How Large a Break Can Hide in a Scaling Law?

Anonymous authors
Paper under double-blind review

Abstract

Scaling laws guide extrapolation from a small set of training runs, but a smooth fit does not establish that the exponent is stable. We study the detection of a continuous change in exponent when the breakpoint is unknown. For designs with 4–20 model sizes, comparing a scanned-breakpoint statistic with a fixed-break F distribution gives false-positive rates of 18.1–32.0% at a nominal 5% level. Simulating the complete scan under the null restores rates to 5.0–5.3%. We characterize the test's invariance to the baseline exponent and noise scale, and quantify detection power over sample size, noise, and break location. Across two decades of model size, designs with 6–8 observations and 2–5% relative noise require exponent changes of 0.1–0.4 on the tested grid to reach 80% power. Power decreases when the break moves toward either end of the range. A six-model Pythia case study illustrates a separate issue: adding a loss floor reduces log-scale residual sum of squares by a factor of 53, showing how smooth curvature can produce a strong apparent break under a pure power law. We provide a reproducible calibration and power-analysis procedure, and recommend reporting detectable changes alongside fitted exponents.

1 Introduction

Neural scaling laws summarize how loss changes with model size, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Their practical value often comes from extrapolation: a curve fitted to smaller runs helps choose the scale and allocation of a larger run. This use depends on the functional form remaining appropriate beyond the observations.

A fitted power law can look smooth even when the exponent changes within the measured range. A continuous change in exponent produces a kink in log–log coordinates, rather than a jump in loss. Detecting that kink is a change-point problem. When only a few model sizes are available, two quantities matter: the false-positive rate of the test and the smallest change it can detect with useful probability.

We study a continuous segmented regression with an unknown breakpoint. The statistic compares a single line with the best two-segment fit over a fixed scan grid. Treating the selected breakpoint as fixed gives the wrong reference distribution: the scan has already selected the location with the largest improvement. This is a classical unidentified-nuisance-parameter problem (Davies, 1987; Andrews & Ploberger, 1994; Hansen, 2000). Our focus is its finite-sample behavior at designs used to measure scaling curves.

Calibration addresses the search over breakpoints. It does not address the choice of scaling function. In particular, the form $L = E + AN^{-\alpha}$ flattens in log–log coordinates as loss approaches E, even with a constant exponent. A test against a pure power law can detect this curvature. We use measured Pythia losses to separate this specification effect from the calibration and power questions.

The paper makes three contributions. First, we give a simulation-calibrated scan test and establish the invariances that allow its null distribution to be reused for a fixed design. Second, we measure its size and power across sample sizes, noise levels, and breakpoint locations, with complete analysis code. Third, we connect the test to a practical reporting procedure that states the fitted form, the calibration, and the detectable exponent change. The goal is to make the sensitivity of a scaling study explicit, rather than infer stability from goodness of fit alone.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 RELATED WORK

Scaling functions and estimation. Power laws describe empirical trends across model and data scales (Kaplan et al., 2020; Hoffmann et al., 2022). Theoretical explanations relate these trends to data geometry and skill acquisition (Bahri et al., 2024; Sharma & Kaplan, 2022; Michaud et al., 2023). Estimation choices also affect the fitted law. Besiroglu et al. (2024) revisit a widely used compute-optimal fit, and Choshen et al. (2024) examine design choices in scaling studies. The survey of Li et al. (2025) documents variation in functional forms, training setups, data collection, and fitting procedures. Parameter counting and optimizer settings help explain differences between published fits (Pearce & Song, 2024; Porian et al., 2024). We study the sensitivity of a specified break test after these modeling choices have been made.

Changes in scaling behavior. Broken Neural Scaling Laws model smooth transitions between power-law regimes (Caballero et al., 2023). Other work studies apparent discontinuities created by the evaluation metric (Schaeffer et al., 2023) and changes in language-model scaling at larger sizes (Chen et al., 2025). Our alternative is a continuous, sharp change in log–log slope. We ask how often a test detects such a change at a given design, rather than select a multi-regime function for extrapolation. A change in fitted slope is also distinct from evidence for a change in the mechanism learned by a model.

Unknown breakpoints. When a parameter exists only under the alternative, standard reference distributions need not apply (Davies, 1987). Tests with an unidentified nuisance parameter and threshold regression have an extensive statistical literature (Andrews & Ploberger, 1994; Hansen, 2000); segmented regression is also widely used in applied work (Muggeo, 2003). We do not propose a new general change-point test. We use this framework to quantify what a small scaling design can detect, including the consequences of searching over the breakpoint.

3 MODEL AND TEST

Let $N_1 < \cdots < N_n$ be measured model sizes and set $x_i = \ln(N_i/N_1)$. We use natural logarithms throughout. The observation model is

$$y_i = \ln L_i = \beta - \alpha x_i - \delta(x_i - c)_+ + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{1}$$

where $(u)_+ = \max(u, 0)$, c is the breakpoint, and δ is the change in exponent. The two segments meet at c. Positive δ means a steeper loss decrease after the break. At the noise levels considered here, σ approximates relative loss noise: the coefficient of variation of the multiplicative lognormal error is $\sqrt{\exp(\sigma^2) - 1}$.

The null is $\delta = 0$. For each candidate c in a prespecified grid $\mathcal{C}$, we fit the three linear coefficients by least squares. Write SSE_0 for the single-line residual sum of squares and SSE_1 for the smallest segmented residual sum of squares over $\mathcal{C}$. The scan statistic is

$$T = \frac{\mathrm{SSE}_0 - \mathrm{SSE}_1}{\mathrm{SSE}_1/(n-3)}. \tag{2}$$

For a fixed, full-rank breakpoint, the corresponding statistic has an $F(1, n-3)$ null distribution under Eq. 1. Selecting the breakpoint by minimizing SSE_1 instead takes the maximum of correlated fixed-break statistics. The resulting distribution depends on the design and the scan grid.

Simulation design. We use $n \in \{4, 6, 8, 12, 20\}$ equally spaced log sizes spanning two decades, so $x_n = 2 \ln 10$. The scan uses 60 equally spaced candidates from x_2 to x_{n-1}, inclusive. Noise levels are $\sigma \in \{0.01, 0.02, 0.05\}$. Breaks lie at 25%, 50%, or 75% of the log-size range. For each sample size, we estimate the null critical value from 50,000 simulations and evaluate size and power on 20,000 further draws. We reuse standard normal draws across effect sizes and locations within a sample size to reduce Monte Carlo variation in comparisons. Appendix A specifies the implementation and random seed.

Table 1: False-positive rates at a nominal 5% level. Each rate uses 20,000 null simulations independent of the 50,000 calibration draws. The Monte Carlo standard error near 0.05 is 0.0015.

n	4	6	8	12	20
Fixed-break F reference	.320	.181	.188	.216	.242
Calibrated scan	.052	.051	.050	.051	.053

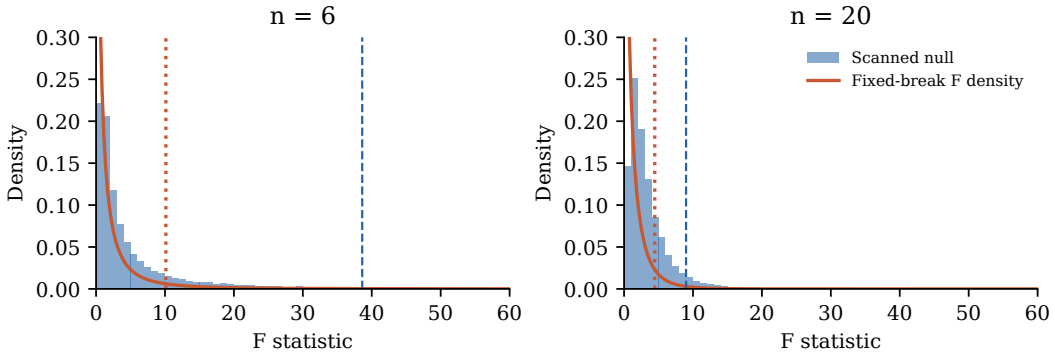

Figure 1: The scanned null distribution and the fixed-break F density. Vertical lines mark the respective 95th percentiles. Scanning shifts probability toward larger statistics; the fixed-break cutoff therefore rejects too often. The plot truncates the right tail at $T = 60$.

4 Calibrating the Breakpoint Scan

The null distribution can be simulated without estimating an exponent or noise scale. This follows from a projection representation of the statistic. Let $X = [\mathbf{1}, x]$ and let $M = I - X(X^\top X)^{-1}X^\top$ project onto the residual space of the single line. For $h_c = ((x_i - c)_+)_{i=1}^n$, define $u_c = Mh_c/\|Mh_c\|$ whenever $Mh_c \neq 0$. Then

$$\mathrm{SSE}_0 = \|My\|^2, \qquad G = \max_{c \in \mathcal{C}} (u_c^\top My)^2, \qquad T = \frac{(n-3)G}{\|My\|^2 - G}. \tag{3}$$

Proposition 1. *For fixed x and $\mathcal{C}$, the null distribution of T under Eq. 1 is independent of β, α, and every $\sigma > 0$. Under a fixed alternative, its distribution depends on the exponent change and noise through δ/σ.*

The projection removes the intercept and common slope. Under the null, $My = \sigma Mz$ for $z \sim \mathcal{N}(0, I)$, and σ^2 cancels in Eq. 3. Under the alternative, $My = -\delta Mh_c + \sigma Mz$, giving the second claim. A complete derivation is in Appendix B.

Procedure. For the observed design and chosen grid:

1. Compute the observed statistic T_{obs} by fitting the null and scanning the segmented alternative.
2. Draw $z^{(b)} \sim \mathcal{N}(0, I_n)$, $b = 1, \ldots, B$, and compute $T^{(b)}$ with the identical design and scan.
3. Reject at level η when T_{obs} exceeds the empirical $(1 - \eta)$ quantile. A Monte Carlo p-value is $[1 + \sum_b \mathbf{1}\{T^{(b)} \geq T_{\mathrm{obs}}\}]/(B + 1)$.

This is equivalent to a parametric bootstrap of the fitted Gaussian linear null: by Proposition 1, its fitted intercept, slope, and positive noise scale do not change the reference distribution. The critical value can therefore be reused for studies with the same relative design and scan grid.

Table 1 shows false-positive rates of 18.1–32.0% when the scan is compared with a fixed-break reference. Calibration gives 5.0–5.3%. At $n = 6$, the calibrated critical value is 38.6, compared with 10.1 for $F(1, 3)$; at $n = 20$, the values are 9.0 and 4.5. Figure 1 displays this difference. The effect is present at every design examined, rather than only at the smallest sample size.

Table 2: Smallest tested exponent change reaching 80% power for a midpoint break. Entries are grid values, not continuously estimated thresholds. All designs span two decades of model size.

$\sigma \setminus n$	4	6	8	12	20
1%	> 0.8	.10	.05	.05	.05
2%	> 0.8	.20	.10	.10	.05
5%	> 0.8	.40	.40	.20	.20

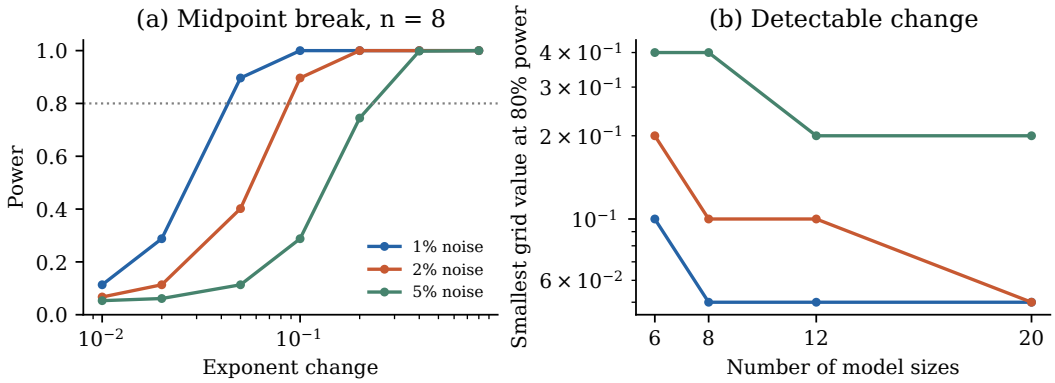

Figure 2: Power and detectable changes for midpoint breaks. (a) Estimated power at eight model sizes. The dotted line marks 80% power. (b) Smallest tested changes reaching that level as the number of sizes increases. Colors denote the same noise levels in both panels.

5 Detection power

We define the detectable change as the smallest tested $|\delta|$ at which estimated power reaches 0.8. The tested grid is $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.8\}$. This is a design-sensitivity summary, not a confidence bound on the exponent change in an observed curve. A non-rejection does not exclude all changes above the reported value.

Sample size and noise. At six to eight model sizes and 2–5% noise, the detectable changes on this grid range from 0.1 to 0.4 (Table 2). For context, Kaplan et al. (2020) report a model-size exponent of 0.076 for a pure power-law fit, while Hoffmann et al. (2022) report 0.34 for the reducible model-size term in a different functional form. These values illustrate the scale of the changes; the simulation does not test the full Chinchilla model.

More observations and lower noise both improve sensitivity. At 2% noise, increasing the number of sizes from six to twenty reduces the smallest detectable grid value from 0.20 to 0.05. At eight sizes, reducing noise from 5% to 2% changes the value from 0.40 to 0.10. The relative cost of these choices depends on the training design.

At four sizes, the scanned fit can nearly interpolate noisy observations. The resulting null has a very long right tail, and no tested change reaches 80% power. This result pertains to the specified grid and statistic; it is not a claim that every test with four observations is uninformative.

Break location. The midpoint is the most sensitive of the three locations evaluated. For $\delta = 0.1$, $n = 8$, and 2% noise, moving the break to the lower quarter reduces power from 0.90 to 0.65. At 5% noise, the corresponding values are 0.29 and 0.16. Table 3 and Figure 3 show the same pattern across sample sizes. The two quarter locations have nearly identical power under the symmetric design.

The location effect has a simple geometric interpretation. For a known breakpoint, the noncentrality parameter of the fixed-break test is

$$\lambda_c = \frac{\delta^2}{\sigma^2} \|M h_c\|^2. \tag{4}$$

Table 3: Power at a fixed exponent change of 0.1. "Quarter" places the break at 25% of the log-size range. Values at 75% differ by less than 0.004 in these simulations.

	$\sigma = 2\%$		$\sigma = 5\%$	
n	Midpoint	Quarter	Midpoint	Quarter
6	.62	.38	.20	.12
8	.90	.65	.29	.16
12	1.00	.89	.44	.24
20	1.00	.99	.69	.37

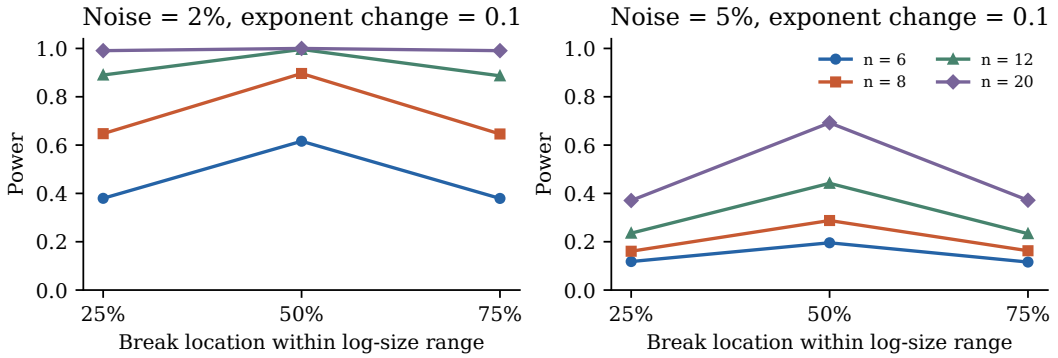

Figure 3: Detection power as the breakpoint moves within the observed range, at exponent change 0.1. Moving toward either end reduces the information available about one segment. Power is estimated from 20,000 simulations per configuration.

Only the part of the hinge that cannot be absorbed by a single line supplies evidence for a break. Near an endpoint, that residual component is small. The scanned test also pays for searching over locations through its larger critical value. A break beyond the measured range cannot be identified by this test because all observed points lie on the same segment.

6 Pythia: curvature from a loss floor

We next examine a measured curve from the Pythia deduplicated suite (Biderman et al., 2023). The six models span 2.12 decades of non-embedding parameter count. Each is evaluated on the WikiText-103 test split (Merity et al., 2017), using 400 consecutive windows of 512 tokens. The supplied per-window losses allow the aggregate curve to be reanalyzed without rerunning model inference. Standard errors use 400 bootstrap resamples of whole windows (Table 4).

Pure power-law fit. Fitting $L = AN^{-\alpha}$ by least squares in $\ln L$ gives $\hat{\alpha} = 0.177$. The scanned statistic is $T = 755.1$. Simulating the null on these actual, unequally spaced log sizes gives a 95th-percentile cutoff of 37.9, compared with the fixed-break value of 10.1. The pure power law is strongly rejected under the working independent Gaussian log-error model. Calibration changes the cutoff but leaves the departure from a single log–log line apparent.

Including a floor. We fit $L = E + AN^{-\alpha}$ using the same log-residual objective. The fit gives $\hat{E} = 2.346$ nats and $\hat{\alpha} = 0.544$, and reduces the residual sum of squares by a factor of 52.8 (Figure 4). Both forms are fitted to the same six means; the floor model has one additional parameter. Its mean absolute residual is 0.026 nats, compared with a mean reported standard error of 0.015 nats.

A floor creates smoothly varying log–log slope without a change in α:

$$\frac{d \ln L}{d \ln N} = -\alpha \frac{AN^{-\alpha}}{E + AN^{-\alpha}}. \tag{5}$$

Table 4: Pythia evaluation losses. Standard errors describe evaluation-window sampling; they do not estimate variation across training seeds.

Model	Non-embedding parameters	NLL (nats)	Bootstrap s.e.
70m	1.89×10^7	6.2344	.0173
160m	8.51×10^7	4.1267	.0162
410m	3.02×10^8	3.1543	.0146
1b	8.06×10^8	2.8565	.0146
1.4b	1.21×10^9	2.7736	.0148
2.8b	2.52×10^9	2.6181	.0154

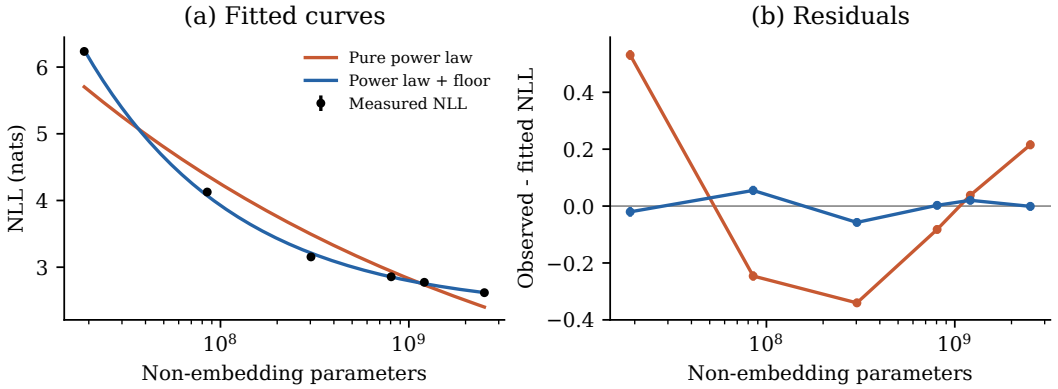

Figure 4: Pythia losses and fits under a common log-residual objective. (a) The floor model follows the flattening of the measured curve. (b) Residuals in loss units, with evaluation bootstrap standard errors. The floor reduces curvature left by the pure power law.

As $AN^{-\alpha}$ decreases relative to E, the magnitude of the slope falls. Thus the rejection of a pure power law does not identify an exponent break: a standard unbroken form already explains much of the curvature. The remaining residuals are not assigned a mechanistic interpretation.

Relation to the power analysis. The mean relative evaluation standard error is 0.46%. Training-run variation is not measured by the released checkpoints. Tables 2 and 3 describe an equally spaced, pure-power-law design, so their thresholds are not confidence bounds for this fitted Pythia curve. Testing a break in the exponent of the reducible term would instead require calibration and power simulations under the floor model, on the actual design. The case study isolates the need to specify that null before interpreting a break test.

7 Reporting Exponent Sensitivity

A scaling study can report sensitivity alongside its fitted exponent in three steps.

Specify the curve and observation model. State whether the loss includes an irreducible term, which variables are held fixed, the loss scale used for fitting, and the assumed noise structure. Distinguish evaluation uncertainty from variation between training runs. A break test assesses a specified null; it cannot select that null on scientific grounds.

Calibrate the complete procedure. Reproduce breakpoint selection inside the null simulation. For the Gaussian linear model in this paper, Proposition 1 simplifies this to standard-normal draws on the observed design. A nonlinear floor model or a correlated noise model requires its own calibration. Any additional search over functional forms must also be included when interpreting the test.

Report a power curve and its design. State the tested changes, break locations, noise levels, and target power. For example: "At eight equally spaced sizes over two decades and 2% relative noise, a midpoint exponent change of 0.1 is detected with approximately 90% power by the calibrated scan." This statement describes a procedure's sensitivity. Excluding changes from observed data requires an appropriate confidence-set or equivalence procedure, rather than a non-significant break test alone.

Scope. Our simulations use one sharp break and independent, homoscedastic Gaussian log errors. Different noise structures, smooth transitions, and alternative scan grids define different power problems. The Pythia analysis illustrates functional-form sensitivity in one model suite. The accompanying code makes the assumed design explicit so that other studies can calibrate and simulate the procedure they intend to use.

8 Conclusion

The sensitivity of an exponent-stability test depends on the search over breakpoints, the sampling design, and the scaling function being tested. Simulating the scan corrects the false-positive rate for the Gaussian piecewise-linear model. Power analysis then quantifies the changes that remain difficult to detect. The Pythia example shows why functional form must be considered separately: a loss floor can explain curvature that a pure power-law test labels as a break. Reporting these choices and the resulting power curve gives a more precise account of what a scaling study establishes.

Reproducibility statement

The supplementary package contains the simulation and fitting script, its numerical outputs, all figure-generation code, and the supplied per-window Pythia losses. Running `code/analyze.py` regenerates the simulation tables, curve fits, and four figures. The original model-evaluation script is also included. Appendix A records the settings, and Appendix B derives the projection formulas.

LLM usage statement

GPT was used for writing polish.

References

Donald W. K. Andrews and Werner Ploberger. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62(6):1383–1414, 1994.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27), 2024.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *International Conference on Learning Representations*, 2023.

Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. Revisiting scaling laws for language models: The role of data quality and training strategies. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Leshem Choshen, Yang Zhang, and Jacob Andreas. A hitchhiker's guide to scaling law estimation. *arXiv preprint arXiv:2410.11840*, 2024.

Robert B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1):33–43, 1987.

Bruce E. Hansen. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

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.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (mis)fitting: A survey of scaling laws. In *International Conference on Learning Representations*, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, 2023.

Vito M. R. Muggeo. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19):3055–3071, 2003.

Tim Pearce and Jinyeop Song. Reconciling Kaplan and Chinchilla scaling laws. *Transactions on Machine Learning Research*, 2024.

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*, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23, 2022.

A Simulation and numerical details

Randomness and sample counts. The analysis uses NumPy's default random generator with seed 20260925. For each n, the script first draws 50,000 standard-normal vectors for calibration and then 20,000 independent vectors for evaluation of test size. It next draws 20,000 vectors reused across noise levels, effect sizes, and break locations for the power comparisons at that n. The Pythia design has a separate set of 50,000 null draws. Quantiles use NumPy's default linear interpolation. The largest binomial Monte Carlo standard error for a power estimate is $\sqrt{0.25/20000} = 0.0036$, conditional on the estimated cutoff.

Fitting and scan grid. The null design matrix contains an intercept and the natural log of relative model size. Each alternative adds one hinge at a grid point. The 60 candidates include the second and penultimate observed log sizes. The implementation first projects out the null design and then evaluates the squared projection onto each normalized residualized hinge. This is algebraically equivalent to refitting each candidate by least squares. It also permits vectorized simulation. The denominator is bounded below by machine epsilon times the null residual sum of squares to avoid division by numerical zero. The power results use the finite grid specified in the main text; changes between grid values are not interpolated.

The four-point design. With four observations, a continuously located hinge and three linear coefficients can nearly interpolate some noise realizations. A dense finite scan can therefore give very small alternative residual sums of squares under the null. In the present grid the simulated 95th percentile of T is approximately 5.7×10^5. Its dependence on the grid is another reason to reproduce the exact scan during calibration.

Pythia fitting. We use the parameter counts and per-window loss values in `data/pythia_curve.json`. The pure power law is fitted by linear least squares in $\ln L$. The floor model is parameterized as $L = E + \exp(b - \alpha x)$ and fitted by nonlinear least squares to the same log losses, subject to $0 \le E < \min_i L_i$ and $0 \le \alpha \le 5$. Both objectives are unweighted. This gives a common scale for comparing residual sums of squares. Figure 4 converts fitted values back to nats for interpretation. The quoted bootstrap standard errors are retained from the supplied evaluation records.

B PROJECTION REPRESENTATION AND POWER

For a fixed breakpoint, orthogonal projection onto the residualized hinge is the additional fitted component beyond the null model. With $r = My$ and $u_c = Mh_c/\|Mh_c\|$, its fitted value is $u_c u_c^\top r$. Orthogonality gives

$$\|r - u_c u_c^\top r\|^2 = \|r\|^2 - (u_c^\top r)^2. \tag{6}$$

Minimizing this expression over the scan grid gives Eq. 3. Under the null, $MX\theta = 0$ and $r = \sigma Mz$. Both the numerator and denominator of the statistic are multiplied by σ^2, proving the null invariance.

Under a break at $c_\star$, $r/\sigma = -(\delta/\sigma)Mh_{c_\star} + Mz$. Thus the distribution depends on the effect and noise through their ratio. Gaussian symmetry also gives identical power for δ and $-\delta$. If the tested breakpoint is fixed at the true location, the added one-dimensional component has noncentrality parameter $\lambda = (\delta/\sigma)^2\|Mh_{c_\star}\|^2$, yielding Eq. 4. Scanning changes the reference distribution, but this quantity still describes how much of the signal remains after the common line has been removed.

Logarithm and span conventions. Using $x = \ln(N/N_1)$ and $y = \ln L$ makes α and δ power-law exponents directly. If base-10 logs are used for both coordinates, the exponents are unchanged but the noise standard deviation becomes $\sigma/\ln 10$. Using base-10 model size with natural-log loss instead multiplies both slope coefficients by $\ln 10$. The supplied analysis uses natural logarithms for both coordinates. Multiplying the log-size span by $a > 0$, while holding the relative design and break location fixed, multiplies Mh_c by a; the alternative therefore depends on $|\delta| \times \text{span}/\sigma$.