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

Joshua Shay Kricheli[‖] Alexander Lawrence Reid[‖] Soumajyoti Sarkar
Venkata Gandikota Paulo Shakarian

Syracuse University: Leibniz Lab, Institute for Artificial Intelligence Amazon AGI Foundations[*]
[‖]Equal contribution. [*]This work does not relate to the author's position at Amazon.

NeurIPS 2026, Sydney, Australia



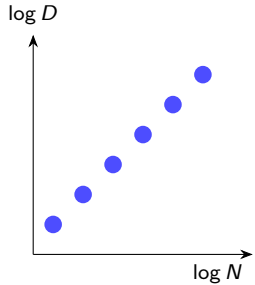




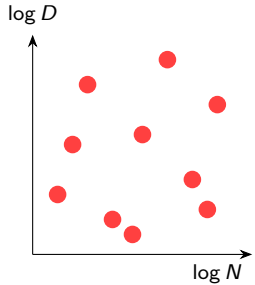


Motivation and collinear designs

- LLM runs cost millions: scaling laws [1, 2] pick N, D
- Chinchilla rule: $k = D/N \approx 20$ [1], one ray
- On a ray [1]: $L = AN^{-\alpha} + Bk^{-\beta}N^{-\beta} + E$
- Only $\varepsilon = |\alpha - \beta|$ separates A, B



Collinear (CO)



Non-collinear (NC)

[1] Hoffmann et al., NeurIPS 2022. [2] Kaplan et al., arXiv 2020.

Theory: when the scale pair is identified

- One ray: $\kappa_{A,B} = \Theta(\varepsilon^{-2})$; CI on A is $\Omega(\varepsilon^{-1})\times$ CI on $\psi = A + Bk^{-\alpha}$
- K rays: $\kappa_{A,B} \le \kappa_{\mathrm{target}}$ (leading order) whenever

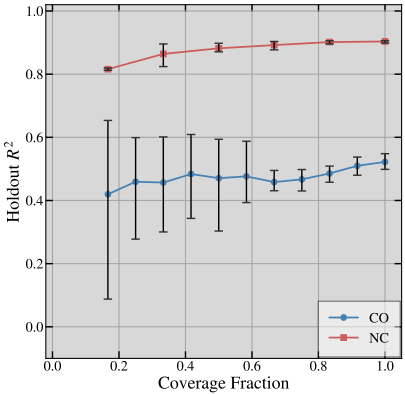
$$\underbrace{\frac{1}{K}\sum_{\ell=1}^{K} k_\ell^{-2\beta_{\mathrm{eff}}} - \left(\frac{1}{K}\sum_{\ell=1}^{K} k_\ell^{-\beta_{\mathrm{eff}}}\right)^2}_{V_K} \;\ge\; \underbrace{\frac{\left(K + \sum_\ell k_\ell^{-2\beta_{\mathrm{eff}}}\right)^2}{K^2\,\kappa_{\mathrm{target}}}}_{\tau_K},$$

- Exact: $\kappa_{A,B} \le \kappa_{\mathrm{target}}$ iff $V_K \ge \tau_K/(1 + 1/\kappa_{\mathrm{target}})^2$
- One ray never passes; two pass once $k_2/k_1 \ge R_{\min}(k_1)$
- Checkable before training

Results: non-collinear designs extrapolate

Comparison (NC vs. CO)	NC wins
Full fan, FP32 (1,460 of 1,500)	97.3%
Full fan, BF16 (296 of 300)	98.7%
Budget-matched, Chinchilla [1]	68.5%
Budget-matched, Droppo-Elibol [3]	57.4%
Budget-matched, Kaplan [2]	80.6%

- 1,900 models (5M-77M), 5 corpora, 4 laws [1-4]
- Budget-matched: CO subsets below threshold
- Wide CO bars: sloppy (A, B) [5], $\kappa_{A,B} = \Theta(\varepsilon^{-2})$; NC bars tight



Holdout R^2 vs. coverage, Kaplan [2] on RedPajama

30 seeds: mean and 5th-95th percentile bars

[1] Hoffmann et al., NeurIPS 2022. [2] Kaplan et al., arXiv 2020. [3] Droppo and Elibol, Interspeech 2021. [4] Muennighoff et al., Scaling data-constrained language models, JMLR 2025. [5] Transtrum et al., PRL 2010.

Takeaway and design recipe

- Fit on designs that span TPP ratios
- Recipe: pick k_1, read $R_{\min}(k_1)$, train both ends, split runs evenly
- Check exact $\kappa_{A,B}$ before training
- Code, data, checkpoints: public

  `lab-v2.github.io/tpp-coverage`

