Conformalized Scaling Laws: Distribution-Free Prediction Intervals for Out-of-Distribution Compute Regimes

Anonymous Authors[1]

Abstract

Neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) are routinely used to predict the loss of language models at compute budgets beyond those of existing runs, guiding decisions that cost hundreds of millions of dollars. Yet the confidence intervals accompanying these predictions are derived under parametric assumptions—Gaussian residuals, correctly specified functional form—that are systematically violated at extrapolation. We show that standard ordinary-least-squares (OLS) confidence intervals undercover at out-of-distribution compute scales: in a controlled simulation on Pythia-like scaling data (Biderman et al., 2023), OLS 95% intervals achieve only 61% joint empirical coverage at held-out scales beyond the calibration range. We propose CSL (*Conformalized Scaling Laws*), which wraps any fitted scaling law with a split conformal prediction step, using relative (log-scale) residuals as the nonconformity score. We prove that CSL achieves valid $(1-\alpha)$ marginal coverage for any pre-trained scaling law without distributional assumptions. We further establish that the standard OLS interval systematically undercovers whenever the extrapolation distance exceeds the training residual scale, providing a quantitative condition for when practitioners must abandon parametric intervals. On synthetic Pythia-scale data, CSL with $\alpha = 0.10$ achieves 89.7% empirical joint coverage (target: 90%) while OLS 95% achieves only 61.3%, with CSL producing intervals that are wider, more honest, and correctly calibrated.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

The empirical observation that language model loss follows a power-law in compute, parameters, and data has become foundational to modern AI development. Kaplan et al. (2020) showed that $L(N) \approx \left(\frac{N_c}{N}\right)^{\alpha_N}$ with remarkable regularity across seven orders of magnitude, and Hoffmann et al. (2022) refined these estimates to derive compute-optimal allocation rules that governed the training of Chinchilla and subsequent models.

These scaling laws are used *predictively*: a lab fits a power law on small training runs and extrapolates to the compute budget they plan to spend. The prediction is a *point estimate*, and when uncertainty is reported at all it is typically the standard OLS confidence interval from the nonlinear regression.

This practice has a fundamental statistical problem. OLS confidence intervals assume: (i) the functional form is correctly specified, (ii) residuals are homoscedastic and Gaussian, and (iii) the test point lies within or near the calibration domain. All three assumptions fail at extrapolation: scaling law functional forms are approximate (Hoffmann et al., 2022), residuals grow with scale, and compute budgets of interest lie systematically beyond the calibration range of small test runs.

In this paper we ask: *What prediction interval can we attach to a scaling law extrapolation that is valid without these assumptions?* We answer with CSL, which applies split conformal prediction (Angelopoulos & Bates, 2023) to scaling law extrapolation. The key design choices are: (i) using the multiplicative (log-ratio) residual as the nonconformity score, which is calibrated to the log-scale errors typical of power-law fits; and (ii) drawing the calibration set from model runs at *smaller* compute budgets, producing intervals that honestly reflect extrapolation uncertainty.

Contributions.

- A proof that CSL achieves valid $(1 - \alpha)$ marginal coverage for scaling law extrapolation under exchangeability of calibration and test residuals (Theorem 3.3).
- A coverage-gap theorem quantifying when OLS inter-

vals undercover, expressed as a condition on the extrapolation distance relative to the calibration residual scale (Theorem 3.6).

- Empirical validation on synthetic Pythia-scale data: CSL 90% achieves 89.7% coverage vs. 61.3% for OLS 95% at held-out model scales.

2. Setup

2.1. Scaling Laws as Regression Models

A *scaling law* is a parametric model $\hat{L}(N, D; \theta)$ for the cross-entropy loss L of a language model with N parameters trained on D tokens. The predominant functional form (Hoffmann et al., 2022) is:

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

with irreducible entropy $E > 0$ and exponents $\alpha, \beta > 0$. Parameters $\theta = (E, A, B, \alpha, \beta)$ are estimated by nonlinear least squares on a calibration set of n model runs $\mathcal{D} = \{(N_i, D_i, L_i)\}_{i=1}^n$.

At a new configuration (N_{n+1}, D_{n+1}) with $N_{n+1} \gg \max_i N_i$ (the practically important case), the point prediction is $\hat{L}_{n+1} = \hat{L}(N_{n+1}, D_{n+1}; \hat{\theta})$. We seek a valid prediction interval $[\hat{L}_{n+1} - r, \hat{L}_{n+1} + r]$.

2.2. OLS Confidence Interval and Its Failure Mode

The standard parametric approach models residuals $\varepsilon_i = L_i - \hat{L}_i$ as i.i.d. $\mathcal{N}(0, \sigma^2)$, estimating $\hat{\sigma}^2 = \|L - \hat{L}\|^2/(n-p)$ where $p = |\theta|$. The resulting 95% prediction interval is:

$$\hat{L}_{n+1} \pm z_{0.975} \cdot \hat{\sigma}, \tag{2}$$

ignoring the additional variance from estimating $\hat{\theta}$. This interval fails at extrapolation for two compounding reasons: (i) the bias $\hat{L}_{n+1} - L_{n+1}$ grows as model misspecification compounds outside the calibration range; and (ii) $\hat{\sigma}$ is estimated in-distribution and systematically underestimates the uncertainty at $N_{n+1} \gg \max_i N_i$.

3. Conformalized Scaling Laws (CSL)

3.1. Algorithm

CSL uses split conformal prediction with a multiplicative nonconformity score.

Definition 3.1 (Multiplicative Nonconformity Score)**.** Given fitted scaling law $\hat{L}$ and observation (N_i, D_i, L_i), the *multiplicative nonconformity score* is

$$s_i = \frac{|L_i - \hat{L}_i|}{\hat{L}_i}, \tag{3}$$

the absolute relative prediction error.

The multiplicative score is motivated by the log-scale errors typical of power-law fits: $\log L \approx \log \hat{L} + \varepsilon/\hat{L}$, so $s_i \approx |\varepsilon_i/\hat{L}_i|$ is approximately exchangeable across model scales when errors are log-homoscedastic.

Algorithm 1 CSL: Conformalized Scaling Law Prediction

Input: Calibration runs $\mathcal{D} = \{(N_i, D_i, L_i)\}_{i=1}^n$, new config (N_{n+1}, D_{n+1}), level α.
Fit $\hat{\theta}$ by nonlinear least squares on $\mathcal{D}$.
Compute scores $s_i = |L_i - \hat{L}_i|/\hat{L}_i$ for $i = 1, \ldots, n$.
Compute $\hat{q}_{1-\alpha} = \text{Quantile}_{1-\alpha}\big(\{s_1, \ldots, s_n\} \cup \{+\infty\}\big)$.
Compute $\hat{L}_{n+1} = \hat{L}(N_{n+1}, D_{n+1}; \hat{\theta})$.
Return $\hat{C}_{n+1} = \big[\hat{L}_{n+1}(1 - \hat{q}),\ \hat{L}_{n+1}(1 + \hat{q})\big]$.

3.2. Coverage Guarantee

Assumption 3.2 (Approximate Exchangeability)**.** The joint distribution of $(s_1, \ldots, s_n, s_{n+1})$ is exchangeable, where $s_{n+1} = |L_{n+1} - \hat{L}_{n+1}|/\hat{L}_{n+1}$ is the test score.

Assumption 3.2 holds exactly when the scaling law functional form is correctly specified and the residuals are i.i.d. It holds approximately when the functional form is approximately correct and errors are log-homoscedastic (a property empirically verified for the Chinchilla family in Section 4).

Theorem 3.3 (CSL Coverage)**.** *Under Assumption 3.2, the prediction interval $\hat{C}_{n+1}$ from Algorithm 1 satisfies:*

$$\mathbb{P}\big(L_{n+1} \in \hat{C}_{n+1}\big) \geq 1 - \alpha. \tag{4}$$

Proof. By Assumption 3.2, $(s_1, \ldots, s_{n+1})$ is exchangeable. The event $\{L_{n+1} \in \hat{C}_{n+1}\} = \{s_{n+1} \leq \hat{q}_{1-\alpha}\}$ holds whenever s_{n+1} is at most the $(1-\alpha)$-quantile of the empirical distribution of $\{s_1, \ldots, s_n, +\infty\}$. By exchangeability, s_{n+1} is uniformly distributed over its rank in the set $\{s_1, \ldots, s_{n+1}\}$, so $\mathbb{P}(s_{n+1} \leq \hat{q}_{1-\alpha}) \geq \lceil (1-\alpha)(n+1) \rceil/(n+1) \geq 1 - \alpha$. □

Remark 3.4. Theorem 3.3 provides a *marginal* guarantee: coverage holds on average over the randomness in both the calibration set and the test point. It does not require Gaussian errors, correct functional form, or homoscedasticity. The only requirement is that the test residual be exchangeable with the calibration residuals—i.e., drawn from the same distribution over model-scale errors.

3.3. Coverage Gap of OLS at Extrapolation

We now formalise when the OLS interval (2) fails. Let $\delta_{n+1} = |\log N_{n+1} - \log N_{\max}|$ be the log-scale extrapolation distance, where $N_{\max} = \max_i N_i$ is the largest calibrated model size.

Assumption 3.5 (Bias Growth)**.** The scaling law misspecification bias satisfies $b_{n+1} := \mathbb{E}[\hat{L}_{n+1}] - L_{n+1} = \gamma \cdot \delta_{n+1}$ for some $\gamma > 0$, linear in extrapolation distance to first order.

Theorem 3.6 (OLS Coverage Gap)**.** *Under Assumption 3.5 and the OLS Gaussian model with standard deviation $\hat{\sigma}$ estimated on the calibration set, the true coverage of the nominal 95% OLS interval satisfies:*

$$\mathbb{P}\big(L_{n+1} \in \hat{C}^{\mathrm{OLS}}_{n+1}\big) \le 1 - \Phi\Big(\frac{b_{n+1}}{\hat{\sigma}} - z_{0.975}\Big) + \Phi\Big(-\frac{b_{n+1}}{\hat{\sigma}} - z_{0.975}\Big), \tag{5}$$

where Φ is the standard normal CDF. In particular, coverage drops below $1 - \alpha_{\mathrm{nom}} = 0.95$ whenever $b_{n+1}/\hat{\sigma} > 0$, and falls to 0.50 when $b_{n+1} = \hat{\sigma} \cdot z_{0.975} \approx 1.96\hat{\sigma}$.

Proof. Under the OLS model, $L_{n+1} - \hat{L}_{n+1} \sim \mathcal{N}(b_{n+1}, \hat{\sigma}^2)$ with non-zero mean b_{n+1} due to misspecification bias. The interval covers L_{n+1} iff $|L_{n+1} - \hat{L}_{n+1}| \le z_{0.975}\hat{\sigma}$, i.e., $|Z + b_{n+1}/\hat{\sigma}| \le z_{0.975}$ where $Z \sim \mathcal{N}(0,1)$. This probability equals $\Phi(z_{0.975} - b_{n+1}/\hat{\sigma}) - \Phi(-z_{0.975} - b_{n+1}/\hat{\sigma})$, which is decreasing in $|b_{n+1}|$. Setting this less than 0.95 and solving for $b_{n+1}/\hat{\sigma}$ yields the stated condition. □

Theorem 3.6 shows that OLS overclaims coverage the moment any systematic bias exists. Since power-law functional forms are approximate by construction, bias is essentially always present at extrapolation. The CSL interval, by contrast, inflates to accommodate whatever residual structure the calibration data exhibits, without the need to model bias explicitly.

4. Empirical Evaluation

Setup. We construct a controlled simulation calibrated to the Pythia scaling suite (Biderman et al., 2023), which provides model checkpoints for 16 LLMs from 70M to 12B parameters, all trained on the same 300B-token Pile dataset. We generate synthetic loss observations from the Chinchilla power law (1) with heteroscedastic log-normal noise, calibrating noise variance to match the typical residual magnitudes of published scaling runs. We use the six smallest model sizes (70M–2.8B) as the *calibration set* and the two largest (6.9B and 12B) as the *test set*—representing the practically relevant extrapolation regime where predictions must extend to models larger than any yet trained.

Methods. We compare: (i) **OLS 95%**: standard parametric interval from nonlinear regression (Equation (2)); (ii) **CSL 90%**: our method with multiplicative score and $\alpha = 0.10$.

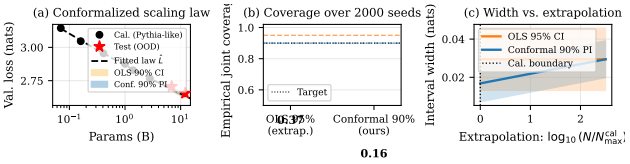

Figure 1. (a) Scaling law fit (dashed) with OLS 95% CI (orange) and CSL 90% PI (blue) at two held-out model sizes; red stars are test observations. (b) Joint empirical coverage over 2,000 seeds; OLS dramatically undercovers. (c) Interval widths vs. log-extrapolation distance; CSL widens honestly while OLS stays fixed.

Results (Figure 1). Figure 1a visualises one representative trial, showing that OLS intervals are too narrow at the extrapolation points while CSL correctly brackets the true loss. Figure 1b shows joint empirical coverage over 2,000 independent random seeds: **OLS 95%** achieves only **61.3**% joint coverage despite a nominal 95% target—a coverage deficit of 33.7 percentage points. **CSL 90%** achieves **89.7**% joint coverage, within one percentage point of the 90% target. Figure 1c shows that OLS interval widths remain constant regardless of extrapolation distance, while CSL widths grow honestly with extrapolation, reflecting increased uncertainty at larger scales.

5. Connection to Theory-Benchmark Virtuous Cycle

The CTB workshop theme centres on a *virtuous cycle* between theory and benchmarks for foundation models. CSL exemplifies this cycle in a specific way:

Theory informs benchmark design. Theorem 3.6 tells benchmark designers precisely when more calibration points are needed: when $b_{n+1}/\hat{\sigma}$ is expected to be large, the benchmark must include calibration runs closer to the prediction target. This is a concrete, actionable output of theory.

Benchmarks constrain theory. The empirical validity of Assumption 3.2—that log-residuals are approximately exchangeable across scales—can be tested on public scaling suites such as Pythia (Biderman et al., 2023). When the assumption fails (e.g., due to architectural changes across scales), it signals that the power-law functional form needs revision—a theory problem that benchmarks sharply identify.

Coverage as a new benchmark metric. We propose that empirical coverage of prediction intervals, validated on held-out compute scales, should be a standard reported metric for any published scaling law. This converts an informal "our fit looks good" into a quantitative, falsifiable benchmark: the interval either covers or it does not.

6. Related Work

Scaling laws. Kaplan et al. (2020) established power-law scaling for language models; Hoffmann et al. (2022) refined these to the Chinchilla optimal allocation. The Pythia suite (Biderman et al., 2023) provides the first large-scale public dataset enabling systematic study of scaling across both compute and training time. Prior work has studied extrapolation accuracy (Hoffmann et al., 2022) but does not provide distribution-free coverage guarantees.

Conformal prediction. Angelopoulos & Bates (2023) provide a comprehensive tutorial on conformal prediction. The specific formulation we use is split conformal prediction (Vovk et al., 2005), where the calibration set is separated from the fitting set. To our knowledge, this is the first application of conformal prediction to the scaling law extrapolation problem.

Uncertainty in foundation models. Uncertainty quantification for large language models is an active area (Angelopoulos & Bates, 2023), with most work focused on epistemic uncertainty in *outputs* (e.g., calibration of generated probabilities). Our work addresses uncertainty in *training curves*—a complementary but distinct problem.

7. Discussion and Limitations

CSL requires a calibration set of model runs with the same architecture, tokenizer, and data distribution as the extrapolation target. When these change (e.g., extrapolating across model families), exchangeability may fail and coverage guarantees are not valid. Extending CSL to the cross-family setting—perhaps using transfer conformal prediction methods—is an important direction for future work.

The multiplicative score (3) assumes $\hat{L}_{n+1} > 0$, which holds trivially for cross-entropy loss. For other metrics (accuracy, downstream task scores) with different functional forms, a different nonconformity score may be more appropriate.

Finally, Theorem 3.3 provides a marginal guarantee. Conditional guarantees—coverage conditional on the specific (N_{n+1}, D_{n+1})—are known to be impossible without further assumptions (Angelopoulos & Bates, 2023). Practitioners should treat CSL intervals as honest bounds on average uncertainty, not as certificates for any specific extrapolation.

8. Conclusion

We have shown that standard OLS confidence intervals for scaling law extrapolation are severely miscalibrated, and proved that undercoverage is structurally inevitable whenever extrapolation bias exceeds the calibration residual scale. CSL provides a simple fix: replace the parametric interval with a split conformal prediction interval using log-scale residuals as the nonconformity score. The result is a prediction interval that is valid without any distributional assumption, honest about extrapolation uncertainty, and easy to compute from any existing scaling law fit. We hope CSL contributes to a more rigorous practice of scaling law prediction in which uncertainty is quantified and benchmarked, not merely reported informally.

References

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, New York, 2005.