TIED, NOT WON: TESTING CLAIMS THAT ONE MODEL SCALES BETTER THAN ANOTHER

Anonymous authors
Paper under double-blind review

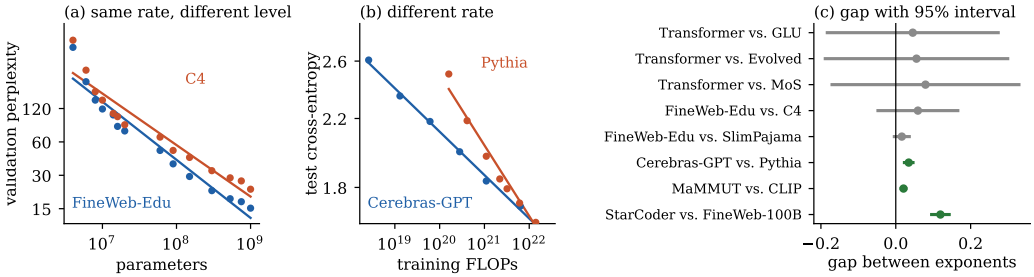

Figure 1: **Testing whether one model scales better than another.** (a) Two pretraining corpora of DataDecide fall at the same rate but at different levels. (b) Cerebras-GPT and Pythia fall at different rates. (c) Eight of the 29 audited comparisons: the gap between the two fitted exponents with its 95% interval; grey intervals cover zero (ties), green intervals do not (separations).

ABSTRACT

Many papers conclude that one architecture, training objective, or pretraining corpus "scales better" than another. Because a scaling rate is the exponent of a fitted power law, such a conclusion is a claim about the difference between two exponents. In practice it is drawn from a handful of model sizes, and neither the exponents nor their difference come with a standard error. We supply the missing statistics. Using only the per-scale results that the original papers released, we refit each power law, check that the refit reproduces the published exponent, and test whether the gap between two exponents is distinguishable from zero. Across widely cited comparisons from eight released model suites, most claimed differences in scaling rate are statistical ties. The conclusion that the vanilla Transformer scales best among ten architectures does not hold against its close contenders, and nearly all pairs of pretraining corpora in the DataDecide suite differ in loss level but not in rate. The probability that the claimed winner would reverse, computed from the same test, separates these ties from large, well-established differences. A closed-form bound on the smallest detectable gap explains the pattern: the small sets of model sizes in common use cannot resolve gaps of the size being claimed. A survey of recent scaling papers further shows that most do not release the data needed to check their own comparisons. We close with simple reporting practices that make comparative scaling claims testable.

1 INTRODUCTION

Scaling laws describe how the loss of a model falls as a power of its size, data, or compute (Hestness et al., 2017; Rosenfeld et al., 2020; Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022). Because this relation holds over many orders of magnitude, a few small models can predict the behavior of much larger ones (Ivgi et al., 2022; Ruan et al., 2024), and scaling laws have become a standard tool for planning training runs, from the size of a compute-optimal model (Hoffmann et al., 2022) to the trade-offs with inference cost (Sardana et al., 2024) and limited data (Muennighoff et al., 2023). They are now also used to choose between methods. New architectures are

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.

compared with the Transformer through their scaling curves (Tay et al., 2023; Gu & Dao, 2024; Sun et al., 2023; De et al., 2024), sparse mixture-of-experts models with dense ones (Clark et al., 2022; Ludziejewski et al., 2024), pretraining corpora with one another (Magnusson et al., 2025; Brandfonbrener et al., 2025), and contrastive with generative vision–language objectives (Nezhurina et al., 2025).

Many of these comparisons make a stronger claim than that method A beats method B at the sizes tested: they conclude that A scales better, so that its advantage grows with compute. The two claims differ in kind. An advantage at a fixed size is a local fact, which anyone can check by evaluating both models. A better scaling rate is an extrapolation beyond the largest model trained, and it shapes where the field invests: which architecture is trained at the next scale, which corpus is curated, and which objective is adopted. Because the next scale is expensive, such a claim is usually acted upon long before it can be checked directly.

A scaling rate is the exponent of a fitted power law, so "A scales better than B" states that one fitted exponent exceeds another. The exponent is not observed but estimated: it is the slope of a line fitted to a few points, each a separately trained model, and it inherits the noise of those points. Checking the claim therefore requires the uncertainty of both exponents and a test on their difference. Such claims are rarely checked this way. They often rest on a handful of model sizes (four or five per architecture in the study of Tay et al. (2023)), the exponents are reported without standard errors, and no test is run on the gap.

Judging such a gap by eye is difficult for two reasons. First, the difference of two exponents is noisier than either exponent, because the uncertainties of two independent fits add. Second, the uncertainty of a slope depends on how many points it is fitted to and how wide a range of sizes they span, and the few points of a typical study give slopes whose uncertainty is as large as the differences being claimed. Two curves can therefore look clearly different on a log–log plot while the data cannot say which one falls faster. Retraining the models to check a claim costs as much as the original study, so the reader cannot tell whether the claimed ordering exceeds the noise of the fits.

In this paper we supply the missing test for published claims, without training any model. For each claim we take the per-scale results released by its authors, refit the power laws in the authors' own coordinates, confirm that the refit reproduces the published exponent, and test whether the difference between the two exponents is distinguishable from zero. When it is not, we call the comparison a tie: the data do not show which method scales faster. When it is, we call it a separation. We express each result as a reversal probability, the probability that the claimed winner would reverse, which grades comparisons from clearly separated to tied. We also derive the detection floor, the smallest gap that a set of model sizes can detect, in closed form, so that it can be computed before any model is trained, and we show where the sizes should be placed.

Applied to comparisons from eight released sources, the test finds mostly ties. The best-known example is the conclusion of Tay et al. (2023) that the vanilla Transformer scales best among ten architectures: against each of its close contenders the difference is within the noise, even on the exponents the authors themselves reported. The pretraining corpora of DataDecide (Magnusson et al., 2025) likewise differ in how low their loss is but not in how fast it falls. Comparisons with large, well-established differences, in contrast, come back as separations, and the reversal probability ranks every tie above every such difference. The detection floor explains the ties: the sets of model sizes in common use are too small to resolve gaps of the size that papers claim. A survey of recent scaling papers shows that most of them do not release the per-scale numbers needed to run the test at all. Our contributions are as follows:

- We formulate a claim that one method scales better than another as a test on the difference of two fitted exponents and apply it to published claims on the authors' own data, finding that most claimed differences in scaling rate, including the Transformer result of Tay et al. (2023) and the corpus comparisons of DataDecide, are statistical ties (Sections 3.1 and 4).
- We introduce the reversal probability, which grades each comparison from clearly separated to tied, and the detection floor, the smallest gap that a set of model sizes can detect, together with where those sizes should be placed (Sections 3.2, 3.3, 4.4, and 4.5).
- We measure how often comparative scaling claims can be audited at all and propose reporting practices that make them testable (Sections 4.6 and 5).

2 Related Work

Scaling laws and their estimation. Power-law decay of loss with model size, data, and compute was established by Hestness et al. (2017) and Rosenfeld et al. (2020), made central for language models by Kaplan et al. (2020) and Hoffmann et al. (2022), and extended to image, video, and multimodal data by Henighan et al. (2020), with Bahri et al. (2024) giving a theoretical account of the exponents. Other work questions whether a single power law is the right form (Alabdulmohsin et al., 2022; Caballero et al., 2023) and how well laws fitted at small scale predict larger models and downstream tasks (Ivgi et al., 2022; Lourie et al., 2025; Schaeffer et al., 2025b); observational scaling laws fit existing public models without training new ones (Ruan et al., 2024), as our audit does. A line of work studies the fitting procedure itself: estimating exponents in log space is statistically delicate (Clauset et al., 2009), conclusions depend on fitting choices and unreported details (Li et al., 2025), and the fitted law shifts with the number of models, seeds, and recipe (Choshen et al., 2025). The Chinchilla compute-optimal exponent in particular has been re-examined for implausibly narrow intervals (Besiroglu et al., 2024), reconciled with Kaplan et al. by correcting parameter counts and small-scale effects (Porian et al., 2024; Pearce & Song, 2024), stress-tested under perturbations (Schaeffer et al., 2025a), and revisited under inference cost (Sardana et al., 2024) and limited data (Muennighoff et al., 2023). These studies ask whether one law is estimated reliably. We ask whether the difference between two laws is resolved, which is the quantity a comparative claim depends on.

Comparative scaling claims. A growing literature states comparisons between methods as differences in scaling rate. For architectures, recurrent and state-space alternatives to attention are compared with Transformers through their scaling curves (Tay et al., 2023; Gu & Dao, 2024; Sun et al., 2023; Peng et al., 2023; Beck et al., 2024; De et al., 2024), as are linear-attention models (Shen et al., 2024) and hybrid stacks (Poli et al., 2024); sparse mixture-of-experts models are compared with dense ones (Fedus et al., 2022; Clark et al., 2022; Ludziejewski et al., 2024). For data, studies compare the scaling of pretraining corpora and curation methods (Magnusson et al., 2025; Brandfonbrener et al., 2025; Mayilvahanan et al., 2025), including claims that data pruning changes the exponent itself (Sorscher et al., 2022). Similar claims compare contrastive and generative vision–language objectives (Nezhurina et al., 2025), the training sets of CLIP models (Cherti et al., 2023), ConvNets and vision Transformers (Smith et al., 2023), and diffusion and autoregressive language models (Prabhudesai et al., 2025). These works report exponents or curves, but rarely an uncertainty on the difference between two exponents. We take their released numbers and test that difference directly, on the authors' own coordinates.

Statistical practice in machine learning evaluation. Many fields of machine learning have found that conclusions drawn from few runs and point estimates do not survive a proper statistical treatment. In reinforcement learning, reported gains shrink or reverse once run-to-run variance is accounted for (Henderson et al., 2018; Agarwal et al., 2021). Benchmark comparisons in general carry variance from data sampling, initialization, and hyperparameters (Bouthillier et al., 2021), and in NLP, significance testing and power analysis show that many reported improvements are below what the evaluation can detect (Dror et al., 2018; Card et al., 2020). Controlled re-evaluations have found that claimed advantages of language-model architectures and of GAN variants largely vanish under a fair comparison (Melis et al., 2018; Lucic et al., 2018), and recent work quantifies the noise of language-model evaluations, including at the small scales used to make design decisions (Madaan et al., 2024; Heineman et al., 2025). We extend this line from comparisons at a single scale to comparisons of scaling rates, where the evidence is a fitted exponent and the experimental design is a set of model sizes.

Re-analysis and meta-science. Our work re-analyzes published claims without retraining, like Schaeffer et al. (2023), who show that several reported emergent abilities come from discontinuous metrics rather than from scale. Large replication efforts (Open Science Collaboration, 2015) and analyses of statistical power (Button et al., 2013; Ioannidis, 2005) show how underpowered designs and flexible, data-dependent analysis (Simmons et al., 2011; Gelman & Loken, 2014) produce unreliable findings. The probability that an estimated effect has the wrong sign (Gelman & Carlin, 2014) and the probability that a replication keeps the same sign (Killeen, 2005) are the closest precedents to our reversal probability. Reproducibility checklists and audits (Gundersen & Kjensmo, 2018; Pineau et al., 2021; Kapoor & Narayanan, 2023) call for machine-readable artifacts, which our sur-

vey supports with a measurement for scaling papers. Concurrently, Zhuang et al. (2026) adapt a classical paired sample-size bound into a minimum-detectable-effect budget for 4-bit quantization benchmarks, and find their own observed gaps below the implied floor.

3 METHOD

A claim that one method scales better than another concerns the rates at which their losses fall. We test such a claim on the difference of the two fitted exponents (Section 3.1), express the result as the probability that the claimed winner reverses (Section 3.2), and compute the smallest difference that a given set of model sizes can detect (Section 3.3).

3.1 THE SLOPE-DIFFERENCE TEST

Two methods can differ in level, how low the loss is at a given size, or in rate, how fast the loss falls with size (Figure 1a,b). A claim that one method scales better is a claim about rate. For each method we refit a power law as a straight line $y' = a + \alpha x'$ in the paper's own transformed coordinates, for example negative log-perplexity against $\log_{10}$ FLOPs for Tay et al. (2023); where a paper uses the Chinchilla form $L(N, D) = AN^{-\alpha} + BD^{-\beta} + E$, as Brandfonbrener et al. (2025) do, we fit that form by nonlinear least squares. Each fit gives an exponent $\hat{\alpha}$ with an analytic standard error, and before any test we check that the refit reproduces the published exponent. The claim "A scales better than B" says that $\alpha_A - \alpha_B$ has a given sign. We test it with Welch's test for two independent estimates (Welch, 1947; Paternoster et al., 1998), using the Welch–Satterthwaite degrees of freedom (Satterthwaite, 1946):

$$d = \alpha_A - \alpha_B, \qquad \mathrm{SE}(d) = \sqrt{\mathrm{SE}_A^2 + \mathrm{SE}_B^2}, \qquad t = d/\mathrm{SE}(d). \tag{1}$$

A comparison is a separation ("sep" in the tables) if the 95% confidence interval of d excludes zero in the claimed direction, and a tie otherwise; we report two-sided p-values. Signs follow each paper's coordinates: for Tay et al. and DataDecide a larger exponent is better, while for slopes of error against compute a more negative slope is better. Fitting power laws in log space is delicate (Clauset et al., 2009), so the standard errors come from each fit's residual scatter; a residual bootstrap (Efron & Tibshirani, 1993) serves as a cross-check but is weak with four to six points, so the analytic test is primary.

3.2 THE REVERSAL PROBABILITY

The test gives each comparison a verdict. To show how close the verdict is, we express the same test as the probability that the claimed winner reverses. Each fitted slope has a sampling distribution $\alpha_A \sim t_{\nu_A}(\hat{\alpha}_A, \mathrm{SE}_A)$, a t distribution whose degrees of freedom account for the few points of the fit, and likewise for α_B. For a claimed direction $s \in \{+1, -1\}$, the reversal probability is

$$P_{\mathrm{rev}} \;=\; \Pr\big[\,\mathrm{sign}(\alpha_A - \alpha_B) \neq s\,\big], \tag{2}$$

which we estimate by drawing each slope from its own distribution. A value near 0 means the claimed winner essentially never changes, a value near 0.5 means the winner is a coin flip, and a value above 0.5 means the data favor the opposite direction. The reversal probability is close to a one-sided p-value, but it keeps the direction of the claim, so a tie that leans toward the claimed winner (below 0.5) is told apart from one that leans against it (above 0.5). We use it because it reads directly as the chance that the claimed ordering is wrong, and it should give large, established differences scores near zero rather than flagging every comparison as uncertain.

3.3 THE DETECTION FLOOR

With few model sizes a fitted slope is uncertain, and the uncertainty can be computed before any model is trained. We call the set of model sizes and runs an experimental design. Each method follows a power law that is a straight line in log–log coordinates, $y = a - \alpha x$ with $x = \log N$, written with $-\alpha$ so that a larger α means faster decay. It is measured at J model sizes spaced evenly in $\log N$ over R decades, with r runs per size and per-point noise of standard deviation σ in y. The

Table 1: **The 29 audited comparisons by source.** For Tay et al. the five close contenders of the Transformer; for DataDecide eight representative pairs. Every comparison is listed in Appendix A.

Source	what is compared	pairs	tie	sep	amb
Tay et al.	architectures	5	5	–	–
DataDecide	pretraining corpora	8	7	–	1
Brandfonbrener et al.	pretraining data (code, text)	4	1	3	–
LAION suite	vision–language models, objectives	4	2	2	–
Pythia	against Mamba; deduplicated data	2	2	–	–
Cerebras-GPT	training regime (against Pythia)	1	–	1	–
OLMo ladder	over-training ratio	2	2	–	–
Over-training suite	pretraining corpora	3	3	–	–
Total		29	22	6	1

least-squares slope has variance $\mathrm{Var}(\hat{\alpha}) = \sigma^2/(r\,S_{xx})$ with $S_{xx} = \sum_i (x_i - \bar{x})^2$, and for J evenly spaced points over a window of width $W = R \ln 10$,

$$S_{xx} = \frac{W^2 J(J+1)}{12(J-1)} \xrightarrow[\text{large } J]{} \frac{W^2 J}{12}. \tag{3}$$

For two independent fits under the same design the variances of the two slopes add, so for large J the gap has standard deviation $\mathrm{SD}(\alpha_A - \alpha_B) \approx \sqrt{24}\,\sigma/(\sqrt{rJ}\,R \ln 10)$: the factor 12 comes from the spread of evenly spaced points, and it is doubled because two slopes are compared. Detecting a true gap $\delta = |\alpha_A - \alpha_B|$ therefore requires, for a two-sided test at level γ and power $1-\eta$ (Cohen, 1988),

$$R\sqrt{rJ} \;\geq\; \frac{(z_{\gamma/2} + z_{\eta})\sqrt{24}\,\sigma}{\delta \ln 10}. \tag{4}$$

We call the smallest gap detected with 80% power at the 5% level the detection floor. With R in decades and one run per size, Equation 4 gives the floor and, inversely, the number of sizes needed to detect a gap δ:

$$\mathrm{MDE}_{80} \approx \frac{5.96\,\sigma}{R\sqrt{J}}, \qquad J_{\min} \approx \left(\frac{5.96\,\sigma}{\delta\,R}\right)^2. \tag{5}$$

The floor thus falls in proportion to the range of sizes but only with the square root of their number. For small J we use the exact S_{xx} and the noncentral t distribution (Appendix B). If the two methods share the same sizes and their per-point noise has correlation ρ, the standard deviation shrinks by a factor $\sqrt{1-\rho}$, so such a paired design is always more powerful.

The floor assumes sizes evenly spaced in $\log N$. Where to place a fixed number of sizes is a classical problem of optimal design (Elfving, 1952; Pukelsheim, 2006; Atkinson et al., 2007). A single slope is estimated most precisely when all sizes sit at the two ends of the range (the c-optimal design, $S_{xx} = JW^2/4$), and even log-spacing, with $S_{xx} \to JW^2/12$, inflates the standard deviation of the slope, and hence of the gap, by a factor that approaches $\sqrt{3}$. An endpoint design, however, cannot check the power-law form on which the meaning of the exponent rests. Placing about a quarter of the sizes in the interior keeps most of the endpoint precision while retaining the power to reject a wrong functional form (Appendix B).

As an example, take the vanilla Transformer and the GLU-Transformer of Tay et al. (2023), each fitted to five model sizes that span 1.68 decades of compute. The gap between their refit exponents, 0.0454, is less than half of its standard error, 0.0947 (Equation 1), so the 95% interval covers zero and the comparison is a tie. The reversal probability is 0.350: in about one draw in three, the GLU-Transformer rather than the Transformer has the larger exponent. The floor of this design (Equation 5) is 0.145, three times the observed gap, so five sizes over this range would usually miss even a real gap of this size.

4 Experiments

We apply the method of Section 3 to published comparisons. Sections 4.2 and 4.3 examine the two most prominent claims in detail, Sections 4.4 and 4.5 evaluate the reversal probability and the detection floor across all audited comparisons, and Section 4.6 measures how many claims can be audited at all.

Table 2: **Tay et al.'s architectures, tested on their own exponents.** The exponents and gaps are Tay et al.'s; the standard errors of the gaps are ours, since none were reported. Welch test of the vanilla Transformer against each contender; t uses unrounded standard errors.

Contender	α (Transf.)	α (contender)	gap	SE of gap	t	p	verdict
GLU-Transformer	0.54	0.49	+0.050	0.095	+0.53	0.617	tie
Universal	0.54	0.50	+0.040	0.118	+0.34	0.750	tie
Evolved	0.54	0.44	+0.100	0.101	+0.99	0.362	tie
MoS-Transformer	0.54	0.43	+0.110	0.104	+1.06	0.330	tie
Funnel	0.54	0.47	+0.070	0.118	+0.60	0.575	tie
LConv	0.54	0.32	+0.220	0.099	+2.23	0.068	tie
Performer	0.54	0.25	+0.290	0.140	+2.08	0.120	tie
Switch	0.54	0.23	+0.310	0.117	+2.64	0.041	sep
MLP-Mixer	0.54	0.32	+0.220	0.083	+2.64	0.049	sep

$7/9$ of the reported gaps are ties; p ranges from 0.041 to 0.750.

4.1 Setup

We audit comparisons whose per-scale results were released, drawn from eight sources that compare architectures, pretraining corpora, training objectives, and training regimes (Table 1): the architecture study of Tay et al. (2023), the DataDecide suite (Magnusson et al., 2025), the data sweep of Brandfonbrener et al. (2025), the LAION comparison suite (Nezhurina et al., 2025), the Pythia (Biderman et al., 2023) and Cerebras-GPT (Dey et al., 2023) model families, the OLMo scaling ladder (Bhagia et al., 2025), and the over-training suite of Gadre et al. (2025). Every number is a fit to the released points. For each comparison we report the gap between the two exponents, its Welch p-value, the reversal probability, and, where the design is known, the detection floor. Six comparisons serve as positive controls: differences in scaling rate that are established by many standard errors. Three are code-versus-text contrasts from the data sweep of Brandfonbrener et al. (2025), for example StarCoder against FineWeb-100B. Two come from the LAION suite (Nezhurina et al., 2025): MaMMUT against CLIP, and a captioning objective (Cap) against CLIP. The last compares Cerebras-GPT, trained compute-optimally, with Pythia, trained on a fixed token budget. For every comparison we extract the per-scale points, refit both laws in the authors' own coordinates, check the refit against the exponents the authors published (Appendix A), and only then run the test.

4.2 The vanilla Transformer ties with its close contenders

Tay et al. (2023) compare ten architectures and conclude that "amongst all ten architectures that we consider, the vanilla Transformer has the best scaling behaviour, even if its absolute performance at each compute region is not the greatest," adding that it "has the highest values of α." The claim rests on the coefficient $\alpha_{F,U}$, the slope of negative log-perplexity on $\log_{10}$ FLOPs, reported as 0.54 for the vanilla Transformer, the largest in their table of exponents. From the raw points they tabulate we recover 0.543, and the refit preserves their ranking of the ten architectures. With our refit exponents, no confidence interval excludes zero in the claimed direction, and Funnel and Universal even refit steeper than the Transformer (Figure 2a; Table 3 in Appendix A).

The ties also hold in the version of the test most favorable to the original claim, which takes Tay et al.'s reported exponents as the point estimates and attaches only our standard errors, since they report none (Table 2). All five close contenders (GLU, Universal, Evolved, MoS, and Funnel) tie. Only Switch and MLP-Mixer reach $p < 0.05$, and only barely; both are among the four architectures that Tay et al. rank weakest. Within the uncertainty of the authors' own numbers, the vanilla Transformer cannot be distinguished from its close contenders.

4.3 DataDecide corpora differ in level, not in rate

DataDecide (Magnusson et al., 2025) releases validation perplexity for 25 pretraining corpora, each trained at the same 14 model sizes, which makes it an unusually clean test bed. Of all 300 corpus pairs, 299 are ties on the scaling exponent; the remaining pair, the steepest against the shallowest corpus, separates marginally on this metric but ties on held-out C4 validation data. The corpora, which include DCLM, FineWeb-Edu, C4, Dolma, Falcon, and code-ablated variants, reorder by level rather than by rate (Figures 1a and 2b). At the largest model size their perplexities differ almost threefold, whereas the whole spread of their exponents is smaller than the 95% interval of a

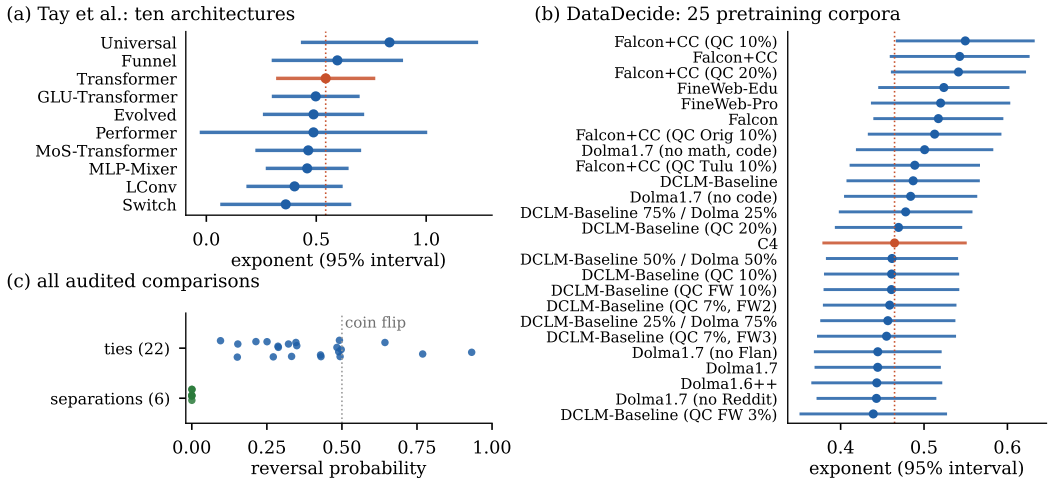

Figure 2: **Fitted exponents in the two case studies, and the reversal probability of every comparison.** (a) The ten architectures of Tay et al., refit from their tabulated points; the Transformer (orange) lies inside the interval of every contender. (b) The 25 pretraining corpora of DataDecide; every interval overlaps that of C4 (orange). (c) In reversal probability, every tie scores above every separation of the audit.

single exponent. Even with 14 sizes, every gap lies below the design's detection floor, so the rate differences are unresolvable rather than shown to be zero (Altman & Bland, 1995). This matches how the suite is meant to be used: DataDecide asks which corpus will perform better at a larger size and how well small experiments predict that ranking, a question about level that its many sizes answer well.

4.4 The reversal probability separates ties from established differences

Across all 29 comparisons the test finds 22 ties, 6 separations, and the ambiguous DataDecide pair (Table 1). Figure 2c shows the reversal probability of each of them. All six positive controls lie at essentially zero, while the ties spread over the range where the winner is uncertain, so the measure ranks every tie above every separation, across comparisons of data, models, objectives, and training regimes. Most ties are far from settled, and for three of them the data favor the opposite of the claimed direction. Because the reversal probability re-expresses the evidence that assigns the labels, this is a consistency check rather than an independent validation. The test behaves the same way on a well-known difference outside the 29 comparisons. Refitting the released analysis of Pearce & Song (2024), the compute-optimal exponent is 0.508 when parameters are counted in total and 0.686 when embedding parameters are excluded, the basis that produced the higher exponent of Kaplan et al. (2020); the two come back as a clear separation.

4.5 Claimed gaps lie below the detection floor

The closed-form floor matches a simulation of the full pipeline for four or more sizes, and the test keeps its nominal false-positive rate (Figure 3a; Appendix B). With only three sizes the closed form over-predicts power, so the smallest designs are even weaker than it says. Figure 3c compares each claimed gap with its design's floor (numbers in Table 7, Appendix C). In our refit, the gaps between the Tay Transformer and its close contenders lie below their floors, whereas the code-versus-text gaps of Brandfonbrener et al. (2025) lie far above theirs. The one Tay gap above its floor, against Universal, points the other way: in our refit Universal has the larger exponent. By Equation 5, at the noise level of the Tay fits, resolving a gap of 0.05 needs about 15 sizes over 3 decades, while the five sizes over 1.7 decades of the original study give a floor three times too coarse. Because the required number of sizes grows with the square of the noise and falls with the square of the range (Equation 5), widening the range is the cheapest way to lower the floor: doubling it cuts the number of sizes needed by a factor of four. The closed form is accurate enough to plan a design: solving

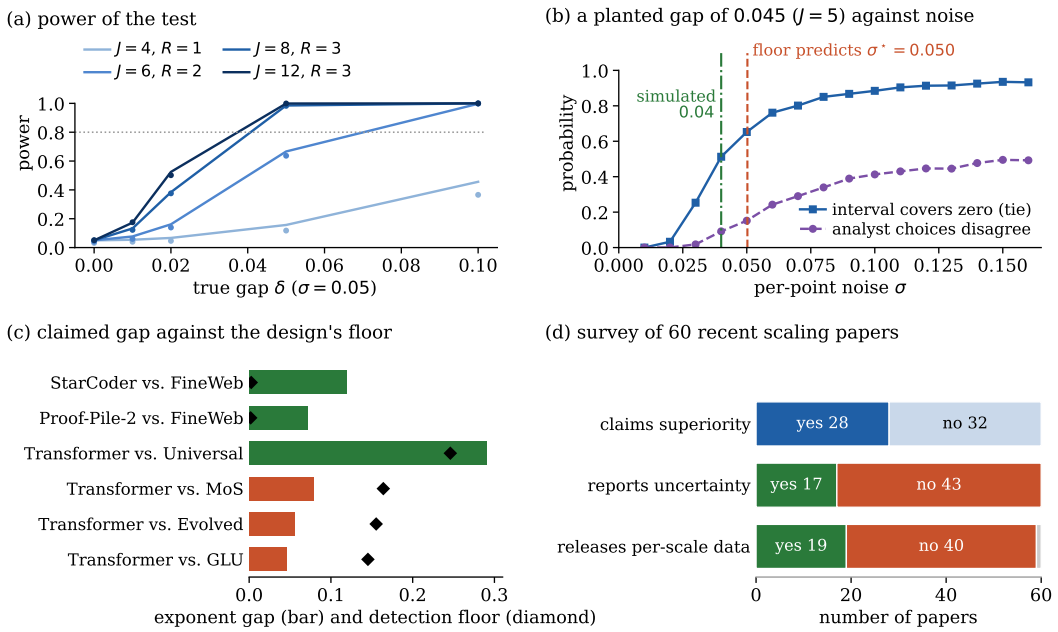

Figure 3: **Few-point designs cannot resolve the claimed gaps.** (a) Simulated power of the test matches the closed-form floor. (b) With a planted gap, the verdict turns into a tie, and irrelevant analysis choices flip the winner, at the noise the floor predicts. (c) Audited gaps below (red) and above (green) their design's floor. (d) Survey of 60 recent scaling papers; one could not be coded for data release.

the power equation exactly changes the required number of sizes by at most one in every setting we checked (Appendix B).

Below the floor, the winner depends on analysis choices. On the Tay data the standard deviation of the estimated gap is as large as the gaps themselves, and dropping the smallest or largest point, subsampling, or switching to a robust (Huber) or a three-parameter fit flips the nominal winner in either direction. To see where this begins, we simulate the design of the Tay comparisons, plant a gap the size of the Transformer–GLU gap, and raise the per-point noise step by step (Figure 3b). The winner turns into a tie close to the noise level the floor predicts. The winner of an underpowered comparison is then set by the analyst's choices, a problem known as researcher degrees of freedom (Simmons et al., 2011; Gelman & Loken, 2014).

4.6 Many claims cannot be audited

Of five further high-profile claims, three cannot be checked because the scaling comparison appears only in a figure without released points (Table 6, Appendix C): RetNet against the Transformer (Sun et al., 2023), ConvNets matching ViTs (Smith et al., 2023), and 4-bit precision being almost universally optimal for a fixed budget of model bits (Dettmers & Zettlemoyer, 2023). The ConvNet comparison also combines results from different papers, datasets, and hardware with an estimated compute conversion. Where part of a claim can be tested, the pattern recurs. On the only numeric scaling data for Mamba (Gu & Dao, 2024), compared with Pythia at the same tokenizer, the rates tie once Pythia's undertrained smallest model is removed, and only an advantage in level survives. For the headline crossover of diffusion against autoregressive models (Prabhudesai et al., 2025), in the data-constrained regime of Muennighoff et al. (2023), the paper reports the fitted crossover law, but the exponents of the two model families and the per-run losses appear only as curves in figures. Missing tables blocked other audits as well. OLMo-2 (Team OLMo et al., 2025) releases checkpoint files that list only the step and a download location, with per-step training loss on WandB but no table indexed by model size; the Mamba repository ships weights and evaluation scripts but no numeric result tables; and one family in the ColPret collection (Choshen et al., 2025) is digitized from a scatter plot rather than measured on distinct trained models (Appendix A lists five such cases).

The same gap appears across the literature. We surveyed 60 recent scaling papers on architectures, objectives, optimizers, data curation, and test-time compute (Figure 3d). Under a criterion fixed before coding, a paper is auditable only if it releases downloadable, machine-readable per-scale numbers for its headline scaling comparison. Two LLM agents coded the papers independently (Cohen's $\kappa = 0.725$; Cohen, 1960; see the AI use statement), and every disagreement was resolved by re-checking the paper's released material. About two thirds of the papers are not auditable, and fewer than a third report any uncertainty on the fitted law. The papers that make the strongest claims are the least checkable: among those that claim superiority, only about one in five reports an uncertainty and one in four releases the numbers behind the claim.

5 Conclusion

We tested published claims that one model scales better than another on the authors' own released data. Most claimed differences in scaling rate are statistical ties, including the conclusion that the vanilla Transformer scales best and the comparisons between pretraining corpora in DataDecide, while large, established differences come back as separations. The detection floor explains the ties: the small sets of model sizes in common use cannot resolve the gaps being claimed. A tie does not show that two methods scale alike; it shows that the released data cannot tell them apart, and a larger or paired design may still resolve a real difference.

Three reporting practices would make such claims testable. First, the per-scale points behind every fitted law should be released as a machine-readable file, so that any reader can refit them. Second, each fitted exponent should come with a standard error, and each comparison with a test on the difference of exponents and its reversal probability. Third, the detection floor of the design should be reported. Because the floor depends only on the number of sizes, the range they span, and the noise of the fits, it can be computed before any model is trained, and a study can choose enough sizes (about 15 over 3 decades at the noise level of the Tay fits) or evaluate both methods at the same sizes, which removes the noise the two fits share. Our analysis adopts each paper's functional form, and extending it to further claims as their per-scale data are released is a natural next step.

Reproducibility statement

All results are fits to per-scale numbers that the audited papers released; no model was trained. The test, the reversal probability, and the detection floor are defined in Section 3. The sources, functional forms, and fitted exponents of every comparison are listed in Appendix A, and the optimal-design results in Appendix B. We will release the analysis code together with the survey codes.

AI use statement

We used large language model assistants for editing text, for coding assistance, for literature search, and for help in developing and checking the derivations. The two coders of the survey in Section 4.6 are two independent model instances. Every result comes from code the authors checked, every citation was verified against its source, and the authors take full responsibility for the paper.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29304–29320, 2021. arXiv:2108.13264.

Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2209.06640.

Douglas G. Altman and J. Martin Bland. Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003):485, 1995.

A. C. Atkinson, A. N. Donev, and R. D. Tobias. *Optimum Experimental Designs, with SAS*. Oxford University Press, Oxford, 2007.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences (PNAS)*, 121(27):e2311878121, 2024. arXiv:2102.06701.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv:2405.04517.

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

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 *Conference on Language Modeling (COLM)*, 2025. arXiv:2412.04403.

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 (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. arXiv:2304.01373.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 3, pp. 747–769, 2021. arXiv:2103.03098.

David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham M. Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *Transactions on Machine Learning Research (TMLR)*, 2025. URL https://openreview.net/forum?id=1Avb4jYjLb. arXiv:2411.12925.

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.14891.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023. arXiv:2212.07143.

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 (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10683–10699, 2025. arXiv:2410.11840.

Aidan Clark, Diego de las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George van den Driessche, Eliza Rutherford, Tom Hennigan, Matthew J. Johnson, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Marc'Aurelio Ranzato, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4057–4086, 2022. arXiv:2202.01169.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7750–7774, 2023. arXiv:2212.09720.

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.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1383–1392, 2018.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research (JMLR)*, 23 (120):1–39, 2022. arXiv:2101.03961.

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, Alex 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 *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.08540.

Andrew Gelman and John Carlin. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 2014.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2312.00752.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi:10.1609/aaai.v32i1.11503.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2508.13144.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. arXiv:1709.06560.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

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.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 30016–30030, 2022. arXiv:2203.15556 (Chinchilla).

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Maor Ivgi, Yair Carmon, and Jonathan Berant. Scaling laws under the microscope: Predicting Transformer performance from small scale experiments. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 7354–7371, 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.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Peter R. Killeen. An alternative to null-hypothesis significance tests. *Psychological Science*, 16(5): 345–353, 2005.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (mis)fitting: A survey of scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.18969.

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16167–16180, 2025.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. arXiv:1711.10337.

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 *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33270–33288, 2024. arXiv:2402.07871.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

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 (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42487–42502, 2025. arXiv:2504.11393.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Sayak Mallick, Matthias Bethge, and Wieland Brendel. LLMs on the line: Data determines loss-to-loss scaling laws. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43302–43330, 2025. arXiv:2502.12120.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1707.05589.

William Merrill, Yanhong Li, Tyler Romero, Anej Svete, Caia Costello, Pradeep Dasigi, Dirk Groeneveld, David Heineman, Bailey Kuehl, Nathan Lambert, Chuan Li, Kyle Lo, Saumya Malik, DJ Matusz, Benjamin Minixhofer, Jacob Morrison, Luca Soldaini, Finbarr Timbers, Pete Walsh, Noah A. Smith, Hannaneh Hajishirzi, and Ashish Sabharwal. Olmo Hybrid: From theory to practice and back. *arXiv preprint arXiv:2604.03444*, 2026.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 50358–50376, 2023. arXiv:2305.16264.

Marianna Nezhurina, Tomer Porian, Giovanni Puccetti, Tommie Kerssies, Romain Beaumont, Mehdi Cherti, and Jenia Jitsev. Scaling laws for robust comparison of open foundation language-vision models and datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.04598.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015.

Raymond Paternoster, Robert Brame, Paul Mazerolle, and Alex Piquero. Using the correct statistical test for the equality of regression coefficients. *Criminology*, 36(4):859–866, 1998.

Tim Pearce and Jinyeop Song. Reconciling Kaplan and Chinchilla scaling laws. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2406.12907.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the Transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, 2023.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research (JMLR)*, 22(164):1–20, 2021. arXiv:2003.12206.

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 (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40908–40950, 2024. arXiv:2403.17844.

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 (NeurIPS)*, volume 37, 2024. arXiv:2406.19146.

Mihir Prabhudesai, Mengning Wu, Amir Zadeh, Katerina Fragkiadaki, and Deepak Pathak. Diffusion beats autoregressive in data-constrained settings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2507.15857.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), 2006.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1909.12673.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 15841–15892, 2024. arXiv:2405.10938.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2401.00448.

F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.15004.

Rylan Schaeffer, Noam Levi, Andreas Kirsch, Theo Guenais, Brando Miranda, Elyas Obbad, and Sanmi Koyejo. Evaluating the robustness of Chinchilla compute-optimal scaling. *arXiv preprint arXiv:2509.23963*, 2025a.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53177–53266, 2025b. arXiv:2406.04391.

Xuyang Shen, Dong Li, Ruitao Leng, Zhen Qin, Weigao Sun, and Yiran Zhong. Scaling laws for linear complexity language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 16377–16426, 2024.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

Samuel L. Smith, Andrew Brock, Leonard Berrada, and Soham De. ConvNets match vision Transformers at scale. *arXiv preprint arXiv:2310.16764*, 2023.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 19523–19536, 2022. arXiv:2206.14486.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to Transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q. Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12342–12364, 2023. doi: 10.18653/v1/2023.findings-emnlp.825. arXiv:2207.10551.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2025.

B. L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.

Zexin Zhuang, Yanhang Li, and Zhichao Fan. Pre-registering the detectable effect: A paired-MDE budget for 4-bit quantization benchmarks, with a pilot audit. *arXiv preprint arXiv:2605.28873*, 2026.

A THE BROADENED AUDIT BED: FULL PER-COMPARISON TABLE

Table 4 lists every comparison in the broadened bed, with the fitted slopes on each paper's own axis, the gap, the Welch slope-difference p-value, the reversal probability, and the group used in the AUROC of Section 4.4. Every number is a fit to released per-scale data (raw tables or released CSVs); no models were retrained. Our refit of the Chinchilla form of Brandfonbrener et al. (2025) reproduces their released exponents within 0.012–0.021 ($R^2 = 0.999$), and their StarCoder against FineWeb-100B contrast has $t = 8.55$. Two rows are flagged SECONDARY (a robustness variant of another row, not an independent AUROC point) and one AMBIGUOUS (metric-dependent), leaving 22 ties and 6 separations in the discrimination set. Sources: Tay (Tay et al., 2023); Brandfonbrener/loss-to-loss (Brandfonbrener et al., 2025); LAION comparison suite (Nezhurina et al., 2025); Mamba (Gu & Dao, 2024) vs. Pythia (Biderman et al., 2023); Cerebras-GPT (Dey et al., 2023); DataDecide (Magnusson et al., 2025); OLMo scaling ladder (Bhagia et al., 2025); over-training corpora (Gadre et al., 2025) via the ColPret collection (Choshen et al., 2025). Two further results outside the 29 comparisons agree with the positive controls: the reconciled Kaplan–Chinchilla exponents (Pearce & Song, 2024) differ strongly when parameters are counted with or without embeddings (0.508 against 0.686, $t \approx 41$), and the OLMo-Hybrid fits (Merrill et al., 2026) tie on the exponent within the authors' own intervals while their Chinchilla coefficient separates.

Table 3: **Tay anchor, our refit.** Per-architecture power-law slope $\alpha_{F,U}$ (neg-log-ppl vs. $\log_{10}$ FLOPs) with our analytic SE (Tay et al. report none), and the Welch slope-difference test of the vanilla Transformer vs. each contender (claim: Transformer slope larger). Every comparison is a TIE. n is the number of model sizes.

Contender	$\hat{\alpha}_{\mathrm{Tf}}$	$\hat{\alpha}_X$	n	diff	Welch t	p	verdict
GLU-Transformer	0.543	0.498	5	+0.045	+0.48	0.649	tie
LConv	0.543	0.401	5	+0.142	+1.43	0.201	tie
Funnel	0.543	0.596	5	−0.053	−0.45	0.669	tie
Switch	0.543	0.361	5	+0.182	+1.55	0.176	tie
Universal	0.543	0.833	4	−0.290	−2.47	0.068	tie
Evolved	0.543	0.488	5	+0.055	+0.55	0.604	tie
Performer	0.543	0.487	4	+0.056	+0.40	0.714	tie
MoS-Transformer	0.543	0.464	5	+0.079	+0.77	0.473	tie
MLP-Mixer	0.543	0.459	4	+0.084	+1.01	0.361	tie

DataDecide full pairwise matrix. Beyond the eight representative pairs in Table 4, we ran the slope-difference test on all 300 recipe pairs on the common 14-size grid: 299 (99.7%) are ties on the exponent and a single extreme steepest-vs-shallowest pair separates ($p = 0.046$, itself metric-dependent). The 25 recipes span DCLM, FineWeb/FineWeb-Edu, C4, Dolma (multiple versions and ablations), Falcon and Falcon+CC (with QC variants), and code-inclusion ablations.

Unauditable / not-comparable cases encountered while broadening. Five headline or adjacent claims could not be added to the auditable bed, each strengthening the auditability finding of Section 4.6: (i) OLMo-2 production runs (Team OLMo et al., 2025) release checkpoint CSVs with only (step, URL) and no per-scale loss table (per-step loss is on WandB, not size-indexed); (ii) Mamba's headline loss-vs-FLOPs Pareto figure (Gu & Dao, 2024) is image-only with no numeric per-method table (we used the adjacent zero-shot perplexity table instead); (iii) the Mamba repository ships weights and eval scripts but no numeric result tables; (iv) t5-Pile (encoder–decoder) vs. Pythia (decoder-only) is a cross-objective pair whose loss magnitudes are not comparable (excluded, not a valid slope-difference); (v) the ColPret "chinchilla" family is digitized from a scatter figure (142 points, not distinct trained models) and excluded from clean fits.

B OPTIMAL-DESIGN DETAILS

The $\sqrt{3}$ log-spacing tax. For a slope estimated over a fixed window of width W, $\mathrm{Var}(\hat{\alpha}) = \sigma^2/(rS_{xx})$, so efficiency is governed entirely by $S_{xx} = \sum_i (x_i - \bar{x})^2$. Placing all J points at the two endpoints $\pm W/2$ gives $S_{xx} = JW^2/4$ (the c-optimal design for a slope (Elfv-

Table 4: **Broadened audit bed, 29 resolved comparisons (+2 secondary).** $\hat{\alpha}_A, \hat{\alpha}_B$ are fitted slopes on the source's own axis (sign per source; a larger magnitude means faster gain per decade). "gap" $= \hat{\alpha}_A - \hat{\alpha}_B$ in the claimed direction; p is the Welch slope-difference test; P_{rev} the reversal probability. Group: TIE/SEP enter the AUROC, SEC/AMB excluded.

Comparison	$\hat{\alpha}_A$	$\hat{\alpha}_B$	gap	p	P_{rev}	grp
Tay: Transformer vs. GLU-Transformer	0.543	0.498	+0.045	0.649	0.350	tie
Tay: Transformer vs. Universal	0.543	0.833	−0.290	0.068	0.932	tie
Tay: Transformer vs. Evolved	0.543	0.488	+0.055	0.604	0.332	tie
Tay: Transformer vs. MoS-Transformer	0.543	0.464	+0.079	0.473	0.271	tie
Tay: Transformer vs. Funnel	0.543	0.596	−0.053	0.669	0.643	tie
Brandfonbrener: FineWeb-Edu vs. SlimPajama (text/text)	0.470	0.454	+0.016	0.189	0.096	tie
Brandfonbrener: StarCoder vs. FineWeb-100B (code/text)	0.558	0.439	+0.119	$< 10^{-4}$	0.000	**sep**
Brandfonbrener: Proof-Pile-2 vs. FineWeb-100B (code/text)	0.511	0.439	+0.072	$< 10^{-4}$	0.000	**sep**
Brandfonbrener: StarCoder vs. SmolLM-Corpus (code/text)	0.558	0.469	+0.089	$< 10^{-4}$	0.000	**sep**
MaMMUT vs. CLIP (self-reproduced)	−0.161	−0.141	−0.021	0.00028	0.0002	**sep**
CoCa vs. CLIP	−0.145	−0.141	−0.004	0.568	0.287	tie
Cap vs. CLIP	−0.100	−0.141	+0.040	$< 10^{-6}$	0.000	**sep**
SigLIP vs. CLIP	−0.134	−0.141	+0.007	0.298	0.153	tie
Mamba vs. Pythia (drop undertrained 160M)	−0.171	−0.148	−0.023	0.534	0.288	tie
Mamba vs. Pythia (all points)	−0.171	−0.366	+0.195	0.155	0.921	sec
Cerebras-GPT vs. Pythia (xent vs. FLOPs)	−0.055	−0.089	+0.035	0.00148	0.0008	**sep**
Cerebras-GPT vs. Pythia (xent vs. N)	−0.104	−0.085	−0.019	0.036	0.021	sec
Pythia vs. Pythia-deduped	−0.085	−0.085	+0.000	0.981	0.492	tie
DataDecide: DCLM-Baseline vs. C4	−0.487	−0.465	−0.022	0.684	0.347	tie
DataDecide: FineWeb-Edu vs. C4	−0.524	−0.465	−0.059	0.282	0.152	tie
DataDecide: FineWeb-Edu vs. DCLM-Baseline	−0.524	−0.487	−0.037	0.482	0.251	tie
DataDecide: Dolma1.7 vs. Dolma1.6++	−0.444	−0.443	−0.001	0.983	0.494	tie
DataDecide: Falcon+CC vs. Falcon	−0.543	−0.517	−0.026	0.631	0.322	tie
DataDecide: Dolma1.7 vs. Dolma1.7 (no code)	−0.444	−0.484	+0.040	0.442	0.769	tie
DataDecide: DCLM-Baseline vs. Dolma1.7	−0.487	−0.444	−0.043	0.408	0.214	tie
DataDecide: Falcon+CC (QC10%) vs. Dolma1.7 (no Reddit)	−0.549	−0.443	−0.107	0.046	0.028	amb
OLMo ladder: 1×C vs. 10×C (N-exp)	−0.083	−0.083	−0.000	0.961	0.483	tie
OLMo ladder: 0.5×C vs. 5×C (N-exp)	−0.083	−0.083	+0.000	0.977	0.489	tie
Over-train: RedPajama vs. C4 (N-exp)	−0.106	−0.106	+0.000	0.993	0.498	tie
Over-train: RedPajama vs. RefinedWeb (N-exp)	−0.106	−0.102	−0.004	0.836	0.430	tie
Over-train: C4 vs. RefinedWeb (N-exp)	−0.106	−0.102	−0.004	0.833	0.429	tie

ing, 1952)); uniform log-spacing gives $S_{xx} = W^2 J(J+1)/[12(J-1)] \to JW^2/12$. The SD ratio is therefore $\sqrt{(JW^2/4)/(JW^2/12)} = \sqrt{3}$ in the large-J limit. Monte-Carlo simulation of both designs (Table 5) confirms the closed form; the ratio approaches $\sqrt{3} \approx 1.732$ from below.

Table 5: **Log-spacing pays a $\sqrt{3}$ efficiency tax.** Ratio of $\mathrm{SD}(\hat{\alpha})$ under uniform log-spacing to the endpoint (c-optimal) design; closed form vs. Monte Carlo (10^5 draws/cell). The limit is $\sqrt{3} \approx 1.732$.

J	4	6	8	12	24	48	$\to \infty$
closed form	1.342	1.464	1.528	1.593	1.661	1.696	1.732
Monte Carlo	1.338	1.475	1.505	1.599	1.655	1.690	—

Form–power tradeoff. The endpoint design that minimizes slope variance cannot detect curvature: with no interior point, a quadratic departure from log-linearity is unidentifiable. Requiring 80% power to reject a curvature of magnitude Δ_{form} imposes a lower bound on the number of interior points, while every point moved to the interior raises $\mathrm{SD}(\hat{\alpha})$. Writing k for the inte-

rior count and minimizing the slope-difference SD subject to the curvature-power constraint yields $k^\star = \frac{1}{2}(J - \sqrt{J^2 - 4Q})$, where Q collects $(\sigma, \Delta_{\text{form}}$, target power, $R)$. For $J = 16$, $R = 1.5$ decades, $\sigma = 0.06$, $\Delta_{\text{form}} = 0.22$ we get $Q = 50.1$, $k^\star = 4.27$ (round to 5; interior fraction 0.27). The resulting design has $\text{SD}(\alpha_A - \alpha_B) = 0.0143$, versus 0.0123 for pure-endpoint (a $1.17\times$ tax) and 0.0200 for uniform log-spacing (a $1.39\times$ tax): the constrained optimum recovers most of the endpoint efficiency while remaining able to validate the power-law form.

Exact finite-sample check. The min-points guidance of Section 3.3 uses the asymptotic normal MDE; replacing z by the exact noncentral-t solution (Welch df $= 2(J-2)$) changes the required J by at most one point over the regime of interest (e.g. $\delta = 0.05$, $\sigma = 0.05$, $R = 2$: asymptotic $J = 9$, exact $J = 9$; at $\sigma = 0.08$: asymptotic $J = 23$, exact $J = 22$), so the "$\gtrsim 15$ sizes over $\gtrsim 3$ decades" prescription is robust to the small-sample correction.

Simulation check of the floor. The closed-form power matches a simulation of the full pipeline (300 designs with 4000 replicates each) for $J \geq 4$ with a mean absolute error of 0.015, and the test holds its nominal false-positive rate (0.0397 against 0.05; conservative below $J = 6$). Only $J = 3$, with a single residual degree of freedom, departs: there the closed form over-predicts power. On the Tay data the estimated gap has a standard deviation of 0.043; with a planted gap of 0.045 in the Tay design, the verdict turns into a tie at $\sigma^\star = 0.050$, where the gap equals $z \cdot \text{SD}(\delta)$, against a simulated crossing near $\sigma \approx 0.04$ (Figure 3b).

C Additional tables

Table 6 is the breadth audit of five further high-profile claims (Section 4.6), and Table 7 tabulates each audited anchor's claimed gap against its own 80%-power detection floor (the numeric companion to Figure 3c, Section 3.3).

Table 6: **Breadth audit of five further high-profile comparative scaling claims.** The dominant verdict is UNAUDITABLE: no released per-scale numbers. "Adjacent" denotes a testable sub-claim when the headline is image-only.

Claim (arXiv)	subfield	verdict
RetNet > Transformer, crossover ∼2B (2307.08621)	retention vs. Transf.	unauditable (fig-only)
Mamba matches/beats strong Transformer scaling (2312.00752)	SSM vs. Transf.	headline unauditable; adjacent **tie on slope**
ConvNets match ViTs at scale (2310.16764)	vision	unauditable (cross-paper, fig-only)
4-bit is optimal for total model bits (2212.09720)	quantization	unauditable (fig-only)
Diffusion beats AR when data-constrained (2507.15857)	diffusion vs. AR	headline unauditable; adjacent robust (direction)
Reconciling Kaplan–Chinchilla [pos. control] (2406.12907)	methodology	robust / agrees

Table 7: **Claimed exponent gap vs. the design's own 80%-power floor.** $3/6$ resolvable-design anchors have their claimed gap below their own detection floor (undetectable by construction). Brandfonbrener et al.'s real gaps sit above its floor—the bound is a discriminator, not a hammer. (The Tay "Universal" row is the mis-extracted architecture, shown for completeness; its wrong-direction gap is not part of the tie argument.)

Anchor comparison	claimed gap	MDE_{80}	(J, R, σ)	verdict
Tay: Transf. vs. GLU-Transformer	0.0454	0.145	$(5, 1.68, 0.0931)$	below floor
Tay: Transf. vs. Evolved	0.0554	0.155	$(5, 1.72, 0.102)$	below floor
Tay: Transf. vs. MoS-Transformer	0.0794	0.164	$(5, 1.72, 0.108)$	below floor
Tay: Transf. vs. Universal	0.290	0.246	$(4, 1.19, 0.0947)$	above (mis-extr.)
Brandfonbrener: StarCoder vs. FineWeb	0.119	0.003	(many pts)	above (real)
Brandfonbrener: Proof-Pile-2 vs. FineWeb	0.072	0.002	(many pts)	above (real)