = c_1\,\mathbf{j}_A + \boldsymbol{\delta}_1$ and $\mathbf{j}_F = c_2\,\mathbf{j}_A + \boldsymbol{\delta}_2$, where $c_1 = k_\ell^{-\beta}$, $c_2 = k_\ell^{-\gamma_D}$, and

$$\delta_{1,i} = -c_1\,N_i^{-\alpha}\,(N_i^{\varepsilon_1} - 1) = O(\varepsilon_1), \qquad \delta_{2,i} = -c_2\,N_i^{-\alpha}\,(N_i^{\varepsilon_2} - 1) = O(\varepsilon_2). \tag{100}$$

Thus $\|\boldsymbol{\delta}_1\| = O(\varepsilon_1)$ and $\|\boldsymbol{\delta}_2\| = O(\varepsilon_2)$ - the same near-proportionality structure as the additive case (Appendix B.5), but now with *three* columns all proportional to $\mathbf{j}_A$ up to $O\big(\max\{|\varepsilon_1|, |\varepsilon_2|\}\big)$ perturbations. Under $D = k_\ell N$, the interaction column $\mathbf{j}_F$ is therefore nearly proportional to both $\mathbf{j}_A$ and $\mathbf{j}_B$, generalizing the additive degeneracy to coupled laws.

3×3 **sub-block analysis.** The Gram matrix of the (A, B, F) sub-block has entries:

$$G_{jk} = \sum_{i=1}^{n} \mathbf{j}_{j,i}\,\mathbf{j}_{k,i}, \qquad j, k \in \{A, B, F\}. \tag{101}$$

Write $\mathbf{j}_B = c_1\,\mathbf{j}_A \circ \mathbf{h}_1$ and $\mathbf{j}_F = c_2\,\mathbf{j}_A \circ \mathbf{h}_2$ where $\circ$ denotes element-wise product, $h_{1,i} = N_i^{\varepsilon_1}$, $h_{2,i} = N_i^{\varepsilon_2}$, $c_1 = k_\ell^{-\beta}$, $c_2 = k_\ell^{-\gamma_D}$. When $\varepsilon_1, \varepsilon_2 \to 0$, $\mathbf{h}_1, \mathbf{h}_2 \to \mathbf{1}$ and the three columns become proportional, giving $\operatorname{rank}(G) = 1$. More precisely:

$$\det(G_{A,B,F}) = O(\varepsilon_1^2\,\varepsilon_2^2) + O(\varepsilon_1^2\,(\varepsilon_1 - \varepsilon_2)^2), \tag{102}$$

$$\varepsilon_{\text{gap}} \coloneqq \min\big(|\alpha - \beta|,\, |\alpha - (\gamma_N{+}\gamma_D)|,\, |(\gamma_N{+}\gamma_D) - \beta|\big). \tag{103}$$

The 3×3 Gram determinant is controlled by the *pairwise* and *triple* volume elements among the three nearly-proportional vectors. The smallest eigenvalue of $G_{A,B,F}$ is at most that of its closest pair, $O(\varepsilon_{\text{gap}}^2)$; the second smallest is $O(\max\{\varepsilon_1^2, \varepsilon_2^2\})$, since all three columns lie within that distance of $\operatorname{span}(\mathbf{j}_A)$.

Condition number. The 3×3 block inherits the Gram-determinant scaling (102); combining with eigenvalue trace-determinant relations for positive semidefinite matrices (as in the proof of Proposition 2) shows that $\kappa(J^T J)$ grows at least as $\Omega\big(\varepsilon_{\text{gap}}^{-2}\big)$ for ε_{gap} from (103), and can be *larger* than the pure (A, B) Chinchilla rate when multiple exponent gaps are small. Thus the problem is *worse* than the additive case: the interaction term introduces a *second* sloppy direction (the null space of the 3×3 sub-block is two-dimensional when all three exponent gaps vanish), requiring two additional constraints (e.g., two distinct TPP ratios and a third quantity) for full identifiability.

Implications. Adding interaction terms does not cure collinearity - it exacerbates it. Under $D = k_\ell N$, the interaction coefficient F is confounded with both A and B: the data cannot distinguish whether a given loss level comes from model capacity, data volume, or their interaction. The design-of-experiments prescription is unchanged: diverse TPP ratios are essential, and with an interaction term even more diversity may be needed ($K \geq 3$ TPP ratios to resolve the two-dimensional degeneracy).

C Appendix - published scaling-law designs classified

This appendix applies the a-priori test $V_K \geq \tau_K$ of Eq. (2) to every design in a fixed set of 206 candidate papers whose training ratios can be recovered (Table 11). The protocol was fixed before

any candidate was collected and amended five times after audits of its verdicts; each amendment and its effect on the counts is logged in `CRITERIA.md`, next to the code, candidate list and per-design records in the released repository (`camera_ready/grid_classification`).

Screening. The candidates and the reasons for excluding them are:

Candidates	206	Excluded	182
surveyed by Li et al. (2025b)	51	law in one variable	102
pipelines this paper cites	6	no parametric law	66
arXiv title searches	149	duplicate entry	6
		ratios not recoverable	7
		protocol not recoverable	1
Classified	24		

The searches are three fixed arXiv title searches for scaling-law papers, top-ranked hits first, and also return papers from other fields. The 32 designs that fit a bivariate law in N and D use one of the four laws of Table 2 (Proposition 1 is proved for Chinchilla and unsaturated repeated-data and is a heuristic for the other two), or a superset that adds terms to one of them; for a superset we apply the test to that core, which is a heuristic. 52 of the 206 decisions rest on the abstract alone. No design was included or excluded because of its verdict. Cerebras-GPT (Dey et al., 2023) fits only a loss-versus-compute frontier and is not classified; a bivariate fit on its single ray $D \approx 20N$ would have $K = 1$ and fail.

Conventions. Each verdict rests on six conventions, listed with what changes under the alternatives.

- *Units.* Eq. (2) takes raw counts (Appendix B.15): N in parameters as each paper counts them (total for the Pythia family of Choshen et al. (2025), non-embedding where a paper fits with those) and D in the data unit of the fitted law, which is tokens except words for Rosenfeld et al. (2020), bytes for Gordon et al. (2021) and characters for Tao et al. (2024). Converting these three to tokens, at any factor from 1/4 to 4 tokens per unit, changes no verdict. Rescaling every k by a factor of 2 flips two verdicts, Gadre et al. (2025) (to failing at $2k$) and Kumar et al. (2025) (to passing at $k/2$); a factor of 4 also flips Sardana et al. (2024) and Sun et al. (2025).
- *Merging.* Adjacent ratios within 1% are merged to their geometric mean, transitively. This chains the dense checkpoint ratios of Pythia into 243 groups, so the tables print the raw range of each design rather than the merged one (for Pythia, a largest group mean of 3636 against a raw maximum of 4284). Merging only ratios within 1% of a group's first member gives margins V_K/τ_K of 5.28 for Hoffmann et al. (2022), 7.30 for Li et al. (2025a) and 2.65 for Pythia; merging nothing gives 4.20, 4.94 and 2.08, and 5.36 for Muennighoff et al. (2025). No verdict changes.
- *Weights.* Eq. (2) counts each distinct ratio once, as for a collinear design (Definition 1), which trains every size at every ratio and is the design Proposition 1 covers. Only 6 of the 24 designs are collinear, and all 6 fail: Hu et al. (2024), Bhagia et al. (2025), Volkova et al. (2026) and the three at one ratio. The other 18, all 11 passes among them, train different sizes at different ratios, so their rays enter $G_{A,B}$ with unequal weights, and the test on their distinct ratios is a heuristic, as for a superset. The proof's Gram block weights each run by $N^{-2\alpha}$, with α the model-side exponent. With these weights, on the 17 of the 18 whose runs can be rebuilt (the released runs of Muennighoff et al. (2025) give ratios, not sizes), two verdicts change. Four of the five runs of Gadre et al. (2025) lie at $k = 20$ and one at $k = 320$; its margin falls to 0.93 (the condition number of the (A, B) block rises from 94 to 108, past the target). Rosenfeld et al. (2020), whose smallest models carry most of the weight, falls from 12.8 to 0.20. No other verdict changes, Choshen et al. (2025) coming closest at 1.05, so with these weights the split would be 9 to 15 if the untested pass of Muennighoff et al. (2025) holds. Both passes are marked w.
- *Exponent.* Each design is evaluated at the data-side exponent its paper reports for its headline fit, and the five that print none use the literature value 0.28; the exponent of Li et al. (2025a), which varies with N, is taken at the geometric-mean size of its grid. The two Kaplan-form designs, Kaplan et al. (2020) and Gordon et al. (2021), take their reported outer α_D rather than the $\beta_{\text{eff}} = 1$ of Table 9, at which they have $\kappa = 50$ and 27 and still pass. This choice was an amendment: at the corpus-wide 0.28 of the original protocol, the split is 14 to 10, with Kumar et al. (2025), Xu et al.

(2026) and Ludziejewski et al. (2024) passing. Nine of the 13 failures use an exponent fitted on the very design under test; six of these nine fail at every exponent we probed (0.10 to 0.50 and the paper's own), so their verdicts do not depend on it.

- *Target.* $\kappa_{\text{target}} = 100$, the main-text value, is a cut, not a property of a design. Table 11 gives each design's own leading-order condition number of the (A, B) block, $\kappa = 100\,\tau_K/V_K$, the smallest target it passes. Nine of the ten multi-ratio failures have κ from 120 to 380 and pass at 10^4; Tao et al. (2024) has 3.2×10^4 and passes at 10^5; the three single-ratio designs fail at every target, since $V_1 = 0$. Our own CO fan of Section 4 sits at $\kappa \approx 250$ at $\beta_{\text{eff}} = 0.28$: it fails at 100 and passes at 10^4, yet NC predicts the holdout far better, and Section 4, since the theorem fixes no target, averages over targets. A larger target is not free: the anisotropy of the confidence intervals grows as $\sqrt{\kappa}$, so 100 caps it near tenfold and 10^4 to 10^6 permit hundred- to thousandfold.
- *Endpoints.* A run table is used as printed, and so is a plotted sweep: the 38 runs of Kaplan et al. (2020) are read from the vector graphics of its Figure 9. When only the ends of each size's token range are recoverable (marked e), the test uses the ratios at those ends; this applies to Shukor et al. (2025) alone. Interior ratios usually lower the margin: in 200,000 random crossed grids (2 to 6 log-even levels per axis, spans of 2 to 100 times, β_{eff} from 0.10 to 0.50), a corner pass failed with all ratios included in 26% of cases (under 1% at a corner margin of 2 or more, none at 6 or more), and the reverse happened 56 times. An earlier count crossed the size and token ranges of Ludziejewski et al. (2024) and Shukor et al. (2025) instead of their run tables, and passed both (`CRITERIA.md`, Amendment 4).

Table 11: The test applied to 24 published designs, grouped by regime at $\kappa_{\text{target}} = 100$ and sorted by K, the number of distinct training ratios $k_\ell = D/N$, which run from the smallest to the largest run, in tokens per parameter except words, bytes and characters per parameter for Rosenfeld et al. (2020), Gordon et al. (2021) and Tao et al. (2024); β_{eff} is the exponent the paper reports (for Li et al. (2025a), its N-dependent exponent at the geometric-mean size), and $\kappa = 100\,\tau_K/V_K$ the smallest target the design passes. [e] only the ends of each size's token range are recoverable; [r] ratios reconstructed (Hoffmann: 245 of about 400 runs digitised from a figure; Kaplan: 38 runs read from a vector figure; Hu: interior checkpoints assumed evenly spaced); [w] passes only with equal weight per ratio, not with the proof's weight per run; [*] the paper prints no exponent, so 0.28 is used; [b] passes at 0.28.

Design	K	k_ℓ	β_{eff}	V_K	τ_K	κ
Above the threshold: 11 designs pass						
Choshen et al. (2025)[*]	243	0.874–4284	0.28	0.0521	0.0139	27
Hoffmann et al. (2022)[r]	151	0.0361–341	0.285	0.104	0.0188	18
Besiroglu et al. (2024)	151	0.0361–341	0.366	0.159	0.0186	12
Muennighoff et al. (2025)	95	0.0912–4.26×10^4	0.353	0.0783	0.0136	17
Gordon et al. (2021)	49	0.0436–6361	0.38	0.576	0.043	7.5
Kaplan et al. (2020)[r]	34	0.0297–5.6×10^4	0.103	0.0644	0.0312	48
Li et al. (2025a)	27	0.314–2570	0.327	0.157	0.0193	12
Aghajanyan et al. (2023)	21	0.746–1.25×10^4	0.22	0.0605	0.0177	29
Rosenfeld et al. (2020)[w]	18	0.0762–9990	0.735	2.97	0.233	7.8
Sardana et al. (2024)	12	10–10^4	0.24	0.0218	0.0129	59
Gadre et al. (2025)[w]	2	20, 320	0.242	0.014	0.0132	94
Below the threshold: 13 fail, 3 at one ratio						
Xu et al. (2026)[b]	20	0.0714–5.56	0.125	0.0245	0.0452	180
Ludziejewski et al. (2024)[b]	16	7.57–488	0.147	0.0103	0.0175	170
Sun et al. (2025)	16	14.7–2439	0.516	0.00353	0.0102	290
Kumar et al. (2025)[b]	14	6.82–867	0.496	0.00878	0.0106	120
Frantar et al. (2024)	10	193–5.04×10^4	0.203	0.0054	0.011	200
Shukor et al. (2025)[e*]	7	6.36–48.5	0.28	0.00919	0.0152	170
Hu et al. (2024)[r]	6	10–60	0.23	0.00452	0.0148	330
Tao et al. (2024)	4	130, 142, 161, 178	0.447	3.17×10^{-5}	0.0102	3.2×10^4
Bhagia et al. (2025)	4	20, 40, 100, 200	0.24	0.00621	0.0132	210
Volkova et al. (2026)	4	30, 50, 100, 200	0.38	0.00283	0.0109	380
Sedova et al. (2026)[*]	1	100	0.28	0	0.0116	∞
Magnusson et al. (2025)[*]	1	100	0.28	0	0.0116	∞
Panferov et al. (2025)[*]	1	100	0.28	0	0.0116	∞

What a verdict implies. The test concerns the split of the loss between the N and D terms, at one target. It is a leading-order verdict, which where $\varepsilon \ln N$ is near 1 can misjudge the exact $\kappa_{A,B}$ several-fold (Appendix B.12).

- *A pass* (above the threshold) of a collinear design means the (A, B) block is conditioned to the target at leading order; a pass is nearly necessary at leading order, not sufficient, for the whole fit. None of the 11 passes here is collinear: for them a pass means the distinct ratios are spread widely enough, and 8 of the 10 that can be rebuilt also pass with the proof's weights (Weights above). A pass does not make the reported coefficients right: the runs of Hoffmann et al. (2022) pass, yet the refit of Besiroglu et al. (2024) on a digitised copy of them gives $\beta = 0.366$ against the reported 0.285, so fitting choices and misspecification still move coefficients. The pooled $L(N, D)$ grid of Kaplan et al. (2020) passes, but its allocation comes from its compute frontier, so the test says nothing about the gap between its allocation and Chinchilla's.
- *A failure* (below the threshold) needs no such assumption. It means the split is weakly determined at that target: the fit pins down the combined trend along its training ratios, while its uncertainty along the split is up to $\sqrt{\kappa}$ times larger than across it. Readings that rest on the split, such as separate N and D exponents, allocations derived from them, or predictions at untrained ratios, inherit that uncertainty. A failure does not make them wrong or off-ray prediction impossible, but for these multi-ratio designs Theorem 1 asserts no ordering (Lemma 3).

Readings that rest on the split. For each failing design, what its paper draws from the split, with κ from Table 11:

- Hu et al. (2024) ($\kappa \approx 330$): data should be 192 times the model size, from a fit spanning ratios 10 to 60.
- Bhagia et al. (2025) (210): the task loss of a 7B model at 4T tokens and a 13B model at 5T (571 and 385 tokens per parameter), outside its ratios 20 to 200. Their reported errors, 5.2% and 6.6% on task loss and within 2 points of accuracy on four tasks, show the predictions landed close; they are not evidence that the fit determines the split.
- Magnusson et al. (2025) ($K = 1$): none. Because their suite has one ratio they use a one-variable law in compute, and their bivariate five-parameter variant predicts the 1B targets with 230.8% relative error against 6.5% for that law.
- Volkova et al. (2026) (380): comparisons of optimizers through fitted exponents, which the paper itself finds not robust. It reports correlated pairs (A, α) and (B, β), which it attributes to the law form and the narrow N and D ranges.
- Kumar et al. (2025) (120; passes at 0.28): that training larger models in lower precision may be compute optimal.
- Xu et al. (2026) (180; passes at 0.28): the gain of its masked-input regularizer, read as about 1.3 times as much unique data.
- Ludziejewski et al. (2024) (170; passes at 0.28): compute-optimal sizes, token counts and expert granularity, at 0.12 to 0.68 tokens per parameter against trained ratios of 7.57 to 488, and an efficiency gap over dense models that widens with scale.
- Frantar et al. (2024) (200): that the optimal sparsity at a fixed number of non-zero parameters rises with training data.
- Sun et al. (2025) (290): a critical data size past which more data hurts low-precision training, and an optimal precision of 4 to 8 bits.
- Tao et al. (2024) (3.2×10^4): compute-optimal vocabulary sizes from its parametric fit; its IsoFLOP and derivative approaches reach the same conclusion.
- Sedova et al. (2026) ($K = 1$): none on the split; its law prescribes mixture ratios.
- Shukor et al. (2025) (170): none on the split; under its additive law the best mixture does not depend on N and D, and its tests at 3B and 7B parameters lie inside its ratios 6.36 to 48.5.
- Panferov et al. (2025) ($K = 1$): capacity factors that scale the N term alone; the paper names its single ratio of 100 tokens per parameter as a caveat.

D Appendix - reanalysis of public training-loss releases

This appendix tests the criterion of Proposition 1 on scaling-law data we did not generate, and asks whether its effect weakens at large model sizes. The protocol was written before any loss was read and then amended six times. The first two amendments repaired defects of the design; the other four

followed results of earlier runs, and the paragraph *Protocol history* gives each with the result that preceded it. The protocol with every amendment (`CRITERIA.md`), the screening log and the code behind every number below are in the released repository (`camera_ready/public_checkpoints/`). The run tables are on HuggingFace (`leibnitz-lab/public_checkpoint_reanalysis`), whose contents the paragraph *Sources and licenses* lists. The fits at 512 seeds are not distributed; the released code regenerates them from the run tables.

Studies. A study qualifies if it publishes each run's loss with recoverable N and D and has a two-dimensional core. The core needs at least 6 model sizes and at least 3 TPP bins (factors of two), each covered by 4 or more sizes. Each analysis cell holds one architecture and one training corpus, so a study with several of either splits into several cells. Ten releases from nine studies qualify (Table 12); the two Gemstones releases are one study. Five release final losses of fully trained runs, the primary corpus of 984 runs. The other five release loss trajectories, admitted by Amendment 4. Each run is thinned to at most 12 checkpoints spaced log-uniformly in D, and each checkpoint enters as a further (N, D) point, giving a wider corpus of 12,531 points including the 984. Most screened releases fail on too few sizes.

Table 12: The ten public releases. Points are fully trained runs (final losses) or checkpoints along a run (trajectories); sizes are distinct N. For Chinchilla, ColPret rounds the digitized sizes to the nearest 10M and sets $D = C/(6N)$ from the digitized compute C, so its TPP range ends at 379, not at the 341 of Table 1, which keeps the digitized sizes.

Study	Losses	Points	Sizes	N range	TPP range
Farseer (Li et al., 2025a)	final	404	25	99.6M–6.37B	0.31–2570
Chinchilla isoFLOP (Hoffmann et al., 2022; Besiroglu et al., 2024; Choshen et al., 2025)	final	245	43	60M–16.2B	0.036–379
Gemstones annealed (McLeish et al., 2025)	final	199	20	48.4M–2.01B	5.5–2274
Over-training suite (Gadre et al., 2025; Choshen et al., 2025)	final	104	6	11M–6.9B	4.8–677
OLMo ladder (Bhagia et al., 2025)	final	32	7	190M–13.2B	10–573
DataDecide (Magnusson et al., 2025)	trajectory	10,479	14	3.74M–1.18B	2.2–110
MAD (Poli et al., 2024; Choshen et al., 2025)	trajectory	648	40	40.2M–3.08B	0.00026–2816
Gemstones constant LR (McLeish et al., 2025)	trajectory	264	22	48.4M–2.01B	1–7237
GPT-3 curves (Brown et al., 2020; Choshen et al., 2025)	trajectory	97	8	125M–175B	0.055–2396
OPT trajectories (Zhang et al., 2022; Xia et al., 2023; Choshen et al., 2025)	trajectory	59	6	125M–175B	0.046–2240

Protocol history. Amendment 1 replaced an on-ray holdout that was empty for every collinear subset, and Amendment 2 added a second non-collinear subset, drawn only from the collinear subset's model sizes. The other four amendments followed results:

- Amendment 3 followed a run that contrasted nominal design labels, and replaced that contrast by the margin regression of the *Test* paragraph.
- Amendment 4 followed a pooled estimate whose interval contained zero on the three studies screened first (Farseer, Chinchilla and Gadre et al. (2025), 753 runs: 8%, $[-1\%, 16\%]$, at 5 seeds). It reopened the candidate list, added two fully trained releases the first screen had missed (Gemstones annealed and the OLMo ladder, 231 runs), which took the primary corpus from 753 to 984 runs, and admitted the five trajectory releases. On the three original studies at 512 seeds the pooled estimate is 6% ($[-4\%, 15\%]$) and the within-cell one 33% ($[17\%, 43\%]$).
- Amendment 5 was written after the interval in windows centred at 1B and above was seen to contain zero. It raised that band's seeds from 5 to 512 and its resamples from 2,000 to 200,000, and fixed further window constructions and estimators for the band, each to be reported (*Scale* paragraph).
- Amendment 6 applied 512 seeds to every estimate in both corpora and fixed the treatment of the band below 100M.

Each amendment was written before its own new numbers were computed.

Sources and licenses. The Chinchilla isoFLOP points are the digitization of Besiroglu et al. (2024) (CC-BY-4.0), read through the ColPret aggregation of Choshen et al. (2025) (MIT), which also

supplies the tables of Gadre et al. (2025), MAD, GPT-3 and OPT. The releases of Gadre et al. (2025), MAD and Gemstones are under MIT, and the OLMo ladder and DataDecide under Apache-2.0. The GPT-3 curves are ColPret's digitization of a published figure, and ColPret takes the OPT losses from Xia et al. (2023), whose repository states no license. Farseer's repository states no license. The HuggingFace dataset, whose card gives each source's license with its text, holds three files:

- `runs_expanded.csv`, 14,065 rows with N, D, the released loss columns and $k = D/N$: those of the ten releases and 1,462 single-epoch runs of Muennighoff et al. (2025) (Apache-2.0, read through ColPret), a study the screen rejects for 3 sizes per corpus;
- `runs_normalised.csv`, 2,215 rows of the same form: Farseer, Chinchilla, Gadre et al. (2025) and the same 1,462 runs;
- `datadecide_ppl_by_group.csv`, an unchanged copy of DataDecide's per-group perplexity table (Apache-2.0).

Of Farseer's 404 runs the current version keeps only N, D and three loss columns, cited to Li et al. (2025a), and the code downloads the original from the Farseer repository. The dataset's history is squashed, so no revision holds Farseer's raw run table.

Test. Five steps, all fixed by the protocol:

1. *Windows.* Within a cell the distinct model sizes are sorted and grouped into windows of adjacent sizes, by the number of sizes in the cell:
   - 12 or more: windows of 8 sizes with stride 2;
   - 8 to 11: windows of 5 sizes with stride 1;
   - fewer: one window.

   Consecutive windows overlap. A window is centred at the geometric mean of its sizes, and the size bands below group windows by that centre.
2. *Holdout.* Within a window the 15% of points with the lowest and the 15% with the highest TPP form a holdout shared by every design. In a trajectory release a holdout point can come from the same run as training points.
3. *Designs.* Every factor-of-two TPP bin of the remaining points that holds enough points for the law and at least 4 sizes gives one collinear subset; every such bin is used and none is chosen. A non-collinear subset of the same size is drawn from the same points by the paper's bounding-box selector. The seed drives this selector and the fitter's restarts. A second non-collinear subset, drawn only from the collinear subset's model sizes, enters companion fits.
4. *Margins.* Each subset gets the margin $m = \log_{10}(V_K/\tau_K)$ at $\beta_{\text{eff}} = 0.28$, fixed a priori. Since κ_{target} shifts every margin by one constant, it cancels in $\Delta = m_{\text{NC}} - m_{\text{CO}}$ and enters only the regime test.
5. *Estimate.* The Chinchilla and Kaplan laws are fitted to each subset; Droppo-Elibol is excluded a priori, since some windows cannot supply its 6-point minimum. The outcome is the gap $g = \log_{10}(\text{RMSE}_{\text{CO}}/\text{RMSE}_{\text{NC}})$ on the holdout, regressed on Δ with a fixed effect per cell. Its slope β_Δ is reported as $10^{\beta_\Delta} - 1$. The value is the change in the collinear subset's relative off-ray error when the non-collinear subset's V_K/τ_K exceeds the collinear one's by a further factor of ten. Intervals are 95% bootstrap intervals that resample windows within cells, with 200,000 resamples (50,000 for one study's own slope). Windows overlap, so these intervals are anti-conservative.

Results. Table 13 gives the estimates for both corpora. On the five fully trained releases, the collinear subset's relative off-ray error rises by 40% ($[24\%, 52\%]$) per tenfold increase in the ratio of the two subsets' V_K/τ_K within a cell, and by 12% ($[5\%, 18\%]$) with no fixed effect; on the three studies of the original protocol alone the latter is 6% ($[-4\%, 15\%]$). The within-cell estimate pools four studies whose own slope is identified: three positive (Chinchilla 32%, Farseer 38%, Gemstones annealed 168%), and the suite of Gadre et al. (2025), -42%, without an interval since its three cells have one window each. The OLMo ladder's single window has almost no spread in Δ and identifies no slope. Every companion in Table 13 excludes zero on the five. The protocol's secondary regressor, the collinear subset's own margin, gives 23% ($[6\%, 35\%]$) more relative error per tenfold larger shortfall of its V_K below τ_K. In post hoc refits that each drop one of the five studies, all ten point estimates (within-cell and pooled) stay positive. The within-cell intervals all exclude zero, and the pooled ones all but the one dropping the OLMo ladder ($[-1\%, 20\%]$).

On all ten releases the primary estimate does not detect the effect: it is 4% ($[-10\%, 21\%]$), the pooled one 0% ($[-11\%, 15\%]$), and the secondary regressor's 7% ($[-5\%, 20\%]$). Of the companions

in Table 13, three exclude zero: the suite-weighted pooled fit (7%, [3%, 13%]) and the second non-collinear subset, within a cell (15%, [2%, 29%]) and with a fixed effect per window (15%, [4%, 27%]). The two corpora disagree, and the disagreement is a finding of this reanalysis. Per study, the effect holds on DataDecide's 25 data recipes (19%, [4%, 35%]) and on the constant learning-rate Gemstones trajectories (188%, [54%, 347%]), and it reverses on MAD's trajectories (−86%, [−95%, −54%]), whose TPP falls to 0.00026. GPT-3 gives 18% ([−34%, 180%]) and OPT, in one window, −63%. Of post hoc refits that each drop one of the ten, only the one dropping MAD excludes zero (28%, [16%, 41%]). Unconditionally the non-collinear subset wins 47% of pairs on the five and 50% on the ten, so any advantage on public data is conditional: the regime test below shows where it appears.

Table 13: Change in the collinear subset's off-ray error relative to the non-collinear one's per tenfold increase in the ratio of the two subsets' V_K/τ_K ($10^{\beta_\Delta}-1$), with 95% intervals that resample windows within cells (anti-conservative, since windows overlap); 512 seeds. The rows from the second to the thirteenth are the protocol's companions: complete cases drop the pairs whose collinear subset has $V_K = 0$, the β_{eff} rows are the ends of its grid (the middle values give 38-40% on the five and 3-5% on the ten), and the second non-collinear subset is fitted within a cell, with a fixed effect per window, and pooled. A window is centred at the geometric mean of its sizes. The last row is $10^{b_1}-1$, where b_1 is the change in β_Δ per tenfold N, with $\log_{10} N$ centred at the pairs' mean (630M on the five, 152M on the ten).

	Five fully trained		All ten	
Specification	Estimate	95% interval	Estimate	95% interval
Within a cell (primary)	40%	[24%, 52%]	4%	[−10%, 21%]
Pooled, no fixed effect	12%	[5%, 18%]	0%	[−11%, 15%]
Pooled, suite-weighted	10%	[7%, 13%]	7%	[3%, 13%]
Within a cell, suite-weighted	28%	[8%, 41%]	12%	[−5%, 27%]
Fixed effect per window	35%	[21%, 44%]	9%	[−2%, 21%]
Complete cases	38%	[20%, 57%]	1%	[−14%, 18%]
Chinchilla law only	48%	[30%, 68%]	5%	[−11%, 26%]
Kaplan law only	30%	[7%, 53%]	3%	[−13%, 19%]
$\beta_{\text{eff}} = 0.10$	39%	[21%, 53%]	0%	[−18%, 22%]
$\beta_{\text{eff}} = 0.50$	33%	[20%, 44%]	6%	[−3%, 16%]
Second subset, within a cell	46%	[36%, 58%]	15%	[2%, 29%]
Second subset, fixed effect per window	45%	[34%, 62%]	15%	[4%, 27%]
Second subset, no fixed effect	14%	[8%, 19%]	10%	[−1%, 24%]
Windows centred below 100M	not established (2 windows, 1 cell)		32%	[15%, 56%]
Windows centred at 100M to 1B	42%	[20%, 58%]	−28%	[−47%, −3%]
Windows centred at 1B and above	28%	[−9%, 41%]	23%	[−12%, 36%]
Size interaction, per tenfold N	−14%	[−40%, 18%]	−39%	[−57%, −18%]

Regimes. In this appendix, Regime A and B name the protocol's threshold classes, $V_K < \tau_K$ and $V_K \geq \tau_K$, not the regimes of Theorem 1. The protocol's regime test classifies both subsets of a pair, at each κ_{target} of its grid (Table 14); Table 7 classifies a pair by its collinear subset alone. These collinear subsets are factor-of-two TPP bins, not single rays, so the comparison is empirical. Under that rule, with the protocol's $\beta_{\text{eff}} = 0.28$ for both laws, every pair at $\kappa_{\text{target}} = 100$, the main-text value, is below the threshold. There the non-collinear subset is not ahead: it wins 47% of pairs on the five and 50% on the ten, both intervals of the median gap containing zero. At 10^4, 10^5 and 10^6 it wins 58%, 61% and 76% of the pairs in Regime A on the five and 58%, 56% and 76% on the ten, every interval excluding zero except the five's at 10^4. Table 7 classifies Kaplan by its exact $\kappa_{A,B}$; here too, with Kaplan's pairs classified by the exact $\kappa_{A,B}$ of their collinear subset at the parameters of Kaplan et al. (2020), the rates become 59%, 64% and 67% on the five and 53%, 56% and 55% on the ten, and the intervals contain zero at 10^4 on the five and at 10^4, 10^5 and 10^6 on the ten. Appendix C passes the full designs of Hoffmann et al. (2022), Li et al. (2025a) and Gadre et al. (2025) at 100 (the last with equal weight per ratio), but no subset of the five passes here. These subsets are narrower: a collinear subset spans one factor-of-two TPP bin and a non-collinear one at most the TPP range of one window without its holdout tails, both at $\beta_{\text{eff}} = 0.28$.

The joint rule separates a pair when only the non-collinear subset meets the threshold. At $\kappa_{\text{target}} = 100$ it separates no pair on the five and pairs in two windows on the ten (Table 14). At 10^4, 10^5 and

10^6 the non-collinear subset wins 59%, 61% and 76% of separated pairs on the five and 59%, 56% and 76% on the ten, and the interval of the median gap excludes zero in each case. The protocol asks for a monotone trend across targets: on the five the win rate rises but the median gap does not (0.074, 0.069, 0.151), and on the ten neither does. Where both subsets fall in one regime the protocol expects no ordering (the theorem predicts none for multi-ratio subsets) and counts a median gap whose interval excludes zero as a failure of this test. Pooled over sizes, the test fails on the five. In the four large regime cells with both subsets in one regime the collinear subset is slightly ahead: NC wins 44-47%, and its median off-ray error is 6-11% higher. The interval excludes zero at 10^4 and, only just, at 10^6; it only just contains zero at 10^5, and contains it at 10^2. On all ten the same four cells contain zero (NC wins 48-50%). A small regime cell at 10^4 with both subsets in Regime A (826 pairs in 3 windows, all from the five) fails in both corpora: NC wins 30%. In the reversed regime cell, where only the collinear subset is certified, NC wins 26% on the five and 19% on the ten.

The protocol also runs both tests inside each size band, with the same intervals over windows; the small cell above is left aside:

- *Windows centred below 100M.* The five give 2 windows in one cell, too few to establish either test. On the ten the collinear subset is ahead where both subsets share a regime (NC wins 36-43%), and the non-collinear one in the separated cell at 10^4 (NC wins 59%), every interval excluding zero.
- *Windows centred at 100M to 1B.* On the five no interval excludes zero: NC wins 47-62% of separated pairs and 46-48% of pairs in one regime. On the ten the non-collinear subset is ahead where both share a regime (NC wins 56-57%, every interval excluding zero), and no separated cell excludes zero.
- *Windows centred at 1B and above.* In both corpora the collinear subset is ahead where both share a regime, at every target (NC wins 41-44%; every interval excludes zero except the five's at 10^2, which only just contains it). The separated pairs come from at most three windows, too few to establish the confirmatory test.

So the pooled pass on the ten averages opposite signs across bands, and at 1B and above the falsification test fails in both corpora.

Table 14: The regime test. For each κ_{target} and regime cell (the regimes of the collinear and the non-collinear subset, classified jointly), the non-collinear win rate and the median gap g with its 95% interval over windows (200,000 resamples); $g > 0$ favours the non-collinear subset. Regime cells without pairs in either corpus are omitted.

			Five fully trained		All ten	
κ_{target}	CO	NC	NC wins	Median g	NC wins	Median g
10^2	A	B		no pairs	53.1%	0.437 $[0.437, 0.437]$
10^2	A	A	46.6%	-0.024 $[-0.059, 0.011]$	49.6%	-0.002 $[-0.027, 0.017]$
10^4	A	B	59.0%	0.074 $[0.003, 0.153]$	58.6%	0.076 $[0.047, 0.103]$
10^4	A	A	29.9%	-0.198 $[-0.207, -0.187]$	29.9%	-0.198 $[-0.207, -0.187]$
10^4	B	B	43.7%	-0.045 $[-0.085, -0.009]$	47.5%	-0.015 $[-0.050, 0.006]$
10^4	B	A	25.8%	-0.543 $[-0.658, 0.036]$	19.4%	-0.379 $[-0.578, -0.206]$
10^5	A	B	61.0%	0.069 $[0.015, 0.150]$	56.4%	0.061 $[0.014, 0.105]$
10^5	B	B	45.0%	-0.037 $[-0.076, 0.001]$	49.3%	-0.004 $[-0.031, 0.016]$
10^6	A	B	76.0%	0.151 $[0.060, 0.187]$	76.0%	0.151 $[0.060, 0.187]$
10^6	B	B	45.1%	-0.035 $[-0.070, -0.001]$	49.4%	-0.004 $[-0.029, 0.015]$

Scale. The leading-order criterion does not depend on absolute model size (Proposition 1), but whether the error effect keeps its size at large N is an empirical question, and these data do not settle it. On the five fully trained studies the effect is 42% in windows centred at 100M to 1B and 28% ($[-9\%, 41\%]$) in windows centred at 1B and above, and the size interaction is -14% ($[-40\%, 18\%]$) per tenfold N. At the point estimate the effect falls from 38% at 630M, where the size term is centred, to 19% ($[-19\%, 62\%]$) at 6.3B, so the five cannot exclude complete attenuation. On all ten the interaction is -39% ($[-57\%, -18\%]$), which the protocol reads as attenuation with size. The bands do not fall monotonically, though (32%, -28%, 23%): in windows centred at 1B and above the estimate is close to the five's.

Amendment 5 fixed further specifications for windows centred at 1B and above and requires each to be reported: Table 15 gives the estimators and Table 16 the window constructions. On the five, the specifications that exclude zero each add an assumption: the second non-collinear subset (36%, [2%, 54%]), windows of 5 sizes with stride 1 under the overlap-aware interval (26%, [2%, 47%]), and a slope smooth in $\log N$ (29%, [5%, 50%] at the band's centroid, 1.8B, with a one-sided 95% floor of 9%). Pair resampling is not admissible, and of the constructions with disjoint windows two contain zero and the one with 4 windows gives a degenerate interval. Under the pre-registered estimator the data do not exclude the loss of the whole effect below 1B (the attenuation's 95th percentile is 112% of it); under the smooth slope they exclude a loss of more than 77%. On the ten the second subset gives 40% ([17%, 61%]), but the smooth slope reverses (−41%, [−64%, −10%] at the band's centroid, 1.9B), and the second subset's own size interaction is −25% ([−43%, −8%]).

Table 15: The specifications Amendment 5 fixed for windows centred at 1B and above ($10^{\beta_\Delta} - 1$, 512 seeds, 200,000 resamples; pair resampling 20,000). Pair resampling treats pairs that share runs and a holdout as independent and is not admissible as an interval. The smooth rows fit one slope linear (or quadratic) in $\log N$ over all windows and read it at the band's pair-weighted centroid or at a given N. The attenuation bounds are the 95th percentiles of $1 - \beta_{\geq 1\text{B}}/\beta_{<1\text{B}}$, the share of the effect below 1B that is lost; on the ten, whose effect below 1B is near zero, the share is undefined.

	Five fully trained		All ten	
Specification	Estimate	95% interval	Estimate	95% interval
Pre-registered (Table 13)	28%	[−9%, 41%]	23%	[−12%, 36%]
Moving-block interval, $L = 4$	28%	[−6%, 41%]	23%	[−10%, 36%]
Clusters of cell, window and ray	28%	[−18%, 46%]	23%	[−20%, 42%]
Pairs resampled (anti-conservative)	28%	[27%, 30%]	23%	[22%, 25%]
Fixed effect per window	32%	[−0.3%, 39%]	27%	[−6%, 35%]
Second subset	36%	[2%, 54%]	40%	[17%, 61%]
Second subset, clusters of cell, window and ray	36%	[−2%, 57%]	40%	[6%, 63%]
Smooth in $\log N$, at the band's centroid	29%	[5%, 50%]	−41%	[−64%, −10%]
Same, quadratic in $\log N$	36%	[5%, 58%]	−26%	[−64%, 43%]
Smooth in $\log N$, at $N = 1$B	34%	[16%, 49%]	−32%	[−55%, −5%]
Smooth in $\log N$, at $N = 3.2$B	25%	[−6%, 55%]	−47%	[−70%, −14%]
Smooth in $\log N$, at $N = 10$B	16%	[−26%, 66%]	−59%	[−80%, −23%]
Second subset, smooth, at the centroid	39%	[17%, 63%]	−18%	[−44%, 9%]
One-sided 95% lower bound	−4%		−8%	
Same, smooth in $\log N$	9%		−61%	
Attenuation bound, share of the effect below 1B	112%		-	
Same, smooth in $\log N$	77%		-	

Table 16: Window constructions of Amendment 5 in windows centred at 1B and above, five fully trained releases, 128 seeds (the first row is the protocol's construction). The moving-block interval resamples blocks of L consecutive windows of a cell, so windows that share runs move together; for disjoint windows ($L = 1$) it equals the window interval.

Width	Stride	Windows	Estimate	Window interval	Moving-block interval
8	2	14	27%	[−9%, 38%]	[−7%, 39%] ($L = 4$)
8	1	26	18%	[−4%, 29%]	[−2%, 32%] ($L = 8$)
8	8	4	2%	[2%, 47%]	same
5	5	6	22%	[−23%, 60%]	same
5	1	31	26%	[8%, 48%]	[2%, 47%] ($L = 5$)
4	4	8	52%	[−27%, 90%]	same

The ratio $(V_K/\tau_K)_{\text{NC}}/(V_K/\tau_K)_{\text{CO}}$, that is 10^{Δ}, has median 5.8 ([5.4, 6.9]) in windows centred below 1B and 3.7 ([2.1, 5.6]) in windows centred at 1B and above, whose largest model is 12.6B (intervals over windows, 300,000 resamples). At fixed β_{eff} this ratio depends on the TPP ratios alone, so it describes the public grids and does not test a size effect. The non-collinear subsets these grids offer stay better identified at large sizes, though less so at the point estimate (the ratio of the two medians is 0.64, [0.2, 1.3]). The ratio is below 1 in three of the 13 windows at 1B and above. It uses

the five fully trained studies. The one window whose collinear subsets all have $V_K = 0$, the OLMo ladder's (190M to 13.2B), is left out, and Chinchilla's 16.2B run, its largest model, enters no window. A conditioning test of scale would need the exponents fitted in each window; we do not run one.

E Appendix - experimental setup

E.1 Appendix - Transformer architecture

All models use the HuggingFace Transformers (Wolf et al., 2020) `LlamaForCausalLM` implementation of the LLaMA architecture (Touvron et al., 2023) with RoPE positional embeddings, RMSNorm, and SwiGLU activations.

E.2 Appendix - Architecture configurations

Table 17: Model architecture configurations. All models use the LLaMA architecture with rotary position embeddings (RoPE), SwiGLU activation, RMSNorm, no bias terms, a context length of 256, and vocabulary size of 100,277 (cl100k_base).

Params	n_{layers}	d_{model}	d_{ff}	n_{heads}	d_{head}
5,035,656	24	24	96	12	2
10,070,160	12	48	192	1	48
14,411,456	24	64	256	32	2
19,400,928	1	96	384	1	96
22,796,832	24	96	384	24	4
28,819,840	12	128	512	4	32
32,498,720	1	160	640	16	10
35,368,160	8	160	640	4	40
40,277,184	3	192	768	12	16
47,949,888	16	192	768	1	192
55,538,432	4	256	1024	32	8
63,931,136	12	256	1024	32	8
68,127,488	16	256	1024	32	8
76,520,192	24	256	1024	1	256

Model sizes were specified as target parameter counts and mapped to concrete architectures via nearest-neighbor matching on a grid of transformer hyperparameters (layer count, hidden dimension, head count), with $d_{\text{ff}} = 4 \times d_{\text{model}}$. Because the grid is discrete, realized parameter counts differ slightly from targets; the 14 configurations span $N \in [5.04,\ 76.5]$ M ($N_{\max}/N_{\min} \approx 15.2$). Table 17 reports the exact parameter counts as verified from trained checkpoints. Following Hoffmann et al. (2022), N includes all trainable parameters (embedding and unembedding layers included). An earlier version of the architecture-search procedure produced configurations with odd per-head dimension $d_{\text{head}} = d_{\text{model}}/n_{\text{heads}}$, which crashed at training time because rotary positional encoding's `rotate_half` operation requires d_{head} to be even. The search was patched to filter to head counts producing even d_{head}, which is why n_{heads} varies across architectures in Table 17.

Architecture heterogeneity, embeddings, and the role of $C \approx 6ND$. The grid spans $n_{\text{layers}} \in \{1, 3, 4, 8, 12, 16, 24\}$, deliberately wider than a constant-shape sweep so that the fits are not informed only by smooth depth-vs-width co-variation; the four scaling laws under study are functions of (N, D) only and are agnostic to the internal architectural decomposition. With the cl100k_base vocabulary ($V = 100{,}277$), the input and output embeddings $2Vd_{\text{model}}$ are 67-99% of N in every configuration of Table 17 (e.g. $2 \cdot 100{,}277 \cdot 96 \approx 19.3$M of the single-layer 19.4M model), so the Kaplan FLOPs approximation $C \approx 6ND$, calibrated for embedding-light transformers, is loose throughout. Two consequences, neither of which compromises the formal results:

- *Conditioning and CI inflation* (Proposition 2, Corollary 1, Theorem 1) are statements about the Jacobian geometry of the regression problem $\hat{L}(N, D; \boldsymbol{\theta})$ as a function of the empirical pairs (N_i, D_i, L_i). They do not assume that the underlying loss is a smooth power law, that

depth is held fixed, or that any specific FLOPs identity holds; the Hoffmann/Chinchilla convention of taking N as *all* trainable parameters is preserved end-to-end. Corollary 1 and Theorem 1 additionally assume a well-specified law with Gaussian noise (misspecification: Appendix B.13). Architectural heterogeneity and any resulting departures of the empirical L_i from a clean power-law surface are absorbed by the misspecification extension (Eq. (86)), which preserves the CO/NC ordering under a shared, design-independent bias floor.

- *IsoFLOP visualization* uses $C_i := 6N_iD_i$ (Definition 4) purely as a relabelling: Proposition 3 and its proof in Appendix B.14 show that $\hat{L}(N_i, C_i/(6N_i); \boldsymbol{\theta}) = \hat{L}(N_i, D_i; \boldsymbol{\theta})$ as an algebraic identity - any constant or model-dependent rescaling of the FLOPs/token coefficient leaves the holdout RMSE unchanged. Embedding dominance therefore shifts where shallow points sit on the C-axis but does not affect any reported metric.

The shallow rows survive in the grid because dropping them would collapse the (N, D) design at the small-N end and weaken the TPP-coverage contrast that the formal analysis targets; treating them as standard (N, D, L) observations under the Hoffmann convention is consistent with how prior fits handle the same regime. Figure 2 shows the empirical n_{layers} distribution across all 1,933 trained models. Since C enters only the isoFLOP visualization, the loose $C \approx 6ND$ changes no reported metric (second item above).

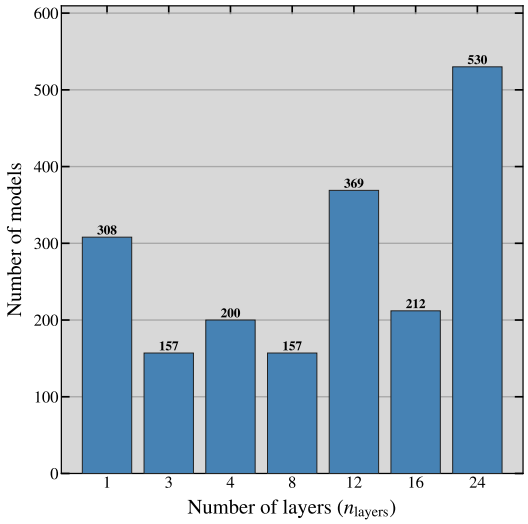


Figure 2: n_{layers} distribution across all 1,933 trained models.

E.3 Appendix - Hyperparameter selection

Table 18 summarizes the training hyperparameters shared across all experimental runs.

Table 18: Training hyperparameters shared across all runs.

Hyperparameter	**Value**
Optimizer	AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8})
Gradient clipping	1.0
Weight decay	0.1
Batch size	96 per GPU
Max sequence length	256
Epochs	3
Warmup	Linear, first 3% of steps (start factor 0.01)
LR schedule	Cosine annealing to 10% of peak LR
Seed	42

E.3.1 Learning rate

We selected $\eta = 3 \times 10^{-4}$ for Wikipedia and RedPajama, $\eta = 1 \times 10^{-3}$ for peS2o, Cosmopedia, and C4.

E.4 Appendix - Loss extraction

We record the converged training loss as the mean of the final 100 batch losses per epoch, to account for batch-to-batch noise.

E.5 Appendix - Dataset selection

We evaluate scaling law fits across five datasets spanning distinct data domains: Cosmopedia (Ben Allal et al., 2024) (synthetic educational text), Wikipedia (Wikimedia Foundation, 2023) (encyclopedic articles), peS2o (Soldaini and Lo, 2023) (academic papers and abstracts), RedPajama-V2 (Weber et al., 2024) (filtered web text), and C4 (Raffel et al., 2020) (web text). Each is used under the license or terms its dataset card states (read 2026-11-03):

- Cosmopedia (`HuggingFaceTB/cosmopedia`, `web_samples_v2`): Apache-2.0;
- Wikipedia (`wikimedia/wikipedia`, `20231101.en`): CC BY-SA 3.0 and GFDL;
- peS2o (`allenai/peS2o`, v2): ODC-By;
- RedPajama-V2 (`togethercomputer/RedPajama-Data-V2`): no license of its own for the data, which falls under the Common Crawl Foundation Terms of Use;
- C4 (`allenai/c4`, `en`): ODC-By, and the Common Crawl Terms of Use for its content.

E.5.1 Train/evaluation splits

Natural language corpora often exhibit systematic ordering biases; Wikipedia sorts by page creation date, peS2o groups abstracts before full papers, and RedPajama clusters by WARC file origin. A naive sequential (left/right) split therefore produces train and evaluation sets with different underlying distributions. *All five text corpora used in this paper (C4, Wikipedia, peS2o, RedPajama, and Cosmopedia) are split by the same procedure*: we randomly permute all document indices with a fixed seed and take contiguous slices, guaranteeing zero overlap while eliminating ordering artifacts. No corpus uses a sequential / left-right / chronological split, and no corpus is treated as a special case. Table 19 confirms this approach.

Table 19: Symmetric KL divergence and cosine similarity between train and validation n-gram distributions ($n = 1, \ldots, 3$). Both train and validation distributions have 600M tokens, drawn from the random-permutation split described above; the same split procedure is applied to every corpus, including Cosmopedia.

Dataset	**Metric**	$n = 1$	$n = 2$	$n = 3$
Wikipedia	Sym. KL	0.0006	0.4446	3.3681
	Cosine	0.999998	0.999952	0.999543
RedPajama	Sym. KL	0.0016	0.5904	4.1139
	Cosine	0.999965	0.997919	0.947550
Cosmopedia	Sym. KL	0.0002	0.2852	2.1356
	Cosine	1.000000	0.999994	0.999933
peS2o	Sym. KL	0.0007	0.4211	3.2188
	Cosine	0.999998	0.999977	0.999717
C4	Sym. KL	0.0013	0.8733	5.2464
	Cosine	0.999998	0.999953	0.998718

E.5.2 Pre-processing

Our training data was pre-tokenized with cl100k_base (tiktoken (OpenAI, 2022)), truncated to 256 tokens, or padded to 256 if an article was shorter. The training and validation losses we report average the cross-entropy over all 256 positions, the padded ones included, as the released code does. A

model learns early to predict the padding token after an article ends, so the padded positions add little loss, and each corpus pads a roughly fixed share p of its positions; the reported loss is then close to $(1-p)$ times the loss on real tokens. That factor shrinks the fitted E, A and B of a corpus by a common amount and leaves its exponents nearly unchanged, and since every comparison in the paper fits the collinear and non-collinear designs to the same corpus, it cancels from the win rates and the conditioning comparisons, and the theorems do not depend on it. Absolute losses are not real-token losses and differ between corpora by their padding share, so they should not be compared across corpora. Our repository's trainer masks padding by default; its `-score-padding` flag reproduces our runs.

E.6 Appendix - Training conditions and hardware

All pre-training used the NCSA DeltaAI GH200 partition (ACCESS allocation CIS230318) with four GH200 GPUs per node via PyTorch Distributed Data Parallel, totaling approximately 3,624 GPU-hours (151 GPU-days) across five primary corpora, BF16, and large-TPP settings. Scaling-law fitting, seed-variance analysis (Tables 4-5, Appendix G-H), and visualization used two AMD EPYC 9755 nodes, three CPUs in all: a dual-socket node (256 cores, 512 threads, 1.5 TB RAM) and a single-socket node (128 cores, 256 threads, 0.75 TB RAM), with 230 parallel workers ($\sim$5 wall-clock hours). The exhaustive subset enumeration (Table 7, 22 independent seeds) used 480 workers on the same hardware ($\sim$24 wall-clock hours, $\sim$65 min per seed). Syracuse University's OrangeGrid high-throughput cluster, both its CPU and its GPU nodes, ran supporting tasks around these stages. Total storage footprint is approximately 2.2 TB (1.5 TB model checkpoints and training metrics, 0.5 TB pre-tokenized datasets, 0.2 TB analysis artifacts and redundancy). Full-project compute was higher than the final reported runs because we also executed pilot and failed sweeps, which these totals do not include. All training uses worldsize=4 (PyTorch DDP); the LR schedule length depends on worldsize through the DistributedSampler batch count, so reproducing identical checkpoints requires exactly 4 GPUs. Training is bitwise reproducible within a fixed CUDA driver version (verified by repeated runs on NVIDIA driver 590.48). Slight cross-driver numerical drift of $O(10^{-3})$ in the training loss over three epochs has been observed following a cluster driver update; we flag it as a minor reproducibility issue. The 30-seed CIs in Tables 4 and 5 are computed over L-BFGS-B optimizer restarts on *fixed* training data (Appendix E.8), so they quantify curve-fitting variance rather than absorbing a systematic shift in the data-generating process. A common driver-induced shift in the training losses, however, does not compromise our conclusions for two reasons. **(i)** All reported metrics are *paired* CO-vs-NC comparisons on the *same* pre-trained loss table per seed and dataset (Tables 4, 5; Appendix E.8); a uniform shift ΔL in the underlying losses adds the same offset to the fitted irreducible loss E in both CO and NC fits of the two additive laws (Chinchilla, repeated-data), and is absorbed only approximately by Droppo-Elibol, whose L_∞ sits inside the outer power, and by Kaplan, which has none; so the head-to-head difference $\mathrm{RMSE}_{\mathcal{H}}^{\mathrm{NC}} - \mathrm{RMSE}_{\mathcal{H}}^{\mathrm{CO}}$ and the full-precision and BF16 win-rate summaries (97.3% and 98.7%, respectively) are first-order invariant to ΔL for the additive laws. **(ii)** The drift magnitude $O(10^{-3})$ is about two orders of magnitude smaller than the CO-vs-NC RMSE gap of Table 4 (0.237 against 0.156), so a shift that propagated unevenly between designs would still contaminate the win rate far less than the gap it estimates.

E.7 Appendix - L-BFGS-B optimizer bounds

All scaling laws are fit via multi-start L-BFGS-B with a differential-evolution global fallback. Parameter bounds are identical across all designs, datasets, and random seeds. For the Kaplan law, $N_c, D_c \in [10^3,\ 10^{14}]$ and $\alpha_N, \alpha_D \in [0.01,\ 2.0]$. For the Chinchilla law, $E \in [0,\ 10]$, $A, B \in [10^{-2},\ 10^{10}]$, and $\alpha, \beta \in [0.01,\ 2.0]$. For the Droppo-Elibol law, $L_\infty \in [10^{-6},\ 0.99 \cdot \min(L)]$, $N_C, D_C \in [10^3,\ 10^{14}]$, and $\alpha_N, \alpha_D, \alpha \in [0.01,\ 2.0]$. For the repeated-data law, $A, B \in [10^{-2},\ 10^{12}]$, $\alpha, \beta \in [0.01,\ 2.0]$, $E \in [0,\ 10]$, and the half-life parameters $R_D^*, R_N^* \in [0.1,\ 50.0]$; the repeated-data law is fit by Huber loss with $\delta = 0.5$, while the other three laws use squared-error loss.

Scale pairs left at a start draw. L-BFGS-B works in native units, where the bounds of a scale coefficient span eleven or twelve orders of magnitude. In 183 of the 900 Kaplan fits (37 CO, 146 NC) and 510 of the 900 Droppo-Elibol fits (365 CO, 145 NC), and in 1 Kaplan and 14 Droppo-Elibol fits of the 180 BF16 fits of each law, the optimizer left the scale pair within a relative 10^{-6} of one of its start draws (the counts are the same at 10^{-2}) and moved only the other parameters. Their mean $R^2_{\mathcal{H}}$ differs from the other fits' by at most 0.07 for both laws and designs: it is higher for Droppo-Elibol

and lower for Kaplan (0.69 against 0.75 on CO). At seed 0 they include the marked rows of Tables 47 and 48 and the NC fit of Figure 1. Chinchilla's (A, B) moves: 107 of its 900 fits stay within 10^{-6} of a start draw, and only 3 match one exactly.

Active lower bound on R_N^*. In the repeated-data fit (Tables 50 and 55), the parameter-side half-life R_N^* frequently terminates at exactly 0.1, the lower bound declared above. This is expected and not a fitting failure: our experimental grid fixes N per run and only D is repeated (epochs $1, 2, 3$), so the data carries strong signal about the data-side half-life R_D^* but is essentially uninformative about R_N^*, whose objective contribution becomes flat at small values. The L-BFGS-B optimizer therefore drifts to the lower bound; the bound itself is the documented value 0.1 and applies uniformly across all designs, datasets, and seeds. The reported R^2 on holdout is unaffected because the loss surface is flat in R_N^* at that point. Reporting the boundary-hitting fits as-is (rather than censoring or post-processing them) is what allows the diagnostic to be visible in the tables.

E.8 Appendix - Confidence intervals and variability bands

Throughout the paper we characterize optimizer-induced variability by re-fitting each scaling law $N_{\text{seeds}} = 30$ times per (dataset, law, epoch mode, design, coverage step) configuration, varying only the optimizer seed that controls the L-BFGS-B random-restart sequence and the differential-evolution fallback. The training data and holdout splits are held fixed across seeds, so the resulting R^2 distribution isolates optimizer-induced variance from data-induced variance.

The importance of seed-to-seed variance. This choice probes a different variance source from Corollary 1, which concerns sampling noise in the losses, while the seeds vary only the optimizer with the data fixed. Both stem from the same flat direction: the corollary bounds parameter CIs below by a factor of order ε^{-1}, and a flat direction also lets optimizer restarts land far apart; that parameter-space spread propagates to prediction uncertainty on holdout, manifesting as wide seed-to-seed spread in holdout R^2. A single CO seed can fit a particular holdout well by luck; the Gauss-Newton step along the sloppy (A, B) direction happens to land near the truth on that draw. What the formal analysis predicts, and the seed-to-seed CI measures, is that this spread is systematically larger for CO than for NC. The CI gap between designs in Tables 4, 6 is the predicted parameter-CI inflation made empirical, observed one layer downstream in prediction error rather than in A directly.

Coverage-fraction plots. The central marker is the *mean* R^2 across the 30 seeds; error bars span the 5th-95th percentile of that distribution. These are empirical percentile bands of the per-seed R^2 values, not Gaussian confidence intervals on the mean.

Summary tables (e.g. Tables 4, 6). Reported R^2 *Mean* is the grand mean of the per-combo seed means; reported R^2 *Median* is the median of those combo means; reported R^2 *Std* is their across-combo standard deviation. The 95% CI is an optimizer-noise interval centered on R^2 Mean with half-width

$$\Delta = 1.96 \cdot \frac{\overline{\sigma_{\text{seed}}}}{\sqrt{N_{\text{seeds}}}},$$

where $\overline{\sigma_{\text{seed}}}$ is the average within-combo seed-to-seed standard deviation. RMSE columns follow the same convention. The CI quantifies how tightly seed-to-seed optimizer noise pins down a typical combo's mean and is independent of the Median column.

Subset analysis (Table 7). For the budget-matched subset enumeration we use 22 seeds rather than 30 due to compute restrictions; each varies the bounding-box construction and the optimizer restarts. The **Dataset Avg.** intervals are 95% bootstrap intervals over the seeds ($5{,}000$ resamples of the per-seed mean across corpora). The preliminary high-TPP win rate (Table 27) carries a Wilson 95% binomial CI.

F Appendix - detailed results

F.1 Appendix - Collinear and non-collinear experimental grid definitions

Figures 3 and 4 visualize the non-collinear and collinear experimental designs as grids. Red cells are holdout models; blue cells are training models. Each dataset comprises 324 runs (120 collinear training $+$ 42 collinear holdout $+$ 120 non-collinear training $+$ 42 non-collinear holdout). Four of the five corpora completed all runs; C4 has 321 (3 non-collinear holdout jobs crashed).

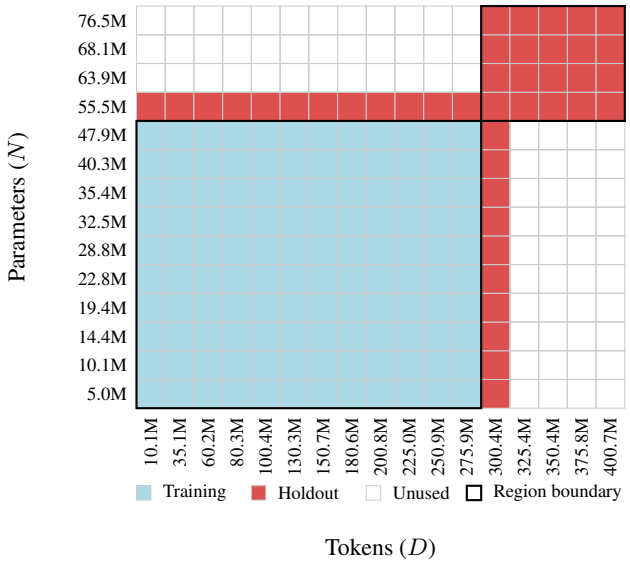


Figure 3: Grid of the non-collinear (NC) experimental design. Each cell represents a unique (N, D) training configuration. The holdout set (red) forms an L-shaped boundary at the next parameter count above the training region ($N = 55.5$M) and at $D = 300.4$M tokens, plus the 3×4 extrapolation block at the largest token and parameter counts. White cells denote configurations not included in either set. The bordered lower-left region contains the 120 in-distribution training configurations.

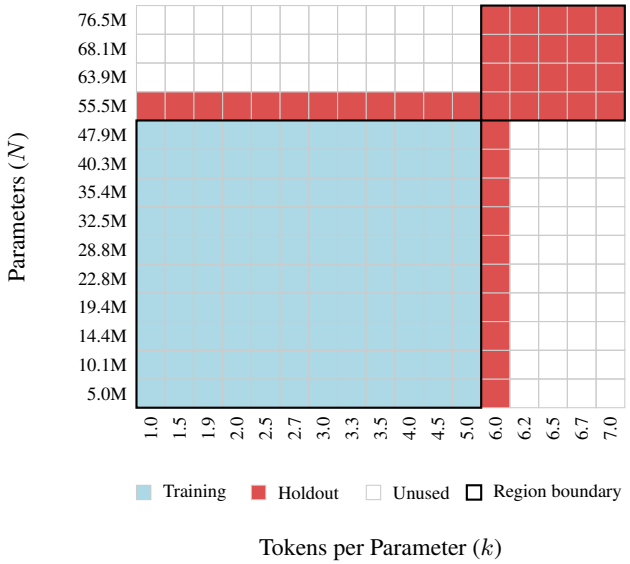


Figure 4: Tokens-per-parameter (TPP) grid of the collinear (CO) experimental design. Each cell represents a unique (N, k) configuration where the token count is $D = k \cdot N$. The holdout set (red) forms an L-shaped boundary at the next parameter count above the training region ($N = 55.5$M) and at $k = 6$, plus the 3×4 extrapolation block. White cells denote configurations not included in either set. The bordered lower-left region contains the 120 in-distribution training configurations.

F.2 Appendix - Table 4 holdout breakdown

Tables 20 and 21 break the main-experiment results down by holdout split, and the win rates also by corpus.

Table 20: R^2 and RMSE summary by training design over 30 optimizer seeds. 95% CI: optimizer-noise interval (Appendix E.8); Std: across (corpus, law, epoch) combinations; best per split in **bold**.

		R^2				RMSE			
Split	**Design**	Mean	95% CI	Median	Std	Mean	95% CI	Median	Std
Train ($\mathcal{D}$)	CO	**0.9848**	**[0.985, 0.985]**	**0.9850**	**0.0047**	**0.1737**	**[0.173, 0.174]**	**0.1616**	**0.0430**
	NC	0.9526	[0.952, 0.953]	0.9542	0.0172	0.2023	[0.202, 0.203]	0.1869	0.0443
Holdout ($\mathcal{H}$)	CO	0.8370	[0.832, 0.842]	0.8811	0.1129	0.2373	[0.233, 0.241]	0.2163	0.1229
	NC	**0.9319**	**[0.930, 0.934]**	**0.9392**	**0.0269**	**0.1561**	**[0.154, 0.158]**	**0.1242**	**0.0599**
Holdout ($\mathcal{H}_{\text{col}}$)	CO	0.9113	[0.907, 0.915]	0.9336	0.0514	0.1885	[0.185, 0.192]	0.1645	0.0903
	NC	**0.9437**	**[0.942, 0.945]**	**0.9439**	**0.0182**	**0.1560**	**[0.154, 0.158]**	**0.1182**	**0.0631**
Holdout ($\mathcal{H}_{\text{nc}}$)	CO	0.6984	[0.688, 0.708]	0.7490	0.2730	0.2757	[0.271, 0.281]	0.2445	0.1516
	NC	**0.9092**	**[0.907, 0.911]**	**0.9297**	**0.0492**	**0.1560**	**[0.154, 0.158]**	**0.1275**	**0.0569**

Table 21: Win rate (NC vs. CO, seed-paired) by holdout split and corpus.

Holdout Split	**NC Wins**	**CO Wins**	**Total**	**NC %**
$\mathcal{H}$	1,460	40	1,500	97.3%
$\mathcal{H}_{\text{col}}$	1082	418	1,500	72.1%
$\mathcal{H}_{\text{nc}}$	1480	20	1,500	98.7%
$\mathcal{H}$ / C4	279	21	300	93.0%
$\mathcal{H}$ / Cosmopedia	284	16	300	94.7%
$\mathcal{H}$ / peS2o	298	2	300	99.3%
$\mathcal{H}$ / RedPajama	300	0	300	100.0%
$\mathcal{H}$ / Wikipedia	299	1	300	99.7%
$\mathcal{H}_{\text{col}}$ / C4	252	48	300	84.0%
$\mathcal{H}_{\text{col}}$ / Cosmopedia	239	61	300	79.7%
$\mathcal{H}_{\text{col}}$ / peS2o	295	5	300	98.3%
$\mathcal{H}_{\text{col}}$ / RedPajama	151	149	300	50.3%
$\mathcal{H}_{\text{col}}$ / Wikipedia	145	155	300	48.3%
$\mathcal{H}_{\text{nc}}$ / C4	292	8	300	97.3%
$\mathcal{H}_{\text{nc}}$ / Cosmopedia	289	11	300	96.3%
$\mathcal{H}_{\text{nc}}$ / peS2o	299	1	300	99.7%
$\mathcal{H}_{\text{nc}}$ / RedPajama	300	0	300	100.0%
$\mathcal{H}_{\text{nc}}$ / Wikipedia	300	0	300	100.0%

F.3 Appendix - Experimental setup for the budget-matched subset comparison

We formalize the empirical protocol that asks, at matched training cardinality, whether NC also beats CO designs with $K \geq 2$ rays, for which Theorem 1 gives no ordering (its Regime B; Lemma 3 and the text after it). The protocol is instantiated independently for each scaling law $\mathcal{L} \in \{\text{Chinchilla, Kaplan, Droppo-Elibol}\}$ (Definition 2), each training corpus, and each training checkpoint mode $\mu \in \{\text{first, second, final}\}$. Throughout this section we reuse the symbols $(N, D, L, \mathcal{D}, K, k_1, \ldots, k_K, \mathcal{H}, \boldsymbol{\theta})$ from Definition 1 without redefinition; the quantities introduced below are the new objects specific to subset enumeration.

Definition 9 (Available TPP ratios)**.** For each corpus and mode μ, let $\mathbf{k}^{\text{all}} = (k_1 < \cdots < k_{12})$ denote the ordered set of 12 TPP ratios at which the collinear training pool $\mathcal{D}^{\text{CO}}_{\text{train}}$ (Definition 1) was collected (so $K = 12$ for the full pool). Each ratio k_ℓ indexes a ray $\{(N, k_\ell N) : N \in \mathcal{N}\}$ in the (N, D) plane, where $\mathcal{N} = \{N_1, \ldots, N_n\}$ is the shared set of model sizes of Definition 1.

Definition 10 (Collinear (CO) subset design)**.** Given any non-empty index set $S \subseteq \{1, \ldots, 12\}$, the induced *collinear subset design* is the union of training runs on the selected rays:

$$\text{CO}(S) \;\coloneqq\; \big\{(N_i, D_i, L_i) \in \mathcal{D}^{\text{CO}}_{\text{train}} \;:\; D_i/N_i = k_\ell \text{ for some } \ell \in S\big\}, \tag{104}$$

with cardinality $n_S \coloneqq |\mathrm{CO}(S)|$ and effective ray count $K_S \coloneqq |S|$ (the design's K in the sense of Definition 1). Enumeration over all non-empty S yields $2^{12} - 1 = 4{,}095$ CO subset designs per (corpus, mode) pair.

Definition 11 (Bounding-box non-collinear (NC) design)**.** Let $\mathcal{D}^{\mathrm{NC}}_{\mathrm{train}}$ be the non-collinear training pool covering the rectangular grid $\mathcal{N}_{\mathrm{NC}} \times \mathcal{D}_{\mathrm{NC}}$, with $M_N \coloneqq |\mathcal{N}_{\mathrm{NC}}|$ rows and $M_D \coloneqq |\mathcal{D}_{\mathrm{NC}}|$ columns. Index the grid cells $(i, j) \in \{0, \ldots, M_N - 1\} \times \{0, \ldots, M_D - 1\}$. Given a target cardinality $n^\star \in \mathbb{N}$ and target TPP ratios $\mathbf{k}_S = \{k_\ell : \ell \in S\}$ from a CO subset S, the *bounding-box NC design* $\mathrm{NC}_\square(n^\star, \mathbf{k}_S)$ is constructed by the following deterministic algorithm (seeded by $\mathrm{hash}(\text{corpus}, \mu, S)$ and the global seed; Appendix F.6):

1. **Center.** Initialize a 2×2 bounding box $[i_{\mathrm{lo}}, i_{\mathrm{hi}}] \times [j_{\mathrm{lo}}, j_{\mathrm{hi}}]$ centered on $(\lfloor M_N/2 \rfloor, \lfloor M_D/2 \rfloor)$ (with a random ± 1 offset when M_N or M_D is even).
2. **Initial absorption.** Include all training runs whose (N, D) falls inside the initial bounding box. If $|\mathrm{NC}_\square| > n^\star$, subsample to exactly $n^\star$ using the priority rule below.
3. **Adaptive growth.** While $|\mathrm{NC}_\square| < n^\star$ and the box has not spanned the full grid:
   (a) Enumerate the two candidate expansions: expand the N-dimension by one row on both edges, or expand the D-dimension by one column on both edges.
   (b) Score each candidate by (s_1, s_2, ξ), where s_1 counts newly-covered CO-target TPP bins, s_2 counts newly-covered non-target TPP bins (induced by the (N, D) combinations on the boundary), and $\xi \sim \mathrm{Uniform}[0, 1]$.
   (c) Select the candidate maximizing (s_1, s_2, ξ) lexicographically. If the ring fits within $n^\star - |\mathrm{NC}_\square|$ slots, absorb it entirely; otherwise absorb a priority subsample and stop.
4. **Priority subsample rule.** When a ring must be trimmed to r training points, partition its runs by TPP bin into three priority buckets: (a) runs at a target TPP not yet covered, (b) runs at a non-target TPP not yet covered, (c) runs at an already-covered TPP. Shuffle each bucket with the seeded RNG and take the first r from their concatenation.

By construction, $|\mathrm{NC}_\square(n^\star, \mathbf{k}_S)| = n^\star$ whenever the NC grid has enough runs, and $\mathrm{NC}_\square$ is a connected rectangular region of $\mathcal{N}_{\mathrm{NC}} \times \mathcal{D}_{\mathrm{NC}}$.

Definition 12 (Paired comparison)**.** For any scaling law $\mathcal{L}$, corpus $\mathcal{D}$, mode μ, and CO subset S, let $n^\star = n_S$ (Definition 10). Define:

$$\begin{aligned} \mathrm{CO}(S) &\text{ - training set from the selected rays.} \\ \mathrm{NC}_\square(n^\star, \mathbf{k}_S) &\text{ - budget-matched bounding-box NC training set.} \\ \hat{\boldsymbol{\theta}}^{\mathrm{CO}}, \hat{\boldsymbol{\theta}}^{\mathrm{NC}} &\text{ - least-squares fits of } \mathcal{L} \text{ on each design.} \end{aligned}$$

Both fits use the same optimization protocol and evaluate on the same fixed 70% subsample of the unified holdout $\mathcal{H}$ (Definition 1; Appendix F.6). The *paired comparison* is the tuple

$$\mathcal{C}(\mathcal{L}, \mathcal{D}, \mu, S) \;\coloneqq\; \big(\mathrm{RMSE}^{\mathrm{CO}}_{\mathcal{H}},\ \mathrm{RMSE}^{\mathrm{NC}}_{\mathcal{H}},\ \hat{\boldsymbol{\theta}}^{\mathrm{CO}},\ \hat{\boldsymbol{\theta}}^{\mathrm{NC}}\big), \tag{105}$$

with $\mathrm{RMSE}^{\mathrm{CO}}_{\mathcal{H}}$ and $\mathrm{RMSE}^{\mathrm{NC}}_{\mathcal{H}}$ as in Theorem 1. The set of all paired comparisons is

$$\begin{aligned} \mathcal{P} \;=\; \big\{\mathcal{C}(\mathcal{L}, \mathcal{D}, \mu, S) \;:\; & \mathcal{L} \in \{\text{Ch., Ka., DE}\}, \\ & \mathcal{D} \in \{\text{C4}, \ldots, \text{Wiki}\}, \\ & \mu \in \{\text{first, second, final}\}, \\ & S \subseteq \{1, \ldots, 12\}\big\}, \end{aligned} \tag{106}$$

with $|\mathcal{P}| = 3 \cdot 5 \cdot 3 \cdot 4095 = 184{,}275$.

Note on the budget match. $\mathrm{NC}_\square(n^\star, \mathbf{k}_S)$ contains exactly $n^\star$ training points *by construction*, matching $\mathrm{CO}(S)$ precisely. Any observed RMSE advantage of $\mathrm{NC}_\square$ over CO therefore cannot arise from a training-data size difference.

Definition 13 (NC win rate and below-threshold NC win rate)**.** For any collection $\mathcal{P}' \subseteq \mathcal{P}$, the *NC win rate* on $\mathcal{P}'$ is

$$\mathrm{WR}(\mathcal{P}') \;\coloneqq\; \frac{1}{|\mathcal{P}'|} \sum_{\mathcal{C} \in \mathcal{P}'} \mathbf{1}\big\{\mathrm{RMSE}^{\mathrm{NC}}_{\mathcal{H}}(\mathcal{C}) < \mathrm{RMSE}^{\mathrm{CO}}_{\mathcal{H}}(\mathcal{C})\big\}\,, \tag{107}$$

where $\mathbf{1}\{\cdot\}$ is the standard indicator (1 if the inequality holds, 0 otherwise) and ties are broken in favor of CO (negligible at our sample sizes).

Given a scaling law $\mathcal{L}$, mode μ, and **target condition number** $\kappa^\star > 0$ (in the sense of κ_{target} from Proposition 1), the *predicted below-threshold set at* $\kappa^\star$ is

$$\mathcal{P}^A_{\mathcal{L},\mu}(\kappa^\star) \;:=\; \big\{\mathcal{C}(\mathcal{L},\mathcal{D},\mu,S) \in \mathcal{P} \;:\; V_{K_S}(S;\beta^{\mathcal{L}}_{\text{eff}}) \;<\; \tau_{K_S}(S;\beta^{\mathcal{L}}_{\text{eff}},\kappa^\star)\big\}, \tag{108}$$

where $V_K(\cdot;\beta)$, $\tau_K(\cdot;\beta,\kappa^\star)$, and $\beta^{\mathcal{L}}_{\text{eff}}$ are exactly the quantities of Proposition 1 and Definition 2, evaluated on the rays selected by S (so $K_S = |S|$). We use the **literature value** of $\beta^{\mathcal{L}}_{\text{eff}}$ (Table 9; $\beta = 0.28$ for Chinchilla and $\gamma_D = 0.08$ for Droppo-Elibol) so that regime classification can be performed *a priori* without fitting. For Chinchilla and Droppo-Elibol the classification is the leading-order one, which at our sizes can misjudge the exact $\kappa_{A,B}$ several-fold, and for Droppo-Elibol it is a heuristic (Appendix B.12). For Kaplan we use the exact rule instead: S is below the threshold iff $\kappa_{A,B}$ of the Gram block of the columns $\partial\hat{L}/\partial N_c$ and $\partial\hat{L}/\partial D_c$ of the fitted law, on the runs of $\mathrm{CO}(S)$, at the parameters of Kaplan et al. (2020) ($N_c = 8.8\times10^{13}$, $D_c = 5.4\times10^{13}$, $\alpha_N = 0.076$, $\alpha_D = 0.095$, from their single-variable fits), exceeds $\kappa^\star$. Droppo-Elibol keeps the leading-order rule because its literature $\gamma_D = 0.08$ comes without the other parameters its $\kappa_{A,B}$ needs. The *below-threshold NC win rate* at target $\kappa^\star$ is

$$\mathrm{WR}_A(\mathcal{L},\mu;\kappa^\star) \;:=\; \mathrm{WR}\big(\mathcal{P}^A_{\mathcal{L},\mu}(\kappa^\star)\big). \tag{109}$$

F.4 Appendix - BF16 results

Tables 22, 23 and 24 report summary statistics (Table 6 split by holdout) and pairwise win rates for the BF16 experiments described in Section 4. These runs use the same 14 model sizes N, the same 12 NC dataset sizes D, and the same 12 collinear training TPP ratios $\mathcal{K}_{\text{train}}$ as the FP32 main experiment, but are restricted to a single corpus (Wikipedia) due to compute. They cast the whole model to BF16 (`model.bfloat16()`): weights, gradients and AdamW states are BF16, and only the logits are cast to FP32 for the loss. Of their 324 runs, 302 completed: 116 of the 120 in each training design, 28 of the 42 collinear holdout runs and all 42 non-collinear holdout runs. They are not to be confused with the preliminary high-TPP (BigTPP) experiment of Appendix F.7, which uses the disjoint set $\mathcal{K}_{\text{big}} = \{10, 11, 12, 13, 14, 15\}$ in FP32.

Table 22: BF16 runs on Wikipedia: R^2 and RMSE by training design (CO: collinear, NC: non-collinear) over 30 optimizer seeds, on the training set $\mathcal{D}$, the unified holdout $\mathcal{H}$ and its collinear and non-collinear parts $\mathcal{H}_{\text{col}}$ and $\mathcal{H}_{\text{nc}}$ (Table 6 with the holdout split). 95% CI: optimizer-noise interval (Appendix E.8); Std: standard deviation across the 10 (law, epoch) combinations; the design with the higher mean R^2 per split is highlighted.

		R^2				**RMSE**			
Split	**Design**	Mean	95% CI	Median	Std	Mean	95% CI	Median	Std
Train ($\mathcal{D}$)	CO	0.9896	[0.990, 0.990]	0.9916	0.0029	0.1854	[0.185, 0.185]	0.1606	0.0353
	NC	**0.9907**	**[0.991, 0.991]**	**0.9915**	**0.0020**	**0.1457**	**[0.145, 0.146]**	**0.1428**	**0.0171**
Holdout ($\mathcal{H}$)	CO	0.9412	[0.940, 0.942]	0.9473	0.0172	0.2546	[0.252, 0.257]	0.2402	0.0396
	NC	**0.9657**	**[0.965, 0.967]**	**0.9663**	**0.0068**	**0.1949**	**[0.192, 0.198]**	**0.1921**	**0.0203**
Holdout ($\mathcal{H}_{\text{col}}$)	CO	0.9650	[0.964, 0.966]	0.9693	0.0089	0.2194	[0.218, 0.221]	0.2063	0.0279
	NC	**0.9734**	**[0.973, 0.974]**	**0.9741**	**0.0037**	**0.1919**	**[0.190, 0.194]**	**0.1897**	**0.0136**
Holdout ($\mathcal{H}_{\text{nc}}$)	CO	0.9153	[0.913, 0.917]	0.9190	0.0254	0.2753	[0.272, 0.279]	0.2673	0.0477
	NC	**0.9568**	**[0.956, 0.958]**	**0.9582**	**0.0106**	**0.1964**	**[0.193, 0.199]**	**0.1906**	**0.0266**

Table 23: BF16 runs on Wikipedia: NC win rate (NC vs. CO) on $\mathcal{H}$ by law and by epoch, seed-paired over 30 optimizer seeds; the epoch column omits repeated-data, fit on all epochs at once. Overall NC win rate: 98.7% (296/300); 95% Wilson CI $[96.6\%, 99.5\%]$ over comparisons that share one training run per (N, D). NC has the lower mean holdout RMSE in all 10 (law, epoch) cells.

Law		Epoch	
Chinchilla	97.8%	first	100.0%
Droppo-Elibol	97.8%	second	96.7%
Kaplan	100.0%	final	98.9%
Repeated-data	100.0%		

Table 24: BF16 runs on Wikipedia: seed-paired NC vs. CO wins on the unified holdout $\mathcal{H}$ and on its collinear and non-collinear parts $\mathcal{H}_{\text{col}}$ and $\mathcal{H}_{\text{nc}}$, over 10 (law, epoch) cells $\times$ 30 optimizer seeds.

Holdout Split	NC Wins	CO Wins	Total	NC %
$\mathcal{H}$	296	4	300	98.7%
$\mathcal{H}_{\text{col}}$	222	78	300	74.0%
$\mathcal{H}_{\text{nc}}$	299	1	300	99.7%

They cover Chinchilla, repeated-data (Muennighoff et al., 2025), Kaplan and Droppo-Elibol, and Appendix G breaks Table 22 down by law, epoch mode and holdout split.

F.5 Appendix - TPP convergence fraction detailed setup

We construct an incremental coverage analysis to measure how scaling law fit quality evolves as the experimental design space is progressively expanded. This analysis reuses the runs from our main experiment rather than conducting independent trials, so the results should be interpreted as a post-hoc decomposition of existing data rather than a controlled ablation. For the collinear (CO) design, we sort the 12 designed TPP levels in ascending order and add them one at a time: step 1 trains on runs from only the lowest TPP, step 2 adds the second-lowest, and so on until all 12 levels are included. Each TPP level contributes $120/12 = 10$ training-eligible runs (one per training model size after the L-shaped holdout removal), growing the training set from 10 to 120 runs across 12 steps. For the non-collinear (NC) design, we expand coverage by pinching inward from the extremes of the D range: step 1 includes runs at both the smallest and largest D, step 2 adds the second-smallest and second-largest, and so on, adding two D levels per step. This pairing ensures that even the earliest step spans the full D range. At each step, all four scaling laws are fit on the current subset and evaluated against fixed holdout sets, so changes in R^2 reflect only the effect of broader coverage. The TPP bins are scaled to match the epoch mode under evaluation ($\times 1$ for first-epoch, $\times 2$ for second-epoch, $\times 3$ for final-epoch), aligning the groupings with the effective tokens-per-parameter that each fitter sees.

A coverage fraction of 1.0 means all 12 TPP levels (for CO) or all 12 D levels (for NC) are included in the training set - this is the same configuration used in the main results tables, and the R^2 values at full coverage correspond exactly to those reported there. The CO and NC coverage fractions are not directly comparable: they traverse different axes of the design space using non-corresponding sets of values. Furthermore, the observed R^2 trajectory may depend not only on the number of levels included but also on their spacing; our designed TPP values $(1.0,\ 1.5,\ 1.9,\ 2.0,\ 2.5,\ 2.7,\ 3.0,\ 3.3,\ 3.5,\ 4.0,\ 4.5,\ 5.0)$ are unevenly distributed, with denser coverage at the low end, so a different spacing (e.g., uniformly from 1 to 5) could yield a different convergence profile. The primary takeaway is qualitative - that collinear designs require substantially more coverage to approach their asymptotic R^2, while non-collinear designs converge rapidly with minimal coverage - rather than a precise quantitative comparison between the two trajectories.

F.6 Appendix - Optimizer seed protocol

All scaling-law parameters are estimated via nonlinear least squares (L-BFGS-B with differential evolution polish) using multiple random restarts seeded by a deterministic sequence. The `seed`

column in released parquet files and CSVs stores the seed *index*; the actual RNG seed passed to the optimizer is computed from this index as described below. All model pre-training uses a fixed seed of 42 for weight initialization and data shuffling, with one run per (N, D); the seeds documented here govern only the scaling-law curve fitting.

Tables 4-5 and convergence-fraction plots. Both the full-coverage analysis (Tables 4-5) and the seed-variance convergence fraction plots are produced by identical seed logic: 30 optimizer seeds at 100 random restarts each, base seed $b = 0$, stride 1:

$$s_i = i, \qquad i \in \{0,\ 1,\ 2,\ \ldots,\ 29\}.$$

Each seed index i is passed directly as the RNG seed to the optimizer. The full-coverage stage fits, for both designs at each seed, five datasets $\times$ ten (law, epoch) cells (Chinchilla, Kaplan and Droppo-Elibol per epoch mode, repeated-data once on all epochs), producing 1,500 paired comparisons (Table 5). The seed-variance stage fits the same configurations at each incremental coverage step, recording the distribution of holdout R^2 across the 30 seeds at every step.

Tables 6 and 23 (BF16 replication). The BF16 analysis uses 30 optimizer seeds at 100 random restarts each, base seed $b = 0$, stride 137:

$$s_i = 137 \cdot i, \qquad i \in \{0,\ 1,\ 2,\ \ldots,\ 29\},$$

yielding the sequence $\{0,\ 137,\ 274,\ 411,\ \ldots,\ 3{,}973\}$. The wider stride avoids accidental overlap with the main analysis seeds; the results are statistically equivalent to any other choice of 30 independent seeds.

Table 7 (subset enumeration). The budget-matched subset enumeration uses 22 independent global seeds $g \in \{0,\ 1,\ \ldots,\ 21\}$, each with 300 random restarts per fit. The actual optimizer seed for global seed g is:

$$s_g = 42 + 100{,}003 \cdot g.$$

Table 25 lists all 22 values. The NC bounding-box construction uses a separate deterministic seed derived via BLAKE2b hash of the dataset name, epoch mode, and subset mask, XORed with $g \cdot$ `0x9E3779B1` (a 32-bit golden-ratio constant ensuring uniform dispersion). The holdout subsample seed is intentionally *not* varied across g: it is fixed at $7{,}777 + \mathrm{hash}(\text{dataset, mode}) \bmod 10^5$ so that the holdout composition remains constant across all 22 runs.

Preliminary Large TPP Seeds. Our preliminary large TPP results use the same stride as in our BF16 results, using a stride of 137.

Table 25: Fit seeds for each of the 22 subset-enumeration runs ($s_g = 42 + 100{,}003 \cdot g$).

g	s_g	g	s_g	g	s_g
0	42	8	800,066	16	1,600,090
1	100,045	9	900,069	17	1,700,093
2	200,048	10	1,000,072	18	1,800,096
3	300,051	11	1,100,075	19	1,900,099
4	400,054	12	1,200,078	20	2,000,102
5	500,057	13	1,300,081	21	2,100,105
6	600,060	14	1,400,084		
7	700,063	15	1,500,087		

Table 26: Seed protocol summary. Count: number of seeds; Restarts: random restarts per fit. Index i ranges over optimizer seeds within a stage; g ranges over independent subset-enumeration runs.

Stage	**Count**	**Restarts**	**Tables/Figures**	**RNG seed**
Full coverage	30	100	4, 5	i
Seed variance	30	100	Convergence plots	i
BF16 replication	30	100	6, 23	$137i$
Subset enumeration	22	300	7	$42 + 100{,}003g$
BigTPP prelim	30	100	27	$137i$

F.7 Appendix - Preliminary large TPP experiment

To check whether the NC advantage persists at higher TPP, we ran a single-corpus preliminary experiment on Wikipedia. Due to compute restrictions, the grid is smaller and we fit only Chinchilla.

Training grid. Both designs share ten model sizes

$$N \in \{5.04, 10.07, 14.41, 19.40, 22.80,\ 28.82, 32.50, 35.37, 40.28, 47.95\}\text{M}.$$

CO sweeps $k \in \mathcal{K}_{\text{big}} := \{10, 11, 12, 13, 14, 15\}$. NC sweeps $D \in \{100, 225, 400, 550, 700, 825\}$ M.

Holdout. Four out-of-grid sizes $N_{\text{holdout}} \in \{55.54, 63.93, 68.13, 76.52\}$ M (Table 17), all above the training maximum. For holdout model size 55M, the holdout model used TPP values $10, 11, 12, 13, 14, 15, 16, 18, 20$. For non-collinear, 55M models used $D \in \{100, 225, 400, 550, 700, 825, 900, 1100, 1400\}$. At the three largest holdout sizes, $\mathcal{H}_{\text{col}}$ only includes TPP $\in \{16, 18, 20\}$ and $\mathcal{H}_{\text{nc}}$ additionally includes $D \in \{900, 1100, 1400\}$ M.

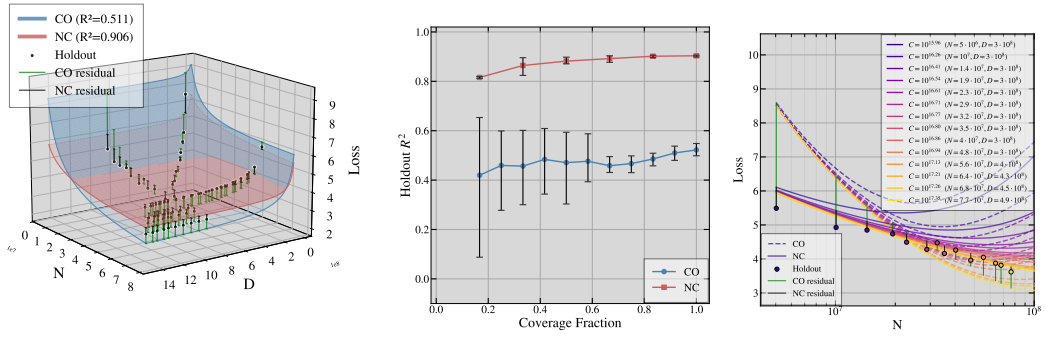
Fitting. Chinchilla, 30 seeds (stride 137, matching Appendix F.4), 100 restarts with differential evolution polish. We report seed-paired NC win rate (Table 27) with Wilson 95% CI. A full BigTPP replication is left to future work.

Table 27: Preliminary BigTPP NC win rate on Wikipedia (Chinchilla, first epoch, CO at $k \in \mathcal{K}_{\text{big}}$): seed-paired NC vs. CO comparisons over 30 optimizer seeds, with a 95% Wilson CI. A compute-limited preliminary experiment, not a full result.

Law	NC Wins	Total	NC % [95% CI]
Chinchilla	19	30	63.3% [45.5%, 78.1%]

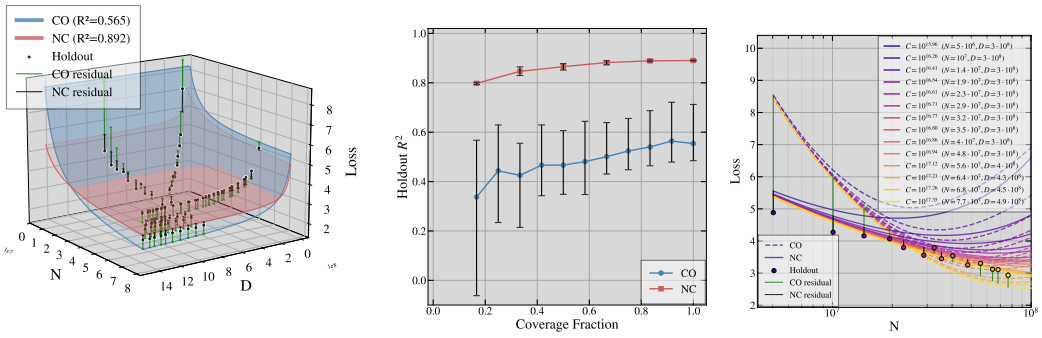
F.8 Appendix - Three-panel plots of the fitted laws

Figures 5-25 continue the three-panel plots of Figure 1, with the same panels. The BF16 figures (Figures 21-25) have no coverage-fraction panel, since the BF16 runs were not part of the coverage tracking of Section 4.



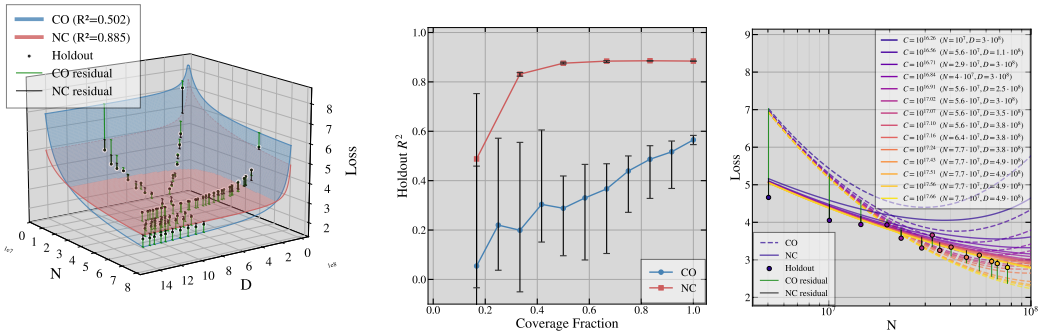
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 5: Kaplan law on RedPajama, first epoch: CO and NC fits, with the panels of Figure 1.



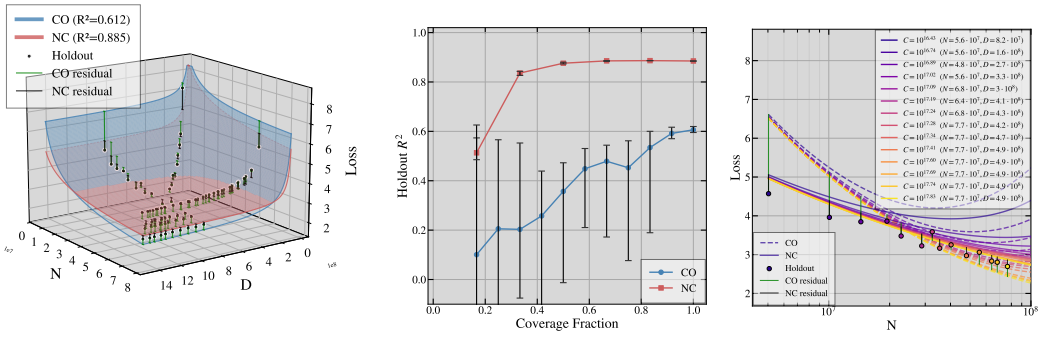
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 6: Kaplan law on Wikipedia, first epoch: CO and NC fits, with the panels of Figure 1.



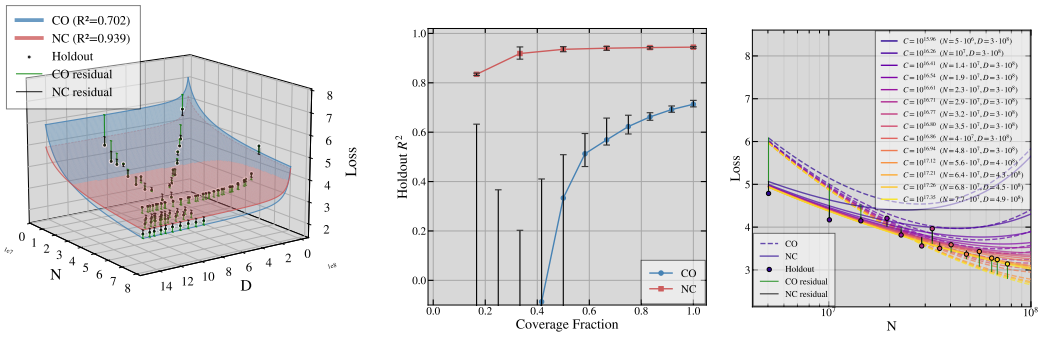
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 7: Kaplan law on Wikipedia, second epoch: CO and NC fits, with the panels of Figure 1.



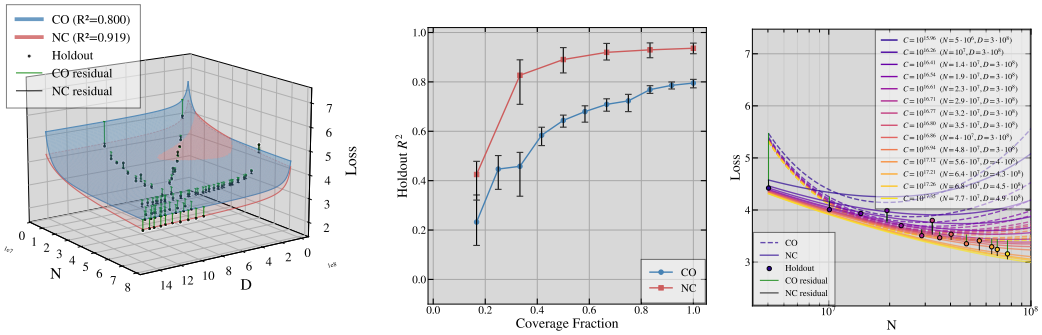
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 8: Kaplan law on Wikipedia, final epoch: CO and NC fits, with the panels of Figure 1.



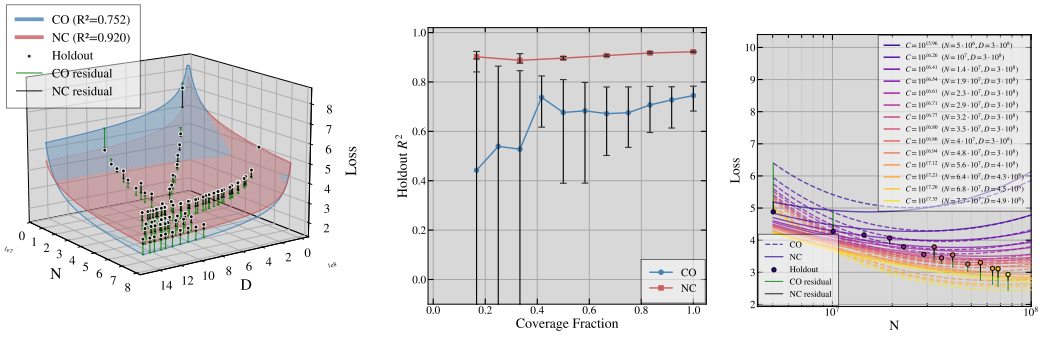
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 9: Kaplan law on peS2o, first epoch: CO and NC fits, with the panels of Figure 1.



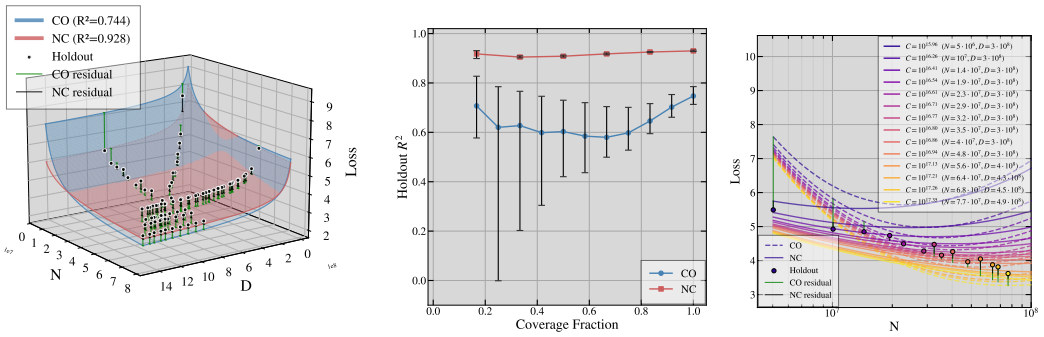
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 10: Chinchilla law on C4, first epoch: CO and NC fits, with the panels of Figure 1.



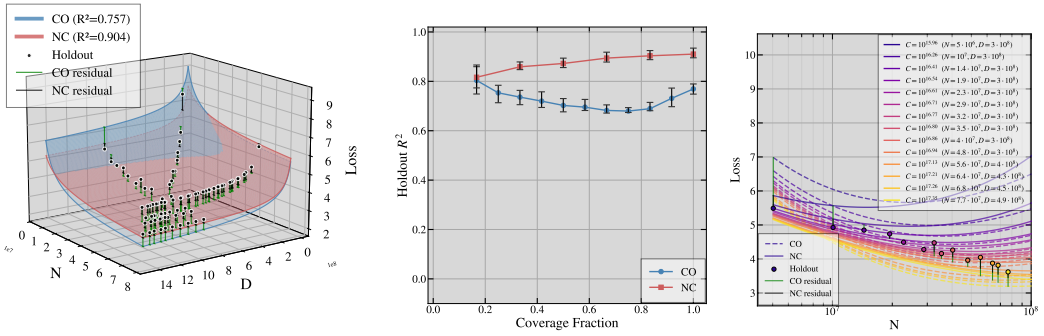
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 11: Droppo-Elibol law on Wikipedia, first epoch: CO and NC fits, with the panels of Figure 1.



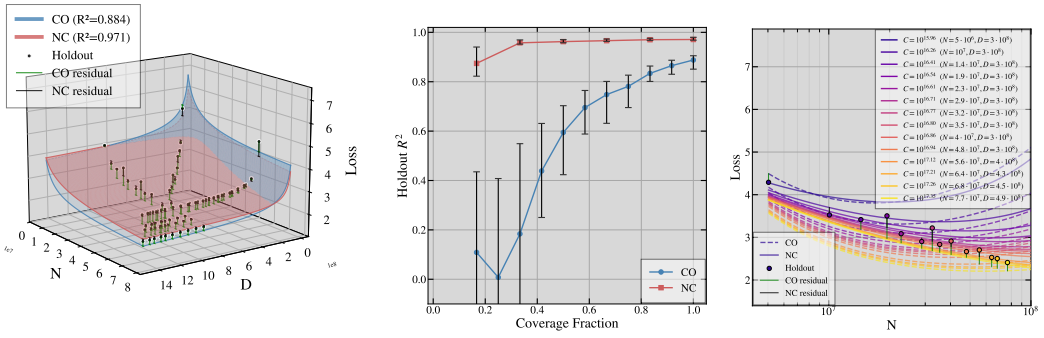
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 12: Droppo-Elibol law on RedPajama, first epoch: CO and NC fits, with the panels of Figure 1.



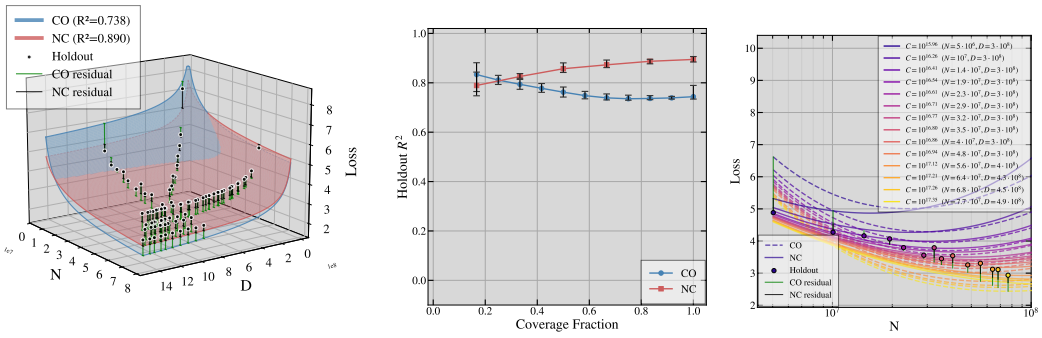
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 13: Chinchilla law on RedPajama, first epoch: CO and NC fits, with the panels of Figure 1.



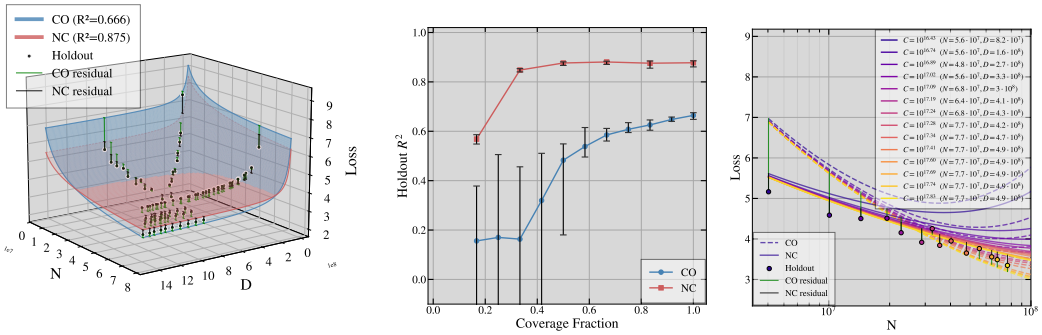
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 14: Droppo-Elibol law on Cosmopedia, first epoch: CO and NC fits, with the panels of Figure 1.



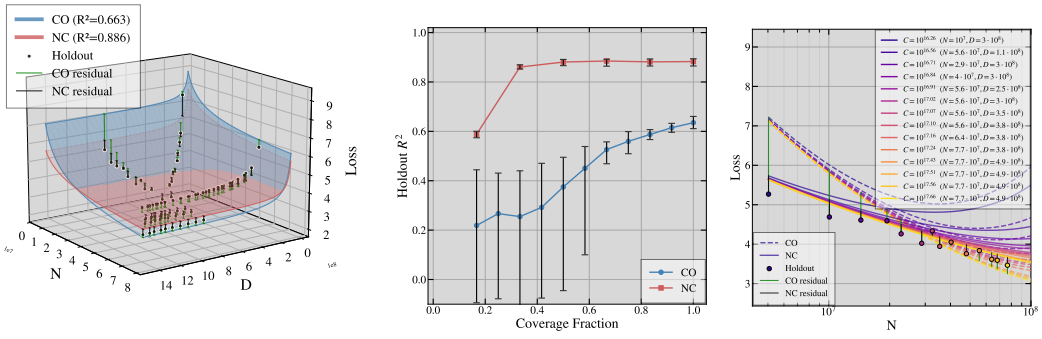
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 15: Chinchilla law on Wikipedia, first epoch: CO and NC fits, with the panels of Figure 1.



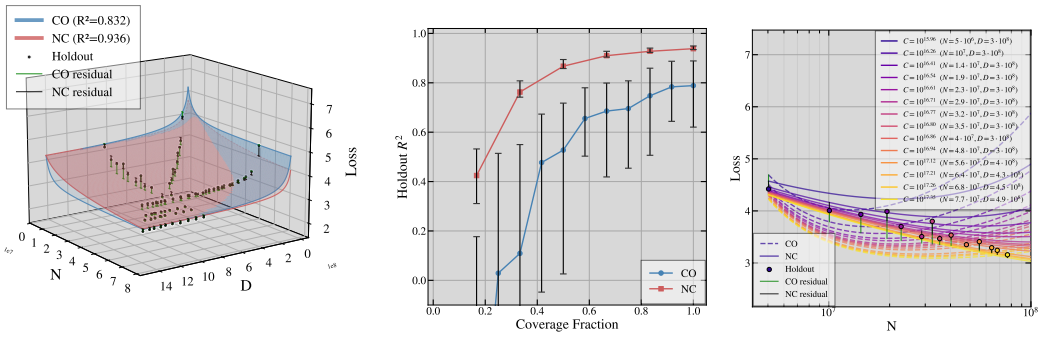
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 16: Kaplan law on RedPajama, final epoch: CO and NC fits, with the panels of Figure 1.



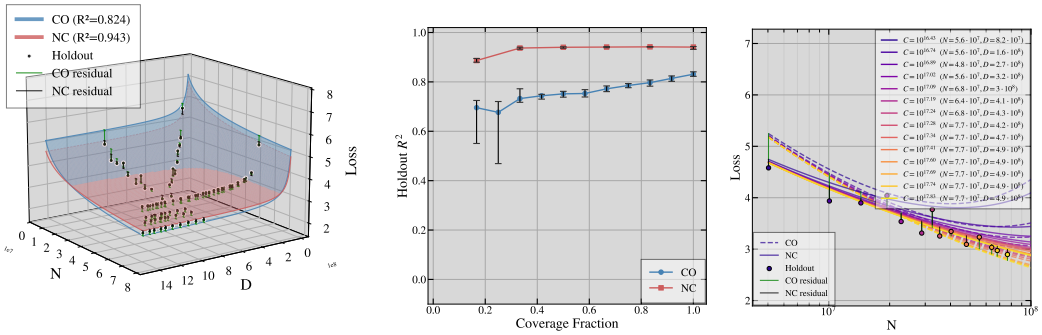
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 17: Kaplan law on RedPajama, second epoch: CO and NC fits, with the panels of Figure 1.



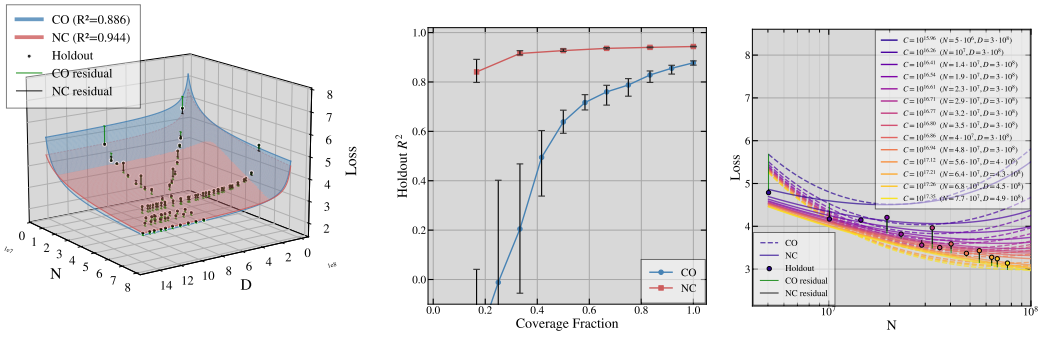
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 18: Droppo-Elibol law on C4, first epoch: CO and NC fits, with the panels of Figure 1.



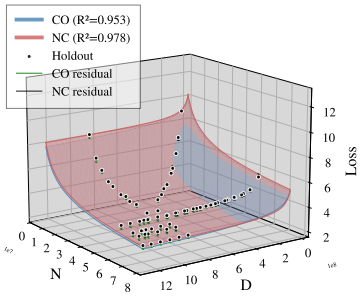
(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 19: Kaplan law on peS2o, final epoch: CO and NC fits, with the panels of Figure 1.

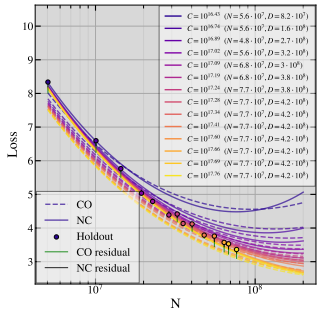


(a) Loss surface. (b) Holdout R^2 vs. TPP coverage. (c) Induced isoFLOP curves.

Figure 20: Droppo-Elibol law on peS2o, first epoch: CO and NC fits, with the panels of Figure 1.

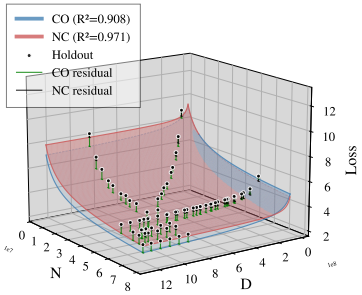


(a) Loss surface.

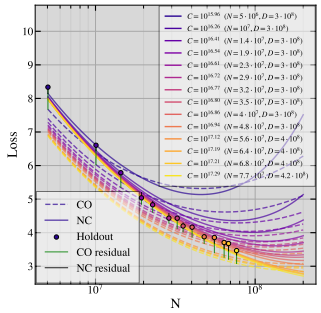


(b) Induced isoFLOP curves.

Figure 21: Chinchilla law on Wikipedia BF16, final epoch: CO and NC fits, with panels (a) and (c) of Figure 1.

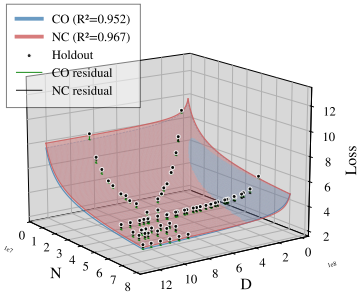


(a) Loss surface.

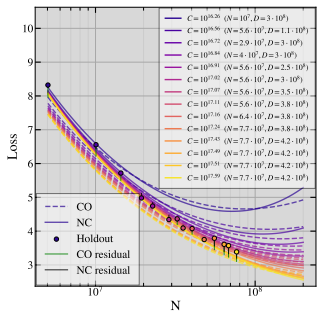


(b) Induced isoFLOP curves.

Figure 22: Chinchilla law on Wikipedia BF16, first epoch: CO and NC fits, with panels (a) and (c) of Figure 1.

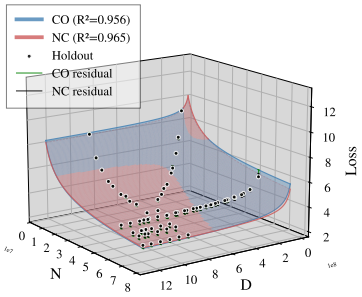


(a) Loss surface.

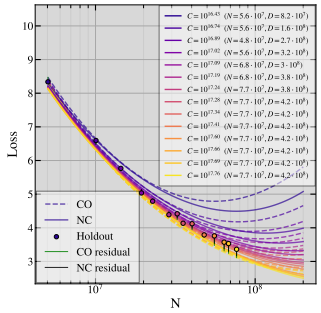


(b) Induced isoFLOP curves.

Figure 23: Chinchilla law on Wikipedia BF16, second epoch: CO and NC fits, with panels (a) and (c) of Figure 1.

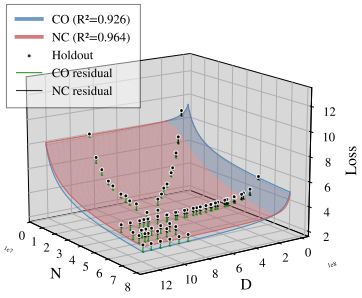


(a) Loss surface.

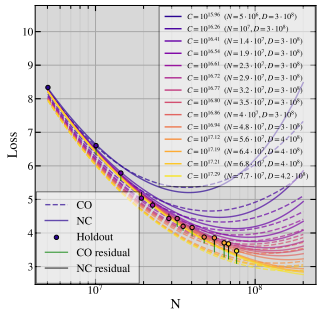


(b) Induced isoFLOP curves.

Figure 24: Droppo-Elibol law on Wikipedia BF16, final epoch: CO and NC fits, with panels (a) and (c) of Figure 1.



(a) Loss surface.



(b) Induced isoFLOP curves.

Figure 25: Droppo-Elibol law on Wikipedia BF16, first epoch: CO and NC fits, with panels (a) and (c) of Figure 1.

G Appendix - Dataset, scaling-law, epoch extended breakdown

The following appendix material shows the full breakdown of our experimental fits at full coverage, disaggregating Table 4 and, in the last three tables, the BF16 Table 22. At full coverage, seed-to-seed variance is small for both designs; the seed spread is widest at lower coverage (convergence panels of Appendix F.8). The CI annotation on each cell is the half-width of the 95% confidence interval on the mean - i.e., the interval is the reported value ± CI.

Table 28: C4 - Unified holdout ($\mathcal{H}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds ± the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}$	RMSE Train	RMSE $\mathcal{H}$
-	Repeated-data	Collinear	0.9781 ± 0.000	0.8183 ± 0.000	0.1644 ± 0.000	0.1844 ± 0.000
		Non-Collinear	0.9361 ± 0.000	**0.9570 ± 0.000**	0.1778 ± 0.000	0.0897 ± 0.000
Final	Chinchilla	Collinear	0.9788 ± 0.000	0.9158 ± 0.014	0.1243 ± 0.001	0.1091 ± 0.009
		Non-Collinear	0.9354 ± 0.000	**0.9520 ± 0.001**	0.1387 ± 0.000	0.0842 ± 0.001
	Droppo-Elibol	Collinear	0.9805 ± 0.000	0.9403 ± 0.003	0.1192 ± 0.000	0.0939 ± 0.002
		Non-Collinear	0.9395 ± 0.000	**0.9544 ± 0.001**	0.1343 ± 0.000	0.0822 ± 0.000
	Kaplan	Collinear	0.9802 ± 0.000	**0.9412 ± 0.001**	0.1203 ± 0.000	0.0933 ± 0.001
		Non-Collinear	0.9271 ± 0.001	0.9385 ± 0.003	0.1473 ± 0.001	0.0953 ± 0.002
First	Chinchilla	Collinear	0.9825 ± 0.000	0.7953 ± 0.004	0.1808 ± 0.000	0.2135 ± 0.002
		Non-Collinear	0.9328 ± 0.000	**0.9362 ± 0.006**	0.2236 ± 0.000	0.1184 ± 0.005
	Droppo-Elibol	Collinear	0.9818 ± 0.001	0.7886 ± 0.029	0.1845 ± 0.003	0.2135 ± 0.014
		Non-Collinear	0.9463 ± 0.000	**0.9382 ± 0.002**	0.1999 ± 0.000	0.1172 ± 0.002
	Kaplan	Collinear	0.9674 ± 0.000	0.7838 ± 0.004	0.2468 ± 0.000	0.2194 ± 0.002
		Non-Collinear	0.9356 ± 0.001	**0.9398 ± 0.001**	0.2189 ± 0.002	0.1158 ± 0.001
Second	Chinchilla	Collinear	0.9799 ± 0.000	0.8857 ± 0.013	0.1345 ± 0.001	0.1303 ± 0.007
		Non-Collinear	0.9380 ± 0.000	**0.9472 ± 0.002**	0.1486 ± 0.000	0.0893 ± 0.001
	Droppo-Elibol	Collinear	0.9808 ± 0.000	0.9282 ± 0.004	0.1316 ± 0.000	0.1040 ± 0.003
		Non-Collinear	0.9432 ± 0.000	**0.9525 ± 0.001**	0.1422 ± 0.000	0.0848 ± 0.001
	Kaplan	Collinear	0.9795 ± 0.000	0.9296 ± 0.001	0.1361 ± 0.000	0.1032 ± 0.001
		Non-Collinear	0.9301 ± 0.001	**0.9395 ± 0.002**	0.1579 ± 0.001	0.0956 ± 0.002

Table 29: C4 - Collinear holdout ($\mathcal{H}_{\text{col}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\text{col}}$	RMSE Train	RMSE $\mathcal{H}_{\text{col}}$
-	Repeated-data	Collinear	0.9781 ± 0.000	0.8571 ± 0.000	0.1644 ± 0.000	0.1642 ± 0.000
		Non-Collinear	0.9361 ± 0.000	**0.9590 ± 0.000**	0.1778 ± 0.000	0.0880 ± 0.000
Final	Chinchilla	Collinear	0.9788 ± 0.000	0.9334 ± 0.014	0.1243 ± 0.001	0.0980 ± 0.009
		Non-Collinear	0.9354 ± 0.000	**0.9595 ± 0.001**	0.1387 ± 0.000	0.0788 ± 0.001
	Droppo-Elibol	Collinear	0.9805 ± 0.000	0.9510 ± 0.002	0.1192 ± 0.000	0.0866 ± 0.002
		Non-Collinear	0.9395 ± 0.000	**0.9631 ± 0.001**	0.1343 ± 0.000	0.0753 ± 0.001
	Kaplan	Collinear	0.9802 ± 0.000	**0.9554 ± 0.001**	0.1203 ± 0.000	0.0827 ± 0.001
		Non-Collinear	0.9271 ± 0.001	0.9456 ± 0.003	0.1473 ± 0.001	0.0912 ± 0.002
First	Chinchilla	Collinear	0.9825 ± 0.000	0.8575 ± 0.003	0.1808 ± 0.000	0.1783 ± 0.002
		Non-Collinear	0.9328 ± 0.000	**0.9402 ± 0.006**	0.2236 ± 0.000	0.1142 ± 0.006
	Droppo-Elibol	Collinear	0.9818 ± 0.001	0.8680 ± 0.013	0.1845 ± 0.003	0.1701 ± 0.008
		Non-Collinear	0.9463 ± 0.000	**0.9398 ± 0.002**	0.1999 ± 0.000	0.1158 ± 0.002
	Kaplan	Collinear	0.9674 ± 0.000	0.8397 ± 0.004	0.2468 ± 0.000	0.1891 ± 0.002
		Non-Collinear	0.9356 ± 0.001	**0.9493 ± 0.001**	0.2189 ± 0.002	0.1064 ± 0.001
Second	Chinchilla	Collinear	0.9799 ± 0.000	0.9023 ± 0.012	0.1345 ± 0.001	0.1193 ± 0.007
		Non-Collinear	0.9380 ± 0.000	**0.9480 ± 0.002**	0.1486 ± 0.000	0.0881 ± 0.001
	Droppo-Elibol	Collinear	0.9808 ± 0.000	0.9343 ± 0.003	0.1316 ± 0.000	0.0990 ± 0.002
		Non-Collinear	0.9432 ± 0.000	**0.9542 ± 0.001**	0.1422 ± 0.000	0.0828 ± 0.001
	Kaplan	Collinear	0.9795 ± 0.000	0.9398 ± 0.001	0.1361 ± 0.000	0.0949 ± 0.001
		Non-Collinear	0.9301 ± 0.001	**0.9417 ± 0.002**	0.1579 ± 0.001	0.0933 ± 0.002

Table 30: C4 - Non-collinear holdout ($\mathcal{H}_{\mathrm{nc}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\mathrm{nc}}$	RMSE Train	RMSE $\mathcal{H}_{\mathrm{nc}}$
-	Repeated-data	Collinear	0.9781 ± 0.000	0.7753 ± 0.000	0.1644 ± 0.000	0.2039 ± 0.000
		Non-Collinear	0.9361 ± 0.000	**0.9547 ± 0.000**	0.1778 ± 0.000	0.0915 ± 0.000
Final	Chinchilla	Collinear	0.9788 ± 0.000	0.8952 ± 0.016	0.1243 ± 0.001	0.1198 ± 0.008
		Non-Collinear	0.9354 ± 0.000	**0.9433 ± 0.001**	0.1387 ± 0.000	0.0897 ± 0.001
	Droppo-Elibol	Collinear	0.9805 ± 0.000	0.9277 ± 0.003	0.1192 ± 0.000	0.1011 ± 0.002
		Non-Collinear	0.9395 ± 0.000	**0.9441 ± 0.001**	0.1343 ± 0.000	0.0890 ± 0.000
	Kaplan	Collinear	0.9802 ± 0.000	0.9244 ± 0.002	0.1203 ± 0.000	0.1035 ± 0.001
		Non-Collinear	0.9271 ± 0.001	**0.9301 ± 0.003**	0.1473 ± 0.001	0.0994 ± 0.002
First	Chinchilla	Collinear	0.9825 ± 0.000	0.7276 ± 0.004	0.1808 ± 0.000	0.2459 ± 0.002
		Non-Collinear	0.9328 ± 0.000	**0.9317 ± 0.005**	0.2236 ± 0.000	0.1226 ± 0.004
	Droppo-Elibol	Collinear	0.9818 ± 0.001	0.7023 ± 0.048	0.1845 ± 0.003	0.2516 ± 0.019
		Non-Collinear	0.9463 ± 0.000	**0.9364 ± 0.002**	0.1999 ± 0.000	0.1187 ± 0.002
	Kaplan	Collinear	0.9674 ± 0.000	0.7231 ± 0.005	0.2468 ± 0.000	0.2480 ± 0.002
		Non-Collinear	0.9356 ± 0.001	**0.9294 ± 0.001**	0.2189 ± 0.002	0.1252 ± 0.001
Second	Chinchilla	Collinear	0.9799 ± 0.000	0.8677 ± 0.013	0.1345 ± 0.001	0.1411 ± 0.006
		Non-Collinear	0.9380 ± 0.000	**0.9462 ± 0.001**	0.1486 ± 0.000	0.0906 ± 0.001
	Droppo-Elibol	Collinear	0.9808 ± 0.000	0.9216 ± 0.005	0.1316 ± 0.000	0.1092 ± 0.003
		Non-Collinear	0.9432 ± 0.000	**0.9505 ± 0.001**	0.1422 ± 0.000	0.0870 ± 0.001
	Kaplan	Collinear	0.9795 ± 0.000	0.9185 ± 0.002	0.1361 ± 0.000	0.1116 ± 0.001
		Non-Collinear	0.9301 ± 0.001	**0.9368 ± 0.003**	0.1579 ± 0.001	0.0981 ± 0.002

Table 31: Cosmopedia - Unified holdout ($\mathcal{H}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}$	RMSE Train	RMSE $\mathcal{H}$
-	Repeated-data	Collinear	0.9841 ± 0.000	0.9496 ± 0.000	0.1805 ± 0.000	0.1333 ± 0.000
		Non-Collinear	0.9656 ± 0.000	**0.9684 ± 0.000**	0.1862 ± 0.000	0.1056 ± 0.000
Final	Chinchilla	Collinear	0.9857 ± 0.000	0.9450 ± 0.005	0.1401 ± 0.001	0.1225 ± 0.005
		Non-Collinear	0.9502 ± 0.000	**0.9580 ± 0.003**	0.1807 ± 0.000	0.1073 ± 0.004
	Droppo-Elibol	Collinear	0.9861 ± 0.000	0.9581 ± 0.001	0.1382 ± 0.000	0.1076 ± 0.002
		Non-Collinear	0.9545 ± 0.000	**0.9624 ± 0.001**	0.1728 ± 0.000	0.1020 ± 0.002
	Kaplan	Collinear	0.9847 ± 0.000	0.9428 ± 0.000	0.1446 ± 0.000	0.1259 ± 0.000
		Non-Collinear	0.9500 ± 0.000	**0.9558 ± 0.001**	0.1811 ± 0.000	0.1105 ± 0.001
First	Chinchilla	Collinear	0.9858 ± 0.000	0.8959 ± 0.003	0.2020 ± 0.000	0.2115 ± 0.003
		Non-Collinear	0.9741 ± 0.000	**0.9626 ± 0.003**	0.1947 ± 0.001	0.1260 ± 0.005
	Droppo-Elibol	Collinear	0.9858 ± 0.000	0.8880 ± 0.009	0.2023 ± 0.001	0.2185 ± 0.008
		Non-Collinear	0.9826 ± 0.000	**0.9711 ± 0.001**	0.1598 ± 0.000	0.1114 ± 0.003
	Kaplan	Collinear	0.9814 ± 0.000	0.7448 ± 0.015	0.2313 ± 0.001	0.3306 ± 0.009
		Non-Collinear	0.9704 ± 0.000	**0.9661 ± 0.001**	0.2083 ± 0.000	0.1207 ± 0.001
Second	Chinchilla	Collinear	0.9867 ± 0.000	0.9504 ± 0.002	0.1464 ± 0.001	0.1214 ± 0.003
		Non-Collinear	0.9574 ± 0.000	**0.9573 ± 0.006**	0.1812 ± 0.001	0.1114 ± 0.006
	Droppo-Elibol	Collinear	0.9869 ± 0.000	0.9562 ± 0.003	0.1451 ± 0.001	0.1138 ± 0.003
		Non-Collinear	0.9615 ± 0.000	**0.9676 ± 0.001**	0.1722 ± 0.000	0.0981 ± 0.002
	Kaplan	Collinear	0.9846 ± 0.000	0.9284 ± 0.001	0.1575 ± 0.000	0.1461 ± 0.001
		Non-Collinear	0.9569 ± 0.000	**0.9618 ± 0.001**	0.1822 ± 0.000	0.1066 ± 0.001

Table 32: Cosmopedia - Collinear holdout ($\mathcal{H}_{\text{col}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	$\mathbf{R^2}$ Train	$\mathbf{R^2}$ $\mathcal{H}_{\text{col}}$	RMSE Train	RMSE $\mathcal{H}_{\text{col}}$
-	Repeated-data	Collinear	0.9841 ± 0.000	0.9646 ± 0.000	0.1805 ± 0.000	0.1137 ± 0.000
		Non-Collinear	0.9656 ± 0.000	$\mathbf{0.9719 \pm 0.000}$	0.1862 ± 0.000	0.1014 ± 0.000
Final	Chinchilla	Collinear	0.9857 ± 0.000	0.9573 ± 0.004	0.1401 ± 0.001	0.1119 ± 0.005
		Non-Collinear	0.9502 ± 0.000	$\mathbf{0.9618 \pm 0.003}$	0.1807 ± 0.000	0.1060 ± 0.004
	Droppo-Elibol	Collinear	0.9861 ± 0.000	0.9651 ± 0.001	0.1382 ± 0.000	0.1019 ± 0.001
		Non-Collinear	0.9545 ± 0.000	$\mathbf{0.9658 \pm 0.001}$	0.1728 ± 0.000	0.1007 ± 0.002
	Kaplan	Collinear	0.9847 ± 0.000	$\mathbf{0.9593 \pm 0.000}$	0.1446 ± 0.000	0.1100 ± 0.000
		Non-Collinear	0.9500 ± 0.000	0.9578 ± 0.001	0.1811 ± 0.000	0.1121 ± 0.001
First	Chinchilla	Collinear	0.9858 ± 0.000	0.9100 ± 0.003	0.2020 ± 0.000	0.1980 ± 0.004
		Non-Collinear	0.9741 ± 0.000	$\mathbf{0.9693 \pm 0.003}$	0.1947 ± 0.001	0.1145 ± 0.006
	Droppo-Elibol	Collinear	0.9858 ± 0.000	0.9041 ± 0.008	0.2023 ± 0.001	0.2034 ± 0.008
		Non-Collinear	0.9826 ± 0.000	$\mathbf{0.9736 \pm 0.001}$	0.1598 ± 0.000	0.1070 ± 0.003
	Kaplan	Collinear	0.9814 ± 0.000	0.8048 ± 0.014	0.2313 ± 0.001	0.2909 ± 0.009
		Non-Collinear	0.9704 ± 0.000	$\mathbf{0.9681 \pm 0.001}$	0.2083 ± 0.000	0.1179 ± 0.001
Second	Chinchilla	Collinear	0.9867 ± 0.000	$\mathbf{0.9612 \pm 0.002}$	0.1464 ± 0.001	0.1093 ± 0.003
		Non-Collinear	0.9574 ± 0.000	0.9590 ± 0.006	0.1812 ± 0.001	0.1110 ± 0.007
	Droppo-Elibol	Collinear	0.9869 ± 0.000	0.9650 ± 0.001	0.1451 ± 0.001	0.1039 ± 0.002
		Non-Collinear	0.9615 ± 0.000	$\mathbf{0.9695 \pm 0.001}$	0.1722 ± 0.000	0.0968 ± 0.002
	Kaplan	Collinear	0.9846 ± 0.000	0.9483 ± 0.001	0.1575 ± 0.000	0.1264 ± 0.001
		Non-Collinear	0.9569 ± 0.000	$\mathbf{0.9621 \pm 0.001}$	0.1822 ± 0.000	0.1081 ± 0.001

Table 33: Cosmopedia - Non-collinear holdout ($\mathcal{H}_{\mathrm{nc}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\mathrm{nc}}$	RMSE Train	RMSE $\mathcal{H}_{\mathrm{nc}}$
-	Repeated-data	Collinear	0.9841 ± 0.000	0.9335 ± 0.000	0.1805 ± 0.000	0.1504 ± 0.000
		Non-Collinear	0.9656 ± 0.000	**0.9646 ± 0.000**	0.1862 ± 0.000	0.1097 ± 0.000
Final	Chinchilla	Collinear	0.9857 ± 0.000	0.9308 ± 0.006	0.1401 ± 0.001	0.1323 ± 0.005
		Non-Collinear	0.9502 ± 0.000	**0.9536 ± 0.004**	0.1807 ± 0.000	0.1085 ± 0.004
	Droppo-Elibol	Collinear	0.9861 ± 0.000	0.9500 ± 0.002	0.1382 ± 0.000	0.1131 ± 0.002
		Non-Collinear	0.9545 ± 0.000	**0.9583 ± 0.001**	0.1728 ± 0.000	0.1033 ± 0.002
	Kaplan	Collinear	0.9847 ± 0.000	0.9235 ± 0.000	0.1446 ± 0.000	0.1400 ± 0.000
		Non-Collinear	0.9500 ± 0.000	**0.9536 ± 0.001**	0.1811 ± 0.000	0.1090 ± 0.001
First	Chinchilla	Collinear	0.9858 ± 0.000	0.8814 ± 0.002	0.2020 ± 0.000	0.2241 ± 0.002
		Non-Collinear	0.9741 ± 0.000	**0.9556 ± 0.003**	0.1947 ± 0.001	0.1365 ± 0.005
	Droppo-Elibol	Collinear	0.9858 ± 0.000	0.8713 ± 0.013	0.2023 ± 0.001	0.2317 ± 0.011
		Non-Collinear	0.9826 ± 0.000	**0.9684 ± 0.001**	0.1598 ± 0.000	0.1155 ± 0.002
	Kaplan	Collinear	0.9814 ± 0.000	0.6829 ± 0.016	0.2313 ± 0.001	0.3660 ± 0.008
		Non-Collinear	0.9704 ± 0.000	**0.9641 ± 0.001**	0.2083 ± 0.000	0.1234 ± 0.001
Second	Chinchilla	Collinear	0.9867 ± 0.000	0.9388 ± 0.003	0.1464 ± 0.001	0.1323 ± 0.003
		Non-Collinear	0.9574 ± 0.000	**0.9555 ± 0.006**	0.1812 ± 0.001	0.1117 ± 0.006
	Droppo-Elibol	Collinear	0.9869 ± 0.000	0.9468 ± 0.005	0.1451 ± 0.001	0.1228 ± 0.005
		Non-Collinear	0.9615 ± 0.000	**0.9656 ± 0.001**	0.1722 ± 0.000	0.0993 ± 0.002
	Kaplan	Collinear	0.9846 ± 0.000	0.9070 ± 0.001	0.1575 ± 0.000	0.1634 ± 0.001
		Non-Collinear	0.9569 ± 0.000	**0.9615 ± 0.001**	0.1822 ± 0.000	0.1051 ± 0.001

Table 34: peS2o - Unified holdout ($\mathcal{H}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}$	RMSE Train	RMSE $\mathcal{H}$
-	Repeated-data	Collinear	0.9862 ± 0.000	0.9232 ± 0.000	0.1588 ± 0.000	0.1545 ± 0.000
		Non-Collinear	0.9616 ± 0.000	**0.9464 ± 0.000**	0.1836 ± 0.000	0.1290 ± 0.000
Final	Chinchilla	Collinear	0.9829 ± 0.000	0.9055 ± 0.002	0.1463 ± 0.000	0.1446 ± 0.002
		Non-Collinear	0.9367 ± 0.000	**0.9421 ± 0.002**	0.1833 ± 0.000	0.1132 ± 0.002
	Droppo-Elibol	Collinear	0.9833 ± 0.000	0.9081 ± 0.004	0.1447 ± 0.000	0.1425 ± 0.003
		Non-Collinear	0.9426 ± 0.000	**0.9520 ± 0.001**	0.1744 ± 0.001	0.1031 ± 0.001
	Kaplan	Collinear	0.9810 ± 0.000	0.8310 ± 0.002	0.1543 ± 0.000	0.1935 ± 0.001
		Non-Collinear	0.9354 ± 0.000	**0.9409 ± 0.002**	0.1851 ± 0.001	0.1143 ± 0.002
First	Chinchilla	Collinear	0.9887 ± 0.000	0.8821 ± 0.004	0.1674 ± 0.000	0.2192 ± 0.003
		Non-Collinear	0.9734 ± 0.000	**0.9272 ± 0.006**	0.1873 ± 0.001	0.1711 ± 0.007
	Droppo-Elibol	Collinear	0.9887 ± 0.000	0.8776 ± 0.002	0.1675 ± 0.000	0.2234 ± 0.002
		Non-Collinear	0.9810 ± 0.000	**0.9435 ± 0.000**	0.1586 ± 0.000	0.1517 ± 0.001
	Kaplan	Collinear	0.9845 ± 0.000	0.7131 ± 0.004	0.1963 ± 0.000	0.3420 ± 0.002
		Non-Collinear	0.9676 ± 0.000	**0.9444 ± 0.001**	0.2068 ± 0.001	0.1506 ± 0.001
Second	Chinchilla	Collinear	0.9863 ± 0.000	0.9108 ± 0.003	0.1408 ± 0.000	0.1483 ± 0.002
		Non-Collinear	0.9484 ± 0.000	**0.9389 ± 0.004**	0.1822 ± 0.001	0.1223 ± 0.004
	Droppo-Elibol	Collinear	0.9864 ± 0.000	0.9105 ± 0.004	0.1404 ± 0.000	0.1484 ± 0.003
		Non-Collinear	0.9546 ± 0.000	**0.9534 ± 0.001**	0.1709 ± 0.000	0.1072 ± 0.001
	Kaplan	Collinear	0.9834 ± 0.000	0.8129 ± 0.002	0.1553 ± 0.000	0.2150 ± 0.001
		Non-Collinear	0.9455 ± 0.000	**0.9457 ± 0.001**	0.1872 ± 0.001	0.1158 ± 0.001

Table 35: peS2o - Collinear holdout ($\mathcal{H}_{\text{col}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\text{col}}$	RMSE Train	RMSE $\mathcal{H}_{\text{col}}$
-	Repeated-data	Collinear	0.9862 ± 0.000	0.9495 ± 0.000	0.1588 ± 0.000	0.1312 ± 0.000
		Non-Collinear	0.9616 ± 0.000	**0.9513 ± 0.000**	0.1836 ± 0.000	0.1290 ± 0.000
Final	Chinchilla	Collinear	0.9829 ± 0.000	0.9370 ± 0.003	0.1463 ± 0.000	0.1263 ± 0.003
		Non-Collinear	0.9367 ± 0.000	**0.9538 ± 0.002**	0.1833 ± 0.000	0.1082 ± 0.002
	Droppo-Elibol	Collinear	0.9833 ± 0.000	0.9302 ± 0.003	0.1447 ± 0.000	0.1330 ± 0.003
		Non-Collinear	0.9426 ± 0.000	**0.9620 ± 0.001**	0.1744 ± 0.001	0.0983 ± 0.001
	Kaplan	Collinear	0.9810 ± 0.000	0.8872 ± 0.002	0.1543 ± 0.000	0.1694 ± 0.001
		Non-Collinear	0.9354 ± 0.000	**0.9526 ± 0.002**	0.1851 ± 0.001	0.1097 ± 0.002
First	Chinchilla	Collinear	0.9887 ± 0.000	0.9107 ± 0.004	0.1674 ± 0.000	0.1979 ± 0.004
		Non-Collinear	0.9734 ± 0.000	**0.9314 ± 0.006**	0.1873 ± 0.001	0.1726 ± 0.007
	Droppo-Elibol	Collinear	0.9887 ± 0.000	0.9077 ± 0.006	0.1675 ± 0.000	0.2009 ± 0.006
		Non-Collinear	0.9810 ± 0.000	**0.9451 ± 0.000**	0.1586 ± 0.000	0.1555 ± 0.001
	Kaplan	Collinear	0.9845 ± 0.000	0.7955 ± 0.003	0.1963 ± 0.000	0.3000 ± 0.002
		Non-Collinear	0.9676 ± 0.000	**0.9417 ± 0.001**	0.2068 ± 0.001	0.1602 ± 0.001
Second	Chinchilla	Collinear	0.9863 ± 0.000	0.9369 ± 0.003	0.1408 ± 0.000	0.1311 ± 0.003
		Non-Collinear	0.9484 ± 0.000	**0.9482 ± 0.004**	0.1822 ± 0.001	0.1185 ± 0.004
	Droppo-Elibol	Collinear	0.9864 ± 0.000	0.9338 ± 0.004	0.1404 ± 0.000	0.1342 ± 0.004
		Non-Collinear	0.9546 ± 0.000	**0.9607 ± 0.001**	0.1709 ± 0.000	0.1037 ± 0.001
	Kaplan	Collinear	0.9834 ± 0.000	0.8692 ± 0.002	0.1553 ± 0.000	0.1893 ± 0.001
		Non-Collinear	0.9455 ± 0.000	**0.9517 ± 0.001**	0.1872 ± 0.001	0.1149 ± 0.002

Table 36: peS2o - Non-collinear holdout ($\mathcal{H}_{\text{nc}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\text{nc}}$	RMSE Train	RMSE $\mathcal{H}_{\text{nc}}$
-	Repeated-data	Collinear	0.9862 ± 0.000	0.8911 ± 0.000	0.1588 ± 0.000	0.1747 ± 0.000
		Non-Collinear	0.9616 ± 0.000	**0.9405 ± 0.000**	0.1836 ± 0.000	0.1291 ± 0.000
Final	Chinchilla	Collinear	0.9829 ± 0.000	0.8631 ± 0.002	0.1463 ± 0.000	0.1607 ± 0.001
		Non-Collinear	0.9367 ± 0.000	**0.9263 ± 0.002**	0.1833 ± 0.000	0.1179 ± 0.002
	Droppo-Elibol	Collinear	0.9833 ± 0.000	0.8783 ± 0.006	0.1447 ± 0.000	0.1513 ± 0.003
		Non-Collinear	0.9426 ± 0.000	**0.9385 ± 0.001**	0.1744 ± 0.001	0.1077 ± 0.001
	Kaplan	Collinear	0.9810 ± 0.000	0.7552 ± 0.002	0.1543 ± 0.000	0.2149 ± 0.001
		Non-Collinear	0.9354 ± 0.000	**0.9252 ± 0.002**	0.1851 ± 0.001	0.1187 ± 0.001
First	Chinchilla	Collinear	0.9887 ± 0.000	0.8484 ± 0.003	0.1674 ± 0.000	0.2385 ± 0.003
		Non-Collinear	0.9734 ± 0.000	**0.9223 ± 0.007**	0.1873 ± 0.001	0.1696 ± 0.007
	Droppo-Elibol	Collinear	0.9887 ± 0.000	0.8423 ± 0.005	0.1675 ± 0.000	0.2431 ± 0.004
		Non-Collinear	0.9810 ± 0.000	**0.9417 ± 0.000**	0.1586 ± 0.000	0.1479 ± 0.001
	Kaplan	Collinear	0.9845 ± 0.000	0.6165 ± 0.004	0.1963 ± 0.000	0.3794 ± 0.002
		Non-Collinear	0.9676 ± 0.000	**0.9476 ± 0.002**	0.2068 ± 0.001	0.1401 ± 0.002
Second	Chinchilla	Collinear	0.9863 ± 0.000	0.8782 ± 0.003	0.1408 ± 0.000	0.1637 ± 0.002
		Non-Collinear	0.9484 ± 0.000	**0.9274 ± 0.005**	0.1822 ± 0.001	0.1260 ± 0.004
	Droppo-Elibol	Collinear	0.9864 ± 0.000	0.8814 ± 0.007	0.1404 ± 0.000	0.1610 ± 0.005
		Non-Collinear	0.9546 ± 0.000	**0.9443 ± 0.001**	0.1709 ± 0.000	0.1106 ± 0.001
	Kaplan	Collinear	0.9834 ± 0.000	0.7428 ± 0.003	0.1553 ± 0.000	0.2379 ± 0.001
		Non-Collinear	0.9455 ± 0.000	**0.9381 ± 0.001**	0.1872 ± 0.001	0.1167 ± 0.001

Table 37: RedPajama - Unified holdout ($\mathcal{H}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}$	RMSE Train	RMSE $\mathcal{H}$
-	Repeated-data	Collinear	0.9879 ± 0.000	0.8905 ± 0.000	0.1842 ± 0.000	0.2391 ± 0.000
		Non-Collinear	0.9644 ± 0.000	**0.9290 ± 0.000**	0.2206 ± 0.000	0.1926 ± 0.000
Final	Chinchilla	Collinear	0.9931 ± 0.000	0.8802 ± 0.001	0.1254 ± 0.000	0.2176 ± 0.001
		Non-Collinear	0.9444 ± 0.000	**0.9079 ± 0.002**	0.2272 ± 0.000	0.1908 ± 0.002
	Droppo-Elibol	Collinear	0.9932 ± 0.000	0.8932 ± 0.008	0.1247 ± 0.000	0.2046 ± 0.007
		Non-Collinear	0.9594 ± 0.000	**0.9177 ± 0.001**	0.1941 ± 0.000	0.1803 ± 0.001
	Kaplan	Collinear	0.9866 ± 0.000	0.6643 ± 0.003	0.1750 ± 0.000	0.3642 ± 0.002
		Non-Collinear	0.9189 ± 0.001	**0.8776 ± 0.003**	0.2741 ± 0.002	0.2199 ± 0.002
First	Chinchilla	Collinear	0.9851 ± 0.000	0.7692 ± 0.006	0.2236 ± 0.000	0.3902 ± 0.005
		Non-Collinear	0.9790 ± 0.000	**0.9102 ± 0.004**	0.1992 ± 0.001	0.2430 ± 0.006
	Droppo-Elibol	Collinear	0.9852 ± 0.000	0.7471 ± 0.009	0.2226 ± 0.000	0.4082 ± 0.007
		Non-Collinear	0.9865 ± 0.000	**0.9299 ± 0.001**	0.1595 ± 0.000	0.2152 ± 0.001
	Kaplan	Collinear	0.9838 ± 0.000	0.5222 ± 0.005	0.2332 ± 0.000	0.5618 ± 0.003
		Non-Collinear	0.9571 ± 0.000	**0.9035 ± 0.001**	0.2844 ± 0.000	0.2525 ± 0.001
Second	Chinchilla	Collinear	0.9923 ± 0.000	0.8823 ± 0.002	0.1388 ± 0.000	0.2252 ± 0.002
		Non-Collinear	0.9538 ± 0.000	**0.9118 ± 0.004**	0.2235 ± 0.001	0.1945 ± 0.004
	Droppo-Elibol	Collinear	0.9923 ± 0.000	0.8877 ± 0.008	0.1385 ± 0.000	0.2190 ± 0.007
		Non-Collinear	0.9678 ± 0.000	**0.9217 ± 0.001**	0.1866 ± 0.000	0.1836 ± 0.001
	Kaplan	Collinear	0.9856 ± 0.000	0.6351 ± 0.006	0.1898 ± 0.000	0.3964 ± 0.003
		Non-Collinear	0.9265 ± 0.001	**0.8827 ± 0.004**	0.2819 ± 0.002	0.2246 ± 0.004

Table 38: RedPajama - Collinear holdout ($\mathcal{H}_{\text{col}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\text{col}}$	RMSE Train	RMSE $\mathcal{H}_{\text{col}}$
-	Repeated-data	Collinear	0.9879 ± 0.000	**0.9604 ± 0.000**	0.1842 ± 0.000	0.1648 ± 0.000
		Non-Collinear	0.9644 ± 0.000	0.9427 ± 0.000	0.2206 ± 0.000	0.1981 ± 0.000
Final	Chinchilla	Collinear	0.9931 ± 0.000	**0.9672 ± 0.001**	0.1254 ± 0.000	0.1329 ± 0.002
		Non-Collinear	0.9444 ± 0.000	0.9324 ± 0.002	0.2272 ± 0.000	0.1908 ± 0.002
	Droppo-Elibol	Collinear	0.9932 ± 0.000	**0.9635 ± 0.002**	0.1247 ± 0.000	0.1401 ± 0.003
		Non-Collinear	0.9594 ± 0.000	0.9379 ± 0.001	0.1941 ± 0.000	0.1830 ± 0.001
	Kaplan	Collinear	0.9866 ± 0.000	0.8875 ± 0.002	0.1750 ± 0.000	0.2463 ± 0.002
		Non-Collinear	0.9189 ± 0.001	**0.9080 ± 0.003**	0.2741 ± 0.002	0.2226 ± 0.004
First	Chinchilla	Collinear	0.9851 ± 0.000	0.8506 ± 0.005	0.2236 ± 0.000	0.3570 ± 0.006
		Non-Collinear	0.9790 ± 0.000	**0.9292 ± 0.004**	0.1992 ± 0.001	0.2454 ± 0.007
	Droppo-Elibol	Collinear	0.9852 ± 0.000	0.8981 ± 0.007	0.2226 ± 0.000	0.2941 ± 0.009
		Non-Collinear	0.9865 ± 0.000	**0.9416 ± 0.001**	0.1595 ± 0.000	0.2234 ± 0.001
	Kaplan	Collinear	0.9838 ± 0.000	0.8069 ± 0.003	0.2332 ± 0.000	0.4062 ± 0.004
		Non-Collinear	0.9571 ± 0.000	**0.9228 ± 0.000**	0.2844 ± 0.000	0.2569 ± 0.001
Second	Chinchilla	Collinear	0.9923 ± 0.000	**0.9625 ± 0.002**	0.1388 ± 0.000	0.1471 ± 0.003
		Non-Collinear	0.9538 ± 0.000	0.9374 ± 0.003	0.2235 ± 0.001	0.1901 ± 0.005
	Droppo-Elibol	Collinear	0.9923 ± 0.000	**0.9577 ± 0.003**	0.1385 ± 0.000	0.1560 ± 0.005
		Non-Collinear	0.9678 ± 0.000	0.9421 ± 0.001	0.1866 ± 0.000	0.1832 ± 0.001
	Kaplan	Collinear	0.9856 ± 0.000	0.8666 ± 0.003	0.1898 ± 0.000	0.2780 ± 0.004
		Non-Collinear	0.9265 ± 0.001	**0.9099 ± 0.005**	0.2819 ± 0.002	0.2280 ± 0.006

Table 39: RedPajama - Non-collinear holdout ($\mathcal{H}_{\mathrm{nc}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	$\mathbf{R^2}$ Train	$\mathbf{R^2}$ $\mathcal{H}_{\mathrm{nc}}$	RMSE Train	RMSE $\mathcal{H}_{\mathrm{nc}}$
-	Repeated-data	Collinear	0.9879 ± 0.000	0.7560 ± 0.000	0.1842 ± 0.000	0.2953 ± 0.000
		Non-Collinear	0.9644 ± 0.000	**0.9023 ± 0.000**	0.2206 ± 0.000	0.1868 ± 0.000
Final	Chinchilla	Collinear	0.9931 ± 0.000	0.6910 ± 0.001	0.1254 ± 0.000	0.2775 ± 0.001
		Non-Collinear	0.9444 ± 0.000	**0.8540 ± 0.003**	0.2272 ± 0.000	0.1907 ± 0.002
	Droppo-Elibol	Collinear	0.9932 ± 0.000	0.7403 ± 0.026	0.1247 ± 0.000	0.2523 ± 0.012
		Non-Collinear	0.9594 ± 0.000	**0.8734 ± 0.002**	0.1941 ± 0.000	0.1776 ± 0.001
	Kaplan	Collinear	0.9866 ± 0.000	0.1789 ± 0.006	0.1750 ± 0.000	0.4524 ± 0.002
		Non-Collinear	0.9189 ± 0.001	**0.8108 ± 0.002**	0.2741 ± 0.002	0.2171 ± 0.001
First	Chinchilla	Collinear	0.9851 ± 0.000	0.6182 ± 0.007	0.2236 ± 0.000	0.4208 ± 0.004
		Non-Collinear	0.9790 ± 0.000	**0.8750 ± 0.005**	0.1992 ± 0.001	0.2405 ± 0.005
	Droppo-Elibol	Collinear	0.9852 ± 0.000	0.4679 ± 0.035	0.2226 ± 0.000	0.4948 ± 0.017
		Non-Collinear	0.9865 ± 0.000	**0.9081 ± 0.001**	0.1595 ± 0.000	0.2065 ± 0.001
	Kaplan	Collinear	0.9838 ± 0.000	-0.0044 ± 0.009	0.2332 ± 0.000	0.6827 ± 0.003
		Non-Collinear	0.9571 ± 0.000	**0.8675 ± 0.003**	0.2844 ± 0.000	0.2479 ± 0.002
Second	Chinchilla	Collinear	0.9923 ± 0.000	0.7156 ± 0.002	0.1388 ± 0.000	0.2823 ± 0.001
		Non-Collinear	0.9538 ± 0.000	**0.8586 ± 0.005**	0.2235 ± 0.001	0.1988 ± 0.004
	Droppo-Elibol	Collinear	0.9923 ± 0.000	0.7423 ± 0.028	0.1385 ± 0.000	0.2661 ± 0.014
		Non-Collinear	0.9678 ± 0.000	**0.8792 ± 0.001**	0.1866 ± 0.000	0.1840 ± 0.001
	Kaplan	Collinear	0.9856 ± 0.000	0.1543 ± 0.010	0.1898 ± 0.000	0.4868 ± 0.003
		Non-Collinear	0.9265 ± 0.001	**0.8258 ± 0.003**	0.2819 ± 0.002	0.2209 ± 0.002

Table 40: Wikipedia - Unified holdout ($\mathcal{H}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}$	RMSE Train	RMSE $\mathcal{H}$
-	Repeated-data	Collinear	0.9853 ± 0.000	0.8666 ± 0.000	0.2247 ± 0.000	0.2779 ± 0.000
		Non-Collinear	0.9585 ± 0.000	**0.9155 ± 0.000**	0.2594 ± 0.000	0.2212 ± 0.000
Final	Chinchilla	Collinear	0.9918 ± 0.000	0.8609 ± 0.002	0.1507 ± 0.000	0.2462 ± 0.002
		Non-Collinear	0.9372 ± 0.000	**0.8974 ± 0.002**	0.2626 ± 0.000	0.2115 ± 0.002
	Droppo-Elibol	Collinear	0.9919 ± 0.000	0.8701 ± 0.011	0.1500 ± 0.000	0.2366 ± 0.010
		Non-Collinear	0.9567 ± 0.000	**0.9136 ± 0.001**	0.2181 ± 0.000	0.1941 ± 0.001
	Kaplan	Collinear	0.9849 ± 0.000	0.6063 ± 0.003	0.2052 ± 0.000	0.4143 ± 0.002
		Non-Collinear	0.9271 ± 0.000	**0.8850 ± 0.000**	0.2830 ± 0.000	0.2240 ± 0.000
First	Chinchilla	Collinear	0.9823 ± 0.000	0.7438 ± 0.006	0.2704 ± 0.001	0.4360 ± 0.006
		Non-Collinear	0.9750 ± 0.000	**0.8947 ± 0.003**	0.2367 ± 0.001	0.2796 ± 0.004
	Droppo-Elibol	Collinear	0.9823 ± 0.000	0.7450 ± 0.013	0.2704 ± 0.000	0.4342 ± 0.011
		Non-Collinear	0.9847 ± 0.000	**0.9222 ± 0.001**	0.1848 ± 0.000	0.2404 ± 0.001
	Kaplan	Collinear	0.9786 ± 0.001	0.5540 ± 0.025	0.2963 ± 0.009	0.5737 ± 0.017
		Non-Collinear	0.9550 ± 0.000	**0.8903 ± 0.001**	0.3173 ± 0.000	0.2855 ± 0.001
Second	Chinchilla	Collinear	0.9906 ± 0.000	0.8575 ± 0.004	0.1691 ± 0.000	0.2623 ± 0.003
		Non-Collinear	0.9461 ± 0.000	**0.8804 ± 0.003**	0.2640 ± 0.000	0.2404 ± 0.003
	Droppo-Elibol	Collinear	0.9906 ± 0.000	0.8467 ± 0.012	0.1692 ± 0.000	0.2708 ± 0.010
		Non-Collinear	0.9643 ± 0.000	**0.8976 ± 0.001**	0.2148 ± 0.000	0.2225 ± 0.001
	Kaplan	Collinear	0.9843 ± 0.000	0.5652 ± 0.006	0.2193 ± 0.000	0.4584 ± 0.003
		Non-Collinear	0.9314 ± 0.001	**0.8843 ± 0.000**	0.2977 ± 0.002	0.2365 ± 0.000

Table 41: Wikipedia - Collinear holdout ($\mathcal{H}_{\text{col}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\text{col}}$	RMSE Train	RMSE $\mathcal{H}_{\text{col}}$
-	Repeated-data	Collinear	0.9853 ± 0.000	**0.9521 ± 0.000**	0.2247 ± 0.000	0.1930 ± 0.000
		Non-Collinear	0.9585 ± 0.000	0.9317 ± 0.000	0.2594 ± 0.000	0.2304 ± 0.000
Final	Chinchilla	Collinear	0.9918 ± 0.000	**0.9632 ± 0.002**	0.1507 ± 0.000	0.1485 ± 0.004
		Non-Collinear	0.9372 ± 0.000	0.9280 ± 0.001	0.2626 ± 0.000	0.2085 ± 0.002
	Droppo-Elibol	Collinear	0.9919 ± 0.000	**0.9611 ± 0.002**	0.1500 ± 0.000	0.1529 ± 0.004
		Non-Collinear	0.9567 ± 0.000	0.9377 ± 0.001	0.2181 ± 0.000	0.1939 ± 0.001
	Kaplan	Collinear	0.9849 ± 0.000	0.8583 ± 0.002	0.2052 ± 0.000	0.2925 ± 0.002
		Non-Collinear	0.9271 ± 0.000	**0.9099 ± 0.000**	0.2830 ± 0.000	0.2333 ± 0.000
First	Chinchilla	Collinear	0.9823 ± 0.000	0.8384 ± 0.005	0.2704 ± 0.001	0.3984 ± 0.007
		Non-Collinear	0.9750 ± 0.000	**0.9201 ± 0.002**	0.2367 ± 0.001	0.2803 ± 0.004
	Droppo-Elibol	Collinear	0.9823 ± 0.000	0.8361 ± 0.002	0.2704 ± 0.000	0.4016 ± 0.002
		Non-Collinear	0.9847 ± 0.000	**0.9371 ± 0.001**	0.1848 ± 0.000	0.2489 ± 0.001
	Kaplan	Collinear	0.9786 ± 0.001	0.8566 ± 0.018	0.2963 ± 0.009	0.3691 ± 0.025
		Non-Collinear	0.9550 ± 0.000	**0.9162 ± 0.000**	0.3173 ± 0.000	0.2872 ± 0.000
Second	Chinchilla	Collinear	0.9906 ± 0.000	**0.9544 ± 0.003**	0.1691 ± 0.000	0.1733 ± 0.006
		Non-Collinear	0.9461 ± 0.000	0.9138 ± 0.002	0.2640 ± 0.000	0.2394 ± 0.003
	Droppo-Elibol	Collinear	0.9906 ± 0.000	**0.9525 ± 0.005**	0.1692 ± 0.000	0.1762 ± 0.008
		Non-Collinear	0.9643 ± 0.000	0.9240 ± 0.001	0.2148 ± 0.000	0.2248 ± 0.001
	Kaplan	Collinear	0.9843 ± 0.000	0.8332 ± 0.004	0.2193 ± 0.000	0.3330 ± 0.003
		Non-Collinear	0.9314 ± 0.001	**0.9059 ± 0.001**	0.2977 ± 0.002	0.2501 ± 0.001

Table 42: Wikipedia - Non-collinear holdout ($\mathcal{H}_{\mathrm{nc}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	$\mathbf{R^2}$ Train	$\mathbf{R^2}$ $\mathcal{H}_{\mathrm{nc}}$	RMSE Train	RMSE $\mathcal{H}_{\mathrm{nc}}$
-	Repeated-data	Collinear	0.9853 ± 0.000	0.6905 ± 0.000	0.2247 ± 0.000	0.3423 ± 0.000
		Non-Collinear	0.9585 ± 0.000	**0.8817 ± 0.000**	0.2594 ± 0.000	0.2116 ± 0.000
Final	Chinchilla	Collinear	0.9918 ± 0.000	0.6288 ± 0.002	0.1507 ± 0.000	0.3148 ± 0.001
		Non-Collinear	0.9372 ± 0.000	**0.8276 ± 0.003**	0.2626 ± 0.000	0.2145 ± 0.002
	Droppo-Elibol	Collinear	0.9919 ± 0.000	0.6637 ± 0.037	0.1500 ± 0.000	0.2963 ± 0.016
		Non-Collinear	0.9567 ± 0.000	**0.8586 ± 0.002**	0.2181 ± 0.000	0.1943 ± 0.001
	Kaplan	Collinear	0.9849 ± 0.000	0.0343 ± 0.006	0.2052 ± 0.000	0.5078 ± 0.002
		Non-Collinear	0.9271 ± 0.000	**0.8281 ± 0.000**	0.2830 ± 0.000	0.2142 ± 0.000
First	Chinchilla	Collinear	0.9823 ± 0.000	0.5555 ± 0.008	0.2704 ± 0.001	0.4707 ± 0.005
		Non-Collinear	0.9750 ± 0.000	**0.8439 ± 0.004**	0.2367 ± 0.001	0.2788 ± 0.004
	Droppo-Elibol	Collinear	0.9823 ± 0.000	0.5638 ± 0.037	0.2704 ± 0.000	0.4637 ± 0.018
		Non-Collinear	0.9847 ± 0.000	**0.8925 ± 0.001**	0.1848 ± 0.000	0.2315 ± 0.001
	Kaplan	Collinear	0.9786 ± 0.001	-0.0459 ± 0.045	0.2963 ± 0.009	0.7208 ± 0.016
		Non-Collinear	0.9550 ± 0.000	**0.8385 ± 0.002**	0.3173 ± 0.000	0.2838 ± 0.002
Second	Chinchilla	Collinear	0.9906 ± 0.000	0.6424 ± 0.004	0.1691 ± 0.000	0.3277 ± 0.002
		Non-Collinear	0.9461 ± 0.000	**0.8058 ± 0.004**	0.2640 ± 0.000	0.2414 ± 0.003
	Droppo-Elibol	Collinear	0.9906 ± 0.000	0.6117 ± 0.044	0.1692 ± 0.000	0.3375 ± 0.019
		Non-Collinear	0.9643 ± 0.000	**0.8386 ± 0.002**	0.2148 ± 0.000	0.2202 ± 0.001
	Kaplan	Collinear	0.9843 ± 0.000	-0.0303 ± 0.011	0.2193 ± 0.000	0.5563 ± 0.003
		Non-Collinear	0.9314 ± 0.001	**0.8359 ± 0.002**	0.2977 ± 0.002	0.2220 ± 0.001

Table 43: Wikipedia (BF16) - Unified holdout ($\mathcal{H}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}$	RMSE Train	RMSE $\mathcal{H}$
-	Repeated-data	Collinear	0.9882 ± 0.000	0.9410 ± 0.000	0.1990 ± 0.000	0.2574 ± 0.000
		Non-Collinear	0.9856 ± 0.000	**0.9729 ± 0.000**	0.1824 ± 0.000	0.1744 ± 0.000
Final	Chinchilla	Collinear	0.9916 ± 0.000	0.9528 ± 0.001	0.1601 ± 0.000	0.2273 ± 0.003
		Non-Collinear	0.9915 ± 0.000	**0.9675 ± 0.002**	0.1334 ± 0.000	0.1882 ± 0.005
	Droppo-Elibol	Collinear	0.9917 ± 0.000	0.9559 ± 0.001	0.1596 ± 0.000	0.2199 ± 0.002
		Non-Collinear	0.9916 ± 0.000	**0.9664 ± 0.001**	0.1328 ± 0.000	0.1918 ± 0.002
	Kaplan	Collinear	0.9915 ± 0.000	0.9453 ± 0.001	0.1612 ± 0.000	0.2450 ± 0.001
		Non-Collinear	0.9898 ± 0.000	**0.9663 ± 0.000**	0.1458 ± 0.000	0.1923 ± 0.001
First	Chinchilla	Collinear	0.9857 ± 0.000	0.9061 ± 0.003	0.2331 ± 0.000	0.3303 ± 0.006
		Non-Collinear	0.9893 ± 0.000	**0.9765 ± 0.002**	0.1689 ± 0.001	0.1647 ± 0.006
	Droppo-Elibol	Collinear	0.9858 ± 0.000	0.9243 ± 0.002	0.2324 ± 0.000	0.2967 ± 0.004
		Non-Collinear	0.9923 ± 0.000	**0.9614 ± 0.001**	0.1438 ± 0.000	0.2118 ± 0.003
	Kaplan	Collinear	0.9856 ± 0.000	0.9249 ± 0.001	0.2342 ± 0.000	0.2957 ± 0.002
		Non-Collinear	0.9922 ± 0.000	**0.9525 ± 0.000**	0.1446 ± 0.000	0.2351 ± 0.001
Second	Chinchilla	Collinear	0.9920 ± 0.000	0.9550 ± 0.001	0.1578 ± 0.000	0.2217 ± 0.003
		Non-Collinear	0.9918 ± 0.000	**0.9698 ± 0.002**	0.1329 ± 0.001	0.1813 ± 0.006
	Droppo-Elibol	Collinear	0.9920 ± 0.000	0.9571 ± 0.002	0.1574 ± 0.000	0.2165 ± 0.004
		Non-Collinear	0.9921 ± 0.000	**0.9635 ± 0.001**	0.1301 ± 0.000	0.1997 ± 0.003
	Kaplan	Collinear	0.9919 ± 0.000	0.9494 ± 0.001	0.1588 ± 0.000	0.2354 ± 0.001
		Non-Collinear	0.9906 ± 0.000	**0.9600 ± 0.000**	0.1419 ± 0.000	0.2091 ± 0.001

Table 44: Wikipedia (BF16) - Collinear holdout ($\mathcal{H}_{\text{col}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\text{col}}$	RMSE Train	RMSE $\mathcal{H}_{\text{col}}$
-	Repeated-data	Collinear	0.9882 ± 0.000	0.9639 ± 0.000	0.1990 ± 0.000	0.2244 ± 0.000
		Non-Collinear	0.9856 ± 0.000	**0.9752 ± 0.000**	0.1824 ± 0.000	0.1861 ± 0.000
Final	Chinchilla	Collinear	0.9916 ± 0.000	0.9699 ± 0.001	0.1601 ± 0.000	0.2046 ± 0.002
		Non-Collinear	0.9915 ± 0.000	**0.9743 ± 0.001**	0.1334 ± 0.000	0.1887 ± 0.004
	Droppo-Elibol	Collinear	0.9917 ± 0.000	0.9716 ± 0.000	0.1596 ± 0.000	0.1987 ± 0.001
		Non-Collinear	0.9916 ± 0.000	**0.9738 ± 0.001**	0.1328 ± 0.000	0.1906 ± 0.002
	Kaplan	Collinear	0.9915 ± 0.000	**0.9735 ± 0.000**	0.1612 ± 0.000	0.1918 ± 0.001
		Non-Collinear	0.9898 ± 0.000	0.9719 ± 0.000	0.1458 ± 0.000	0.1978 ± 0.001
First	Chinchilla	Collinear	0.9857 ± 0.000	0.9497 ± 0.002	0.2331 ± 0.000	0.2655 ± 0.004
		Non-Collinear	0.9893 ± 0.000	**0.9809 ± 0.001**	0.1689 ± 0.001	0.1633 ± 0.005
	Droppo-Elibol	Collinear	0.9858 ± 0.000	0.9526 ± 0.001	0.2324 ± 0.000	0.2579 ± 0.002
		Non-Collinear	0.9923 ± 0.000	**0.9750 ± 0.001**	0.1438 ± 0.000	0.1871 ± 0.003
	Kaplan	Collinear	0.9856 ± 0.000	0.9560 ± 0.001	0.2342 ± 0.000	0.2484 ± 0.002
		Non-Collinear	0.9922 ± 0.000	**0.9688 ± 0.000**	0.1446 ± 0.000	0.2094 ± 0.001
Second	Chinchilla	Collinear	0.9920 ± 0.000	0.9687 ± 0.001	0.1578 ± 0.000	0.2079 ± 0.002
		Non-Collinear	0.9918 ± 0.000	**0.9748 ± 0.001**	0.1329 ± 0.001	0.1862 ± 0.005
	Droppo-Elibol	Collinear	0.9920 ± 0.000	0.9704 ± 0.001	0.1574 ± 0.000	0.2023 ± 0.002
		Non-Collinear	0.9921 ± 0.000	**0.9712 ± 0.001**	0.1301 ± 0.000	0.1996 ± 0.002
	Kaplan	Collinear	0.9919 ± 0.000	**0.9731 ± 0.000**	0.1588 ± 0.000	0.1929 ± 0.001
		Non-Collinear	0.9906 ± 0.000	0.9679 ± 0.000	0.1419 ± 0.000	0.2106 ± 0.001

Table 45: Wikipedia (BF16) - Non-collinear holdout ($\mathcal{H}_{\mathrm{nc}}$). R^2 and RMSE on the training set and the holdout per (law, epoch) and design: mean over 30 optimizer seeds $\pm$ the half-width of its 95% CI; highlighted: the design with the higher holdout R^2; repeated-data (Epoch -) is fit on all epochs at once.

Epoch	Law	Design	R^2 Train	R^2 $\mathcal{H}_{\mathrm{nc}}$	RMSE Train	RMSE $\mathcal{H}_{\mathrm{nc}}$
-	Repeated-data	Collinear	0.9882 ± 0.000	0.9157 ± 0.000	0.1990 ± 0.000	0.2772 ± 0.000
		Non-Collinear	0.9856 ± 0.000	**0.9697 ± 0.000**	0.1824 ± 0.000	0.1662 ± 0.000
Final	Chinchilla	Collinear	0.9916 ± 0.000	0.9330 ± 0.002	0.1601 ± 0.000	0.2413 ± 0.004
		Non-Collinear	0.9915 ± 0.000	**0.9592 ± 0.002**	0.1334 ± 0.000	0.1879 ± 0.006
	Droppo-Elibol	Collinear	0.9917 ± 0.000	0.9376 ± 0.001	0.1596 ± 0.000	0.2330 ± 0.002
		Non-Collinear	0.9916 ± 0.000	**0.9573 ± 0.001**	0.1328 ± 0.000	0.1926 ± 0.003
	Kaplan	Collinear	0.9915 ± 0.000	0.9133 ± 0.001	0.1612 ± 0.000	0.2747 ± 0.002
		Non-Collinear	0.9898 ± 0.000	**0.9591 ± 0.001**	0.1458 ± 0.000	0.1886 ± 0.001
First	Chinchilla	Collinear	0.9857 ± 0.000	0.8621 ± 0.005	0.2331 ± 0.000	0.3672 ± 0.007
		Non-Collinear	0.9893 ± 0.000	**0.9717 ± 0.002**	0.1689 ± 0.001	0.1656 ± 0.006
	Droppo-Elibol	Collinear	0.9858 ± 0.000	0.8954 ± 0.004	0.2324 ± 0.000	0.3200 ± 0.005
		Non-Collinear	0.9923 ± 0.000	**0.9475 ± 0.001**	0.1438 ± 0.000	0.2267 ± 0.003
	Kaplan	Collinear	0.9856 ± 0.000	0.8934 ± 0.002	0.2342 ± 0.000	0.3233 ± 0.003
		Non-Collinear	0.9922 ± 0.000	**0.9358 ± 0.000**	0.1446 ± 0.000	0.2508 ± 0.001
Second	Chinchilla	Collinear	0.9920 ± 0.000	0.9387 ± 0.002	0.1578 ± 0.000	0.2305 ± 0.004
		Non-Collinear	0.9918 ± 0.000	**0.9632 ± 0.002**	0.1329 ± 0.001	0.1779 ± 0.006
	Droppo-Elibol	Collinear	0.9920 ± 0.000	0.9412 ± 0.003	0.1574 ± 0.000	0.2254 ± 0.006
		Non-Collinear	0.9921 ± 0.000	**0.9539 ± 0.001**	0.1301 ± 0.000	0.1998 ± 0.003
	Kaplan	Collinear	0.9919 ± 0.000	0.9222 ± 0.001	0.1588 ± 0.000	0.2599 ± 0.001
		Non-Collinear	0.9906 ± 0.000	**0.9501 ± 0.001**	0.1419 ± 0.000	0.2081 ± 0.001

H Appendix - Fitted scaling law coefficients

Tables 46-50 report the best-fit parameters obtained by multi-start L-BFGS-B with differential evolution polish, on seed 0, and Tables 51-55 report them for the BF16 runs on Wikipedia. R^2 is reported on four splits: training (R^2_{train}), unified holdout ($R^2_{\mathcal{H}}$), collinear holdout ($R^2_{\mathcal{H}_{\text{col}}}$), and non-collinear holdout ($R^2_{\mathcal{H}_{\text{nc}}}$). Please refer to Appendix E.7 for the optimizer bounds we used. Both designs were fit with the same bounds.

Table 46: Chinchilla fitted parameters: $L(N, D) = E + A\,N^{-\alpha} + B\,D^{-\beta}$. Final epoch.

Dataset	**Design**	E	A	α	B	β	R^2_{train}	$R^2_{\mathcal{H}}$	$R^2_{\mathcal{H}_{\text{col}}}$	$R^2_{\mathcal{H}_{\text{nc}}}$
Wikipedia	CO	1.857	4.26e6	0.908	2.99e4	0.518	0.992	0.864	0.966	0.633
	NC	0	239.156	0.260	5.02e4	0.552	0.937	0.898	0.929	0.829
Cosmopedia	CO	1.367	4249.535	0.471	8.53e6	0.874	0.986	0.950	0.961	0.937
	NC	1.174	871.957	0.370	1.37e7	0.902	0.950	0.952	0.956	0.947
RedPajama	CO	2.400	1.87e5	0.706	2.68e4	0.517	0.993	0.880	0.967	0.692
	NC	0.242	118.949	0.209	5.02e4	0.557	0.944	0.916	0.939	0.865
peS2o	CO	2.117	1.66e4	0.566	2.45e5	0.667	0.983	0.896	0.924	0.857
	NC	0	68.799	0.174	2.73e7	0.951	0.935	0.945	0.957	0.929
C4	CO	0.410	47.594	0.161	3.59e7	0.971	0.979	0.945	0.960	0.927
	NC	1.131	82.645	0.216	3.82e4	0.580	0.934	0.954	0.963	0.944

Table 47: Kaplan fitted parameters: $L(N, D) = \left[(N_c/N)^{\alpha_N/\alpha_D} + D_c/D\right]^{\alpha_D}$. Final epoch. †: (N_c, D_c) is a start draw of the optimizer left unmoved; only the exponents moved (Appendix B.8).

Dataset	**Design**	N_c	D_c	α_N	α_D	α_N/α_D	R^2_{train}	$R^2_{\mathcal{H}}$	$R^2_{\mathcal{H}_{\text{col}}}$	$R^2_{\mathcal{H}_{\text{nc}}}$
Wikipedia	CO	6.78e8	6.90e8	0.382	0.563	0.679	0.985	0.612	0.862	0.045
	NC	1.11e10	2.11e7	0.208	2.000	0.104	0.927	0.884	0.909	0.829
Cosmopedia	CO	1.17e9	1.15e8	0.274	0.841	0.325	0.985	0.942	0.959	0.923
	NC†	2.53e9	6.15e7	0.227	1.075	0.211	0.950	0.959	0.961	0.956
RedPajama	CO	3.41e9	10^9	0.296	0.512	0.577	0.987	0.666	0.888	0.181
	NC†	1.89e11	3.74e7	0.162	1.411	0.115	0.920	0.875	0.904	0.811
peS2o	CO	5.32e9	2.10e8	0.236	0.685	0.345	0.981	0.824	0.881	0.747
	NC	4.67e10	2.76e7	0.169	1.435	0.118	0.936	0.943	0.954	0.928
C4	CO†	1.89e11	3.74e7	0.141	1.192	0.118	0.980	0.943	0.956	0.927
	NC†	1.89e11	3.74e7	0.140	1.100	0.128	0.928	0.940	0.950	0.929

Table 48: Droppo-Elibol fitted parameters (law in Appendix B.9): irreducible loss L_∞, scale parameters (N_C, D_C) and exponents $\alpha_N, \alpha_D, \alpha$. Final epoch. †: (N_C, D_C) is a start draw of the optimizer left unmoved (Appendix E.7).

Dataset	**Design**	L_∞	N_C	D_C	α_N	α_D	α	R^2_{train}	$R^2_{\mathcal{H}}$	$R^2_{\mathcal{H}_{\text{col}}}$	$R^2_{\mathcal{H}_{\text{nc}}}$
Wikipedia	CO	2.094	4.64e7	7.72e8	0.735	0.505	0.720	0.992	0.810	0.967	0.456
	NC	0.114	1.36e8	2.57e7	0.298	0.853	2.000	0.957	0.916	0.940	0.861
Cosmopedia	CO†	0.509	8.02e7	3.34e7	0.358	1.095	1.476	0.986	0.960	0.966	0.952
	NC†	0.556	8.02e7	3.34e7	0.329	1.158	1.481	0.954	0.963	0.966	0.959
RedPajama	CO†	1.995	3.69e6	2.61e7	1.016	0.675	1.918	0.993	0.915	0.963	0.811
	NC	0.054	1.61e9	1.97e7	0.215	0.776	2.000	0.959	0.918	0.939	0.875
peS2o	CO†	1.047	8.02e7	3.34e7	0.363	0.983	1.393	0.983	0.894	0.916	0.863
	NC†	0.534	1.10e8	1.52e7	0.264	1.312	1.833	0.943	0.954	0.964	0.940
C4	CO†	0.468	3.48e8	1.49e7	0.205	1.638	1.794	0.981	0.945	0.955	0.933
	NC†	0.050	1.23e10	7.80e6	0.153	1.266	1.984	0.940	0.954	0.963	0.944

Table 49: Repeated-data fitted scaling-law parameters A, α, B, β, E (law in Appendix B.7). Final epoch.

Dataset	**Design**	A	α	B	β	E
Wikipedia	CO	4.22e5	0.746	2.19e4	0.515	1.538
	NC	194.407	0.244	9.09e4	0.601	1.86e−11
Cosmopedia	CO	7.22e4	0.656	1.45e6	0.787	1.752
	NC	303.693	0.290	1.46e6	0.790	0.618
RedPajama	CO	1.15e5	0.667	1.26e4	0.485	2.164
	NC	97.241	0.192	4.86e4	0.566	5.23e−12
peS2o	CO	3.07e4	0.600	4.75e5	0.723	2.265
	NC	70.386	0.177	8.92e5	0.762	2.25e−11
C4	CO	1.68e6	0.879	1.84e7	0.963	3.088
	NC	34.408	0.136	1.82e5	0.684	2.94e−11

Table 50: Repeated-data fit quality and decay constants R^*_D, R^*_N. Final epoch. Values reported as 0.1 for R^*_N indicate the L-BFGS-B optimizer terminating at the documented lower bound $R^*_N \in [0.1,\ 50.0]$; see Appendix E.7 for the full bound list and an explanation of why R^*_N is not informatively constrained by our grid (epochs vary D but not N).

Dataset	**Design**	R^*_D	R^*_N	R^2_{train}	$R^2_{\mathcal{H}}$	$R^2_{\mathcal{H}_{\text{col}}}$	$R^2_{\mathcal{H}_{\text{nc}}}$
Wikipedia	CO	1.074	0.100	0.985	0.867	0.952	0.691
	NC	1.469	0.100	0.959	0.915	0.932	0.882
Cosmopedia	CO	1.500	0.153	0.984	0.950	0.965	0.933
	NC	1.286	0.100	0.966	0.968	0.972	0.965
RedPajama	CO	1.162	0.100	0.988	0.890	0.960	0.756
	NC	1.624	0.100	0.964	0.929	0.943	0.902
peS2o	CO	1.897	0.119	0.986	0.923	0.950	0.891
	NC	2.188	0.100	0.962	0.946	0.951	0.941
C4	CO	0.801	0.564	0.978	0.818	0.857	0.775
	NC	2.352	0.100	0.936	0.957	0.959	0.955

Table 51: BF16 Chinchilla fitted parameters: $L(N, D) = E + A\,N^{-\alpha} + B\,D^{-\beta}$. Final epoch.

Dataset	Design	E	A	α	B	β	R^2_{train}	$R^2_{\mathcal{H}}$	$R^2_{\mathcal{H}_{\text{col}}}$	$R^2_{\mathcal{H}_{\text{nc}}}$
Wikipedia (BF16)	CO	0	3428.447	0.399	2366.171	0.392	0.992	0.953	0.970	0.933
	NC	0.758	4904.551	0.421	4.82e5	0.713	0.991	0.978	0.981	0.973

Table 52: BF16 Kaplan fitted parameters: $L(N, D) = \left[(N_c/N)^{\alpha_N/\alpha_D} + D_c/D\right]^{\alpha_D}$. Final epoch.

Dataset	Design	N_c	D_c	α_N	α_D	α_N/α_D	R^2_{train}	$R^2_{\mathcal{H}}$	$R^2_{\mathcal{H}_{\text{col}}}$	$R^2_{\mathcal{H}_{\text{nc}}}$
Wikipedia (BF16)	CO	1.19e9	2.84e9	0.397	0.394	1.008	0.992	0.946	0.974	0.914
	NC	1.98e9	1.12e8	0.354	0.823	0.430	0.990	0.966	0.972	0.959

Table 53: BF16 Droppo-Elibol fitted parameters (law in Appendix B.9): irreducible loss L_∞, scale parameters (N_C, D_C) and exponents $\alpha_N, \alpha_D, \alpha$. Final epoch.

Dataset	Design	L_∞	N_C	D_C	α_N	α_D	α	R^2_{train}	$R^2_{\mathcal{H}}$	$R^2_{\mathcal{H}_{\text{col}}}$	$R^2_{\mathcal{H}_{\text{nc}}}$
Wikipedia (BF16)	CO	10^{-6}	9.97e8	1.67e9	0.400	0.388	0.583	0.992	0.956	0.972	0.938
	NC	0.224	1.21e9	1.65e8	0.379	0.612	0.882	0.992	0.965	0.973	0.955

Table 54: BF16 repeated-data fitted scaling-law parameters A, α, B, β, E (law in Appendix B.7). Final epoch.

Dataset	Design	A	α	B	β	E
Wikipedia (BF16)	CO	3125.835	0.392	3300.851	0.427	1.29e−12
	NC	3045.878	0.387	3.49e7	0.995	0.612

Table 55: BF16 repeated-data fit quality and decay constants R^*_D, R^*_N. Final epoch. Values reported as 0.1 for R^*_N indicate the L-BFGS-B optimizer terminating at the documented lower bound $R^*_N \in [0.1,\ 50.0]$; see Appendix E.7.

Dataset	Design	R^*_D	R^*_N	R^2_{train}	$R^2_{\mathcal{H}}$	$R^2_{\mathcal{H}_{\text{col}}}$	$R^2_{\mathcal{H}_{\text{nc}}}$
Wikipedia (BF16)	CO	0.283	0.100	0.988	0.941	0.964	0.916
	NC	0.723	0.100	0.986	0.973	0.975	0.970

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 state five claims: a TPP-diversity threshold for multi-ray designs (Proposition 1), ill-conditioning in its one-ray limit (Proposition 2), CI inflation (Corollary 1), NC-vs-CO holdout ordering (Theorem 1 for one ray, Lemma 3 a sufficient margin for several; Tables 4-7 for multi-ray designs), and Jacobian-geometry generality (Appendix B.3) - each backed by a formal result or empirical table; they are the main contributions of the paper. The introduction states the evidence scale (5M-77M parameters), and the criterion is also checked on published designs (Appendix C) and public training-loss releases (Appendix D).

   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: The Limitations paragraph of Section 5 gives each limit with its reason: Theorem 1 orders one ray and gives no gap size, so the multi-ray ordering is measured; the models are small (5.04-76.5M), and while the rates and the leading-order threshold have no size term, the exact $\kappa_{A,B}$ of a design and the measured gap depend on size, and public training-loss releases do not establish the gap at 1B and above; Corollary 1 assumes i.i.d. Gaussian residuals, and correlated noise with covariance Σ moves its ratio by at most the square root of $\kappa(\Sigma)$, not its rate (Appendix B.11); one architecture, English text, one fixed recipe and pre-training loss bound the breadth, with downstream and cross-domain transfer not validated; the budget-matched advantage softens on Cosmopedia and peS2o, not the full-fan one; no isoFLOP-layout arm and no design at the recipe's spread were run. Section 3.1 states that the leading-order threshold is a screen, the exact $\kappa_{A,B}$ of a planned design deciding. The introduction's contributions state what is proved, with size-free rates, and what is measured at 5M-77M.

   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: All assumptions are stated in each theorem environment. Complete proofs are in Appendix B: Proposition 1 in B.12, Proposition 2 in B.3, Corollary 1 in B.11, Theorem 1 in B.13, Proposition 3 in B.14, and Proposition 4 in B.10.1.

   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: All architectures, hyperparameters, data splits, fitting protocols, and seed strategies are specified in Section 4 and Appendices E-F.6, with fitted coefficients in Appendix H. The released code (`https://github.com/lab-v2/tpp-coverage`) has instructions to reproduce all tables either from pre-computed metrics (CPU only) or from scratch ($\sim$3,624 GPU-hours).

   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: Code: https://github.com/lab-v2/tpp-coverage. Trained LLMs checkpoints and metrics: https://huggingface.co/datasets/leibnitz-lab/colinear_scaling_models. The run tables of the public-release reanalysis: https://huggingface.co/datasets/leibnitz-lab/public_checkpoint_reanalysis. All five training corpora are publicly available under open licenses or terms of use (Appendix E).

   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: All 14 architectures (Table 17), training hyperparameters (Table 18), per-dataset learning rates, data-split procedures, CO/NC grid definitions, holdout partitions, fitting protocol (restart counts, seed strategy), and compute infrastructure are specified in Section 4 and Appendix E.

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: [Yes]

Justification: All summary results report across-combination standard deviations and 95% optimizer-noise intervals over 30 optimizer seeds (Tables 4 and 6; Appendix E.8), the overall win rate includes a 95% Wilson CI over 1,500 seed-paired comparisons, which share one training run per (N, D) and so measure fitting variability only (NC has the lower mean holdout RMSE in 49 of the 50 cells), and the budget-matched subset enumeration (Table 7) reports bootstrap CIs across 22 seeds. A pure-BF16 replication on Wikipedia (Table 6 in the main text and Tables 23-24 in Appendix F.4) independently confirms the same patterns with seed-paired NC win rate 98.7% (296/300). The public-release reanalysis reports 95% bootstrap intervals clustered by size window over 512 seeds (Appendix D).

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.6 reports worker type, memory, runtime, and storage for each experiment family: ∼3,624 GPU-hours for pre-training, ∼5 wall-clock hours (230 workers) for fitting/visualization, and ∼24 hours (480 workers) for subset enumeration, plus

$\sim$2.2 TB total storage. We also disclose that full-project compute exceeded the final reported experiment runs due to pilot and failed sweeps, which these totals do not include (Appendix E.6).

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: The work uses only publicly available corpora, under open licenses or terms of use (Appendix E), involves no human subjects, and releases small-scale checkpoints (5.04-76.5M parameters) that pose no dual-use risk. See Section 5 for a detailed assessment.

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: Section 5 discusses the corrective intent of the work (deflating overconfident scaling-law extrapolations) and Section 5 addresses the negligible misuse risk of the released checkpoints and the marginal indirect risk of accelerating capability development.

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: [Yes]

    Justification: Section 5 describes the safeguards: all corpora are publicly available and pre-filtered, released checkpoints are too small for dual-use capabilities, and the accompanying code targets fitting diagnostics rather than model serving.

    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: Every asset is credited and its license or terms are stated. The card of our reanalysis dataset gives each source's license and ships the Apache-2.0, MIT and CC-BY-4.0 texts; its history is squashed, and no revision holds Farseer's raw run table, whose repository states no license. The five corpora are cited in Appendix E.5 and the four scaling-law formalisms in Table 2. The LLaMA architecture is credited to Touvron et al. (2023). Appendix E gives each training corpus's license or terms of use, read from its dataset card: Cosmopedia Apache-2.0; peS2o ODC-By; C4 ODC-By, with the Common Crawl Terms of Use for its content; Wikipedia CC BY-SA 3.0 and GFDL; RedPajama-V2 no license of its own, the Common Crawl Foundation Terms of Use. The ten public training-loss releases of Appendix D are cited in Table 12, with their sources and licenses in that appendix:

    - the Gemstones releases: MIT;
    - the over-training suite of Gadre et al. (2025) and MAD: MIT, both read through ColPret (Choshen et al., 2025), also MIT;
    - the OLMo ladder and DataDecide: Apache-2.0;
    - the Chinchilla isoFLOP points: the digitization of Besiroglu et al. (2024), CC-BY-4.0, read through ColPret;
    - the GPT-3 curves: ColPret's digitization of a figure of Brown et al. (2020), MIT;
    - the OPT trajectories: ColPret (MIT), which takes the losses from Xia et al. (2023), whose repository states no license;
    - Farseer: its repository states no license; the current version of our dataset keeps its 404 runs restricted to N, D and three loss columns, cited to Li et al. (2025a), and the code downloads the original; the dataset's history is squashed, so no revision holds the raw table;
    - the 1,462 single-epoch runs of Muennighoff et al. (2025) in our run tables, which the screen rejects: Apache-2.0, read through ColPret (MIT);
    - DataDecide's per-group perplexity table, which our dataset copies unchanged: Apache-2.0.

The classification of published designs (Appendix C) reads only training ratios; where it reads them from a released file, that file comes from Farseer (no license stated), Epoch AI (CC-BY-4.0), Muennighoff et al. (Apache-2.0) or ColPret (MIT; its Pythia schedule comes from Pythia, under Apache-2.0).

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: We release trained model checkpoints on HuggingFace (`https://huggingface.co/datasets/leibnitz-lab/colinear_scaling_models`) with accompanying documentation describing model configurations, training corpora, and hyperparameters. Experimental code for reproducing scaling law fits and analyses is released at `https://github.com/lab-v2/tpp-coverage`, including the published-design classification and the public-release reanalysis (`camera_ready/public_checkpoints/`), whose run tables are at `https://huggingface.co/datasets/leibnitz-lab/public_checkpoint_reanalysis`. The training data of our models (RedPajama-V2, Cosmopedia, Wikipedia, peS2o, C4) are publicly available under open licenses or terms of use (Appendix E). The reanalysis dataset holds the rows of the ten public releases and the 1,462 single-epoch runs of Muennighoff et al. (2025) that the screen rejects, each with N, D, the released loss columns and k, and an unchanged copy of DataDecide's per-group perplexity table; Appendix D lists each source and its license. Of Farseer's 404 runs, whose repository states no license, the dataset keeps only N, D and three loss columns, cited to Li et al. (2025a). The dataset card gives each source's license and releases our normalisation under CC-BY-4.0; the code and the checkpoints are under GPL-2.0.

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: This work does not involve crowdsourcing or human subjects.

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: This paper does not involve crowdsourcing nor research with human subjects.

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: [Yes]

Justification: LLM usage is declared in the LLM usage paragraph of Section 5. In short: all research questions, theoretical results, experimental design, and empirical analyses originated with the authors. The primary non-trivial use of LLMs was as a sounding board for candidate proof steps - the authors drafted the structure of each derivation, asked an LLM to propose intermediate algebraic manipulations, and independently verified, rewrote, or discarded each suggestion before incorporation. Auxiliary uses included exposition, LaTeX typesetting, boilerplate code for data loading, training, analysis, and visualization, and prose editing - all reviewed by the authors. We used Anthropic's Claude Opus 4.7, 5 and 5.5 and the Cursor coding agent.

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.