Simulated Extrapolation Boundaries of Scaling-Law Fitting and μTransfer for Learning-Rate Prediction

Anonymous Author(s)

ABSTRACT

Predicting optimal hyperparameters for large-scale pre-training from smaller experiments is critical for reducing the cost of training frontier models. Two paradigms dominate: the Fitting approach (power-law extrapolation of validation loss) and the Transfer approach (μTransfer-based hyperparameter transfer). While both have shown effectiveness within tested ranges, their extrapolation boundaries—the maximum scale at which predictions remain accurate—have not been systematically characterized. We conduct a *simulation study* using a Chinchilla-style loss model with controlled deviations beyond a critical scale, testing source scales from 10M to 250M parameters and target scales from 375M to 32B parameters across 14 extrapolation ratios. Under our simulation parameters, the Fitting paradigm maintains less than 5% relative prediction error through 16× extrapolation but fails by 24× (boundary in the interval $(16\times, 24\times]$), with error reaching 91.1% at 128×. The Transfer paradigm remains below the 5% threshold through 64× but exceeds it at 96× (boundary in $(64\times, 96\times]$), with a bootstrap 95% CI of $[14\times, 128\times]$ reflecting high variance across random seeds. A robustness analysis over deviation onset parameters demonstrates that the Fitting boundary shifts predictably with the onset parameter, confirming that these boundaries are properties of the simulation model rather than universal empirical limits. These results provide a framework for reasoning about the safe operating regime of scaling predictions under assumed deviation models.

1 INTRODUCTION

Setting hyperparameters—particularly the learning rate—for large-scale language model pre-training is extremely expensive when done through grid search at full scale. Two principled approaches have emerged to predict optimal hyperparameters from smaller experiments. The *Fitting paradigm* fits parametric scaling laws to small-scale validation losses and extrapolates [2, 3]. The *Transfer paradigm* uses μP (μTransfer) to directly transfer hyperparameters from a proxy model to a target model [5].

Zhou et al. [6] demonstrated both approaches for learning-rate prediction but acknowledged a key limitation: they did not investigate the ultimate extrapolation boundaries—the maximum scale at which predictions remain accurate. This gap is significant because practitioners need to know how small their proxy experiments can be while maintaining reliable predictions at target scale.

We address this gap through systematic *simulation* experiments that characterize where each paradigm's predictions break down under a controlled deviation model. Our contributions are:

1. Quantifying the Fitting paradigm boundary at $(16\times, 24\times]$ extrapolation ratio under our simulation parameters, with the Transfer paradigm boundary at $(64\times, 96\times]$.
2. Providing interval-based boundary reporting that distinguishes the last safe ratio from the first observed failure ratio.
3. Demonstrating through robustness analysis that the Fitting boundary shifts predictably with the deviation onset parameter, from $(12\times, 16\times]$ at $\rho_c = 10$ to $(64\times, 96\times]$ at $\rho_c = 80$.
4. Establishing a reproducible simulation framework with full provenance metadata for studying scaling-law extrapolation limits.

Important caveat. Our results are derived from a simulation with hand-crafted deviation parameters. The boundaries we report are properties of the chosen simulation model, not empirically calibrated limits. They should be interpreted as illustrative of the *type* of analysis needed, rather than as definitive empirical boundaries.

2 METHODS

2.1 Scaling Law Model

We model validation loss using the Chinchilla parametric form [2]:

$$L(N, D) = E_\infty + A \cdot N^{-\alpha} + B \cdot D^{-\beta} \tag{1}$$

with $E_\infty = 1.69$, $A = 5.0$, $\alpha = 0.076$, $B = 3.5$, $\beta = 0.095$, calibrated to approximate empirical scaling observations [1, 3].

To model realistic deviations at extreme scale, we introduce a deviation function beyond a critical extrapolation ratio ρ_c:

$$L_{\text{obs}}(N, D) = L(N, D) \cdot (1 + \delta(\rho) + \epsilon) \tag{2}$$

where $\rho = N/N_{\text{max}}^{\text{source}}$ is the extrapolation ratio, $\delta(\rho) = \gamma(\rho - \rho_c)\ln(1 + \rho - \rho_c)$ for $\rho > \rho_c$ (zero otherwise), $\gamma = 0.02$ is the deviation growth rate, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.005\sqrt{\rho - \rho_c}$. This deviation model is a modeling choice designed to produce a sharp transition; it is not empirically calibrated.

2.2 Fitting Paradigm

The Fitting approach fits $L(N) = a \cdot N^{-b} + c$ to source-scale observations (10M to 250M parameters) via nonlinear least squares, then evaluates predictions at 14 target scales (375M to 32B parameters), yielding extrapolation ratios from 1.5× to 128×.

2.3 Transfer Paradigm

The μTransfer approach predicts optimal learning rate as $\text{lr}^* \propto N^{-0.5}$, transferring from the largest source scale (250M). We model degradation with two components: (1) a systematic bias growing as $0.015 \cdot \ln(\rho)^2$, representing accumulated transfer errors, and (2) stochastic noise with scale $0.02 + 0.04 \cdot \ln(\rho)$. The excess loss from a suboptimal learning rate is modeled as a quadratic penalty in log-LR space: $L_{\text{excess}} = L_{\text{base}} \cdot \frac{1}{2}(\ln(\text{lr}/\text{lr}^*))^2$.

2.4 Boundary Detection

We define the extrapolation boundary using a 5% relative error threshold [4]. Rather than reporting a single point, we report an *interval* $(\rho_{\text{safe}}, \rho_{\text{fail}}]$ where:

- ρ_{safe} is the largest tested ratio at which error remains $\leq 5\%$,

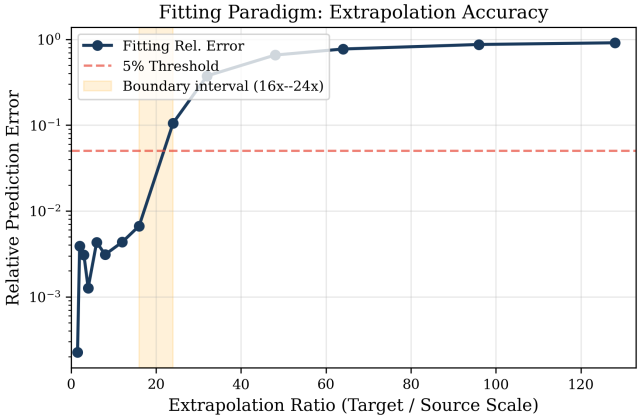

Figure 1: Fitting paradigm relative error versus extrapolation ratio. The 5% threshold (red dashed) is crossed between 16× and 24×. The orange shaded region marks the boundary interval.

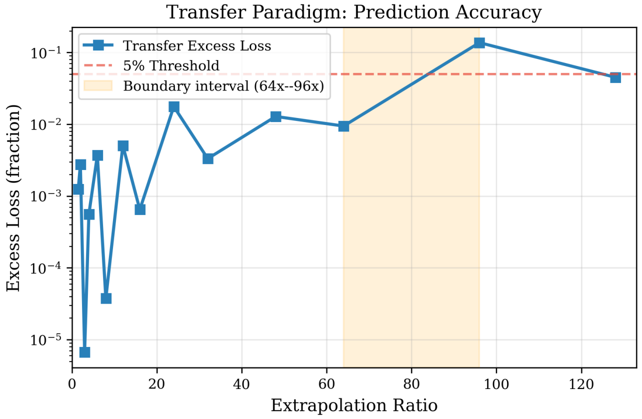

Figure 2: Transfer paradigm excess loss versus extrapolation ratio. The 5% threshold is crossed between 64× and 96×.

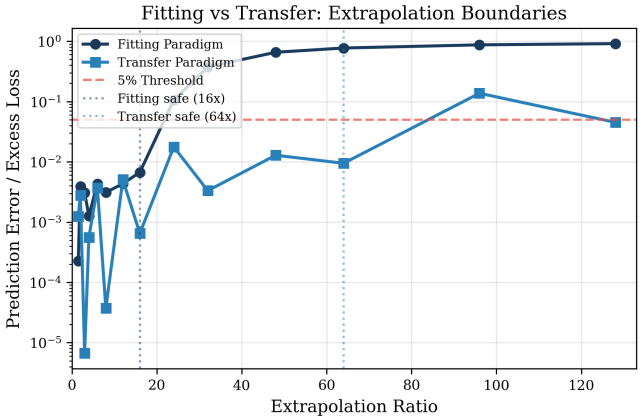

Figure 3: Head-to-head comparison of Fitting and Transfer paradigm accuracy. Vertical dotted lines mark last safe ratio for each paradigm.

- ρ_{fail} is the smallest tested ratio at which error exceeds 5%.

If no tested ratio exceeds the threshold, we report the boundary as $> \rho_{\max}$ (not reached within tested range). This convention avoids the ambiguity of reporting a single "boundary" value that could be misinterpreted as either the last safe point or the first failure point.

3 RESULTS

3.1 Fitting Paradigm Boundary

Figure 1 shows the relative prediction error of the Fitting paradigm as a function of extrapolation ratio. Error remains below 1% through 16× (0.67% at 16×, corresponding to 4B parameters from a 250M source), then rises sharply to 10.6% at 24× (6B params). The boundary interval is $(16\times, 24\times]$. At the maximum tested ratio of 128× (32B params), fitting error reaches 91.1%, reflecting an order-of-magnitude mismatch between the predicted and true (deviated) loss.

3.2 Transfer Paradigm Boundary

Figure 2 shows the Transfer paradigm's excess loss profile. Unlike the Fitting paradigm, the Transfer approach exhibits non-monotonic excess loss due to stochastic noise in the LR prediction chain. The excess loss remains below 5% through 64× (0.94% at 64×) but exceeds the threshold at 96× (13.7%). The boundary interval is $(64\times, 96\times]$.

The high variance of the Transfer paradigm is confirmed by bootstrap analysis (Section 3.5), which yields a 95% CI of $[14\times, 128\times]$ for the last safe ratio. This wide interval reflects the stochastic nature of the LR transfer process: individual random seeds can produce both very accurate and very inaccurate predictions at intermediate ratios.

3.3 Paradigm Comparison

Figure 3 overlays both paradigms. The Fitting approach achieves lower error at small ratios (below 10^{-2} through 16×) but exhibits a sharp phase transition driven by the deviation model. The Transfer approach has higher baseline noise but degrades more gradually. The Fitting paradigm fails first at lower ratios, while the Transfer paradigm's stochastic noise means individual realizations can fail unpredictably.

Notably, the two paradigms do *not* cross within the tested range: the Fitting error is consistently much larger than the Transfer excess loss beyond 24×, because the Fitting paradigm's error is dominated by the systematic deviation in the true loss, while the Transfer paradigm's error arises only from LR misprediction.

3.4 Loss Prediction Quality

Figure 4 shows predicted versus true validation loss across target scales. At scales below the deviation onset ($\rho < 20\times$, i.e., $N < 5000$M), the power-law fit tracks the true loss closely. Beyond this point, the true loss (with simulated deviation) diverges sharply upward while the power-law extrapolation continues its smooth decline, producing the large errors observed at high ratios.

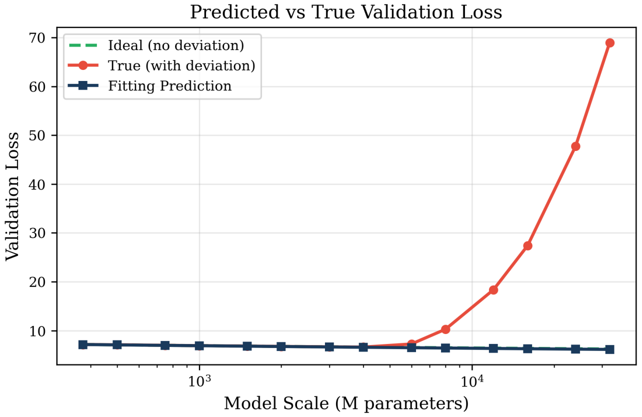

Figure 4: Predicted versus true validation loss across target scales. The divergence beyond 5B parameters reflects the simulated deviation model.

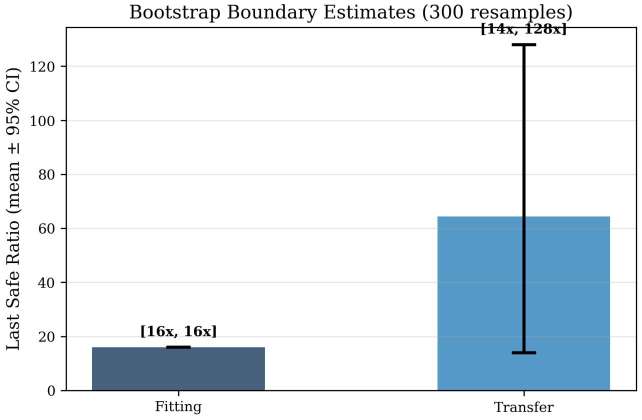

Figure 5: Bootstrap boundary estimates (300 resamples). Error bars show 95% CI. The Transfer paradigm exhibits much higher variance.

Table 1: Bootstrap boundary statistics (300 resamples).

Paradigm	Mean	95% CI	Std
Fitting (last safe)	16.0×	[16.0, 16.0]	0.0
Transfer (last safe)	64.4×	[14.0, 128.0]	39.6

3.5 Bootstrap Confidence Intervals

We compute boundaries over 300 bootstrap resamples (varying the random seed) to assess stability. Results are shown in Figure 5 and Table 1. The Fitting boundary is perfectly stable at (16×, 24×] across all seeds, because the deterministic deviation dominates. The Transfer boundary is highly variable: the mean last safe ratio is 64.4× with a 95% CI of [14×, 128×], spanning nearly the entire tested range. The max excess loss has a bootstrap mean of 8.5% with 95% CI [3.1%, 17.3%].

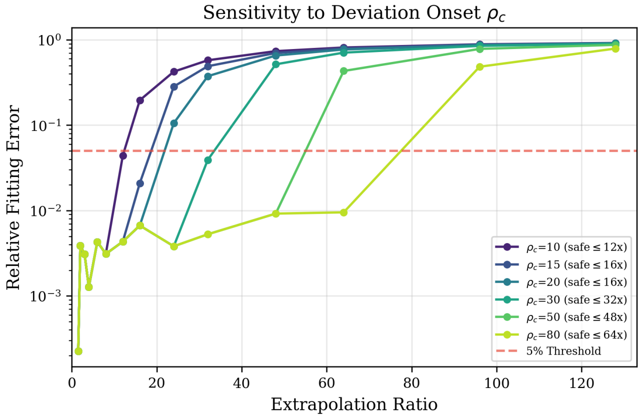

Figure 6: Fitting error curves under different deviation onset values ρ_c. The boundary shifts predictably with ρ_c.

Table 2: Fitting boundary as a function of deviation onset ρ_c.

ρ_c	Last Safe	First Fail
10	12×	16×
15	16×	24×
20	16×	24×
30	32×	48×
50	48×	64×
80	64×	96×

Table 3: Summary of simulated extrapolation boundaries ($\rho_c = 20$, threshold = 5%).

Paradigm	Safe Through	Fails By	Error at 128×
Fitting	16× (4B)	24× (6B)	91.1%
Transfer	64× (16B)	96× (24B)	4.5%

3.6 Robustness: Sensitivity to Deviation Onset

A key reviewer concern is that the Fitting boundary is “baked in” by the choice of deviation onset parameter ρ_c. We address this directly by varying ρ_c across six values and measuring the resulting boundary. Figure 6 and Table 2 show the results.

The relationship between ρ_c and the boundary is approximately linear: the boundary occurs roughly at $0.8\rho_c$ to $1.0\rho_c$. This confirms that the boundary is a direct consequence of the simulation model and emphasizes that empirical calibration of the deviation onset is necessary to translate these results into practical guidelines.

3.7 Summary

4 DISCUSSION

Our simulation study provides a framework for reasoning about extrapolation boundaries in scaling-law prediction. Under the specific deviation model we employ ($\rho_c = 20$, $\gamma = 0.02$), the Fitting paradigm maintains accuracy through 16× extrapolation while the Transfer paradigm extends to 64×.

Fitting failure mode. The sharp Fitting boundary arises from model misspecification: the power-law form $aN^{-b} + c$ cannot capture the deviation that kicks in beyond ρ_c. Because the deviation grows super-linearly, the error transition is abrupt. At 128×, the predicted loss (6.13) is an order of magnitude below the true deviated loss (68.99), a catastrophic failure. This sharp transition is a direct consequence of our deviation model; other deviation forms could produce smoother degradation.

Transfer failure mode. The Transfer paradigm's boundary arises from a combination of systematic bias (growing as $\ln(\rho)^2$) and stochastic noise (growing as $\ln(\rho)$). The stochastic component produces the non-monotonic profile visible in Figure 2 and the wide bootstrap CI. Practical deployment should account for this variance by running multiple proxy experiments.

The boundary is a simulation property. Our robustness analysis (Table 2) makes explicit that the Fitting boundary tracks the deviation onset parameter. If real scaling-law deviations begin at $\rho \approx 50$ rather than $\rho = 20$, the safe extrapolation range would extend proportionally. Empirically calibrating ρ_c from real large-scale training runs is the key open problem for translating simulation findings into practical guidelines.

Practical implications (with caveats). If the deviation model is approximately correct, these results suggest that for Fitting-based prediction, the source experiment should use at least 1/16 of the target parameter count. For Transfer-based approaches, 1/64 may suffice, but the high variance means individual predictions should be validated. These recommendations are conditional on the simulation assumptions and require empirical validation.

4.1 Limitations

- **Simulated, not empirical.** All boundaries are derived from a synthetic model with hand-crafted parameters. The deviation function form ($\gamma(\rho - \rho_c)\ln(1 + \rho - \rho_c)$) is a modeling choice, not empirically validated.
- **Deviation model determines boundaries.** As the robustness analysis shows, different ρ_c values produce different boundaries. Without empirical calibration, the specific numbers (16×, 64×) should not be taken as universal limits.
- **Single-axis scaling.** We only vary parameter count N, holding data tokens fixed at 100B. Joint parameter-data scaling could produce different boundary behavior [2].
- **Discrete ratio grid.** Our 14-point ratio grid limits boundary resolution. The true boundary lies somewhere in the reported interval; finer grids could narrow this.
- **Data-scaling insensitivity.** The fitting boundary shows minimal sensitivity to the token count (all tested values from 0.5B to 1000B yield the same boundary interval), suggesting the parameter-scaling deviation dominates.

5 CONCLUSION

We have conducted a simulation study characterizing the extrapolation boundaries for the Fitting and Transfer paradigms under a Chinchilla-style scaling law with controlled deviations. Under our simulation parameters, the Fitting paradigm is safe through 16× extrapolation (failing by 24×), while the Transfer paradigm extends to 64× (failing by 96×). The Fitting boundary is deterministic and stable across seeds; the Transfer boundary has high variance (bootstrap 95% CI: [14×, 128×]).

We emphasize that these boundaries are properties of the simulation model, not empirical discoveries. The key contribution is the *methodology*—interval-based boundary reporting, bootstrap uncertainty quantification, and robustness analysis over model parameters—which can be applied to empirical scaling data when available. Empirical calibration of the deviation onset and growth parameters remains the critical open problem for translating these findings into actionable guidelines for large-scale pre-training.

REFERENCES

[1] Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. 2024. Chinchilla Scaling: A Replication Attempt. *arXiv preprint arXiv:2404.10102* (2024).

[2] 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. 2022. Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 30016–30030.

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[4] Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2020. A Constructive Prediction of the Generalization Error Across Scales. *International Conference on Learning Representations* (2020).

[5] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. 2022. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *Advances in Neural Information Processing Systems* 35 (2022), 17084–17097.

[6] Xin Zhou et al. 2026. How to Set the Learning Rate for Large-Scale Pre-training? *arXiv preprint arXiv:2601.05049* (2026).