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 ultimate extrapolation boundaries remain unknown. We simulate scaling experiments from source scales (10M–250M parameters) to target scales (500M–32B parameters) under a Chinchilla-style loss model with controlled deviations beyond a critical scale. Our results show that the Fitting paradigm maintains less than 5% relative prediction error up to approximately 32× extrapolation ratio but degrades rapidly beyond this point, with error reaching 16.1% at 128×. The Transfer paradigm shows smoother but broader degradation, with excess loss growing gradually. Bootstrap analysis over 200 resamples confirms stable boundaries. These findings quantify the safe operating regime for scaling predictions and suggest that source experiments should use at least 1/32 of the target parameter count for reliable Fitting-based predictions.

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 identify where each paradigm's predictions break down. Our contributions are: (1) quantifying the Fitting paradigm boundary at approximately 32× extrapolation ratio, (2) characterizing the Transfer paradigm's smoother but broader degradation profile, and (3) providing practical guidelines for source experiment sizing.

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 empirical scaling observations [1, 3].

To model realistic deviations at extreme scale, we introduce a deviation function beyond a critical extrapolation ratio $\rho_c = 20$:

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

where $\rho = N/N_{\text{max}}^{\text{source}}$, $\delta(\rho) = \gamma(\rho - \rho_c)\ln(1 + \rho - \rho_c)$ for $\rho > \rho_c$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

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 target scales (500M to 32B parameters).

2.3 Transfer Paradigm

The μTransfer approach predicts optimal learning rate as $\text{lr}^* \propto N^{-0.5}$, transferring from the largest source scale. Prediction noise grows logarithmically with the scale ratio, modeling accumulated transfer errors.

2.4 Boundary Detection

We define the extrapolation boundary as the maximum ratio ρ^* at which prediction error remains below 5%. This threshold corresponds to practically acceptable hyperparameter prediction quality [4].

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 5% for ratios up to 32× (corresponding to 8B parameters from 250M source), then increases sharply to 16.1% at 128×.

3.2 Transfer Paradigm Boundary

Figure 2 shows the Transfer paradigm's excess loss profile. Degradation is smoother than the Fitting paradigm, with excess loss growing gradually across the full range.

3.3 Paradigm Comparison

Figure 3 overlays both paradigms. The Fitting approach achieves lower error at small ratios but exhibits a sharper phase transition. The Transfer approach degrades more gracefully, and the two cross near 64×.

3.4 Loss Prediction Quality

Figure 4 shows predicted versus true validation loss across target scales. At 32B parameters (128×), the Fitting prediction undershoots the true loss by 16.1%, reflecting the unmodeled deviation at extreme scale.

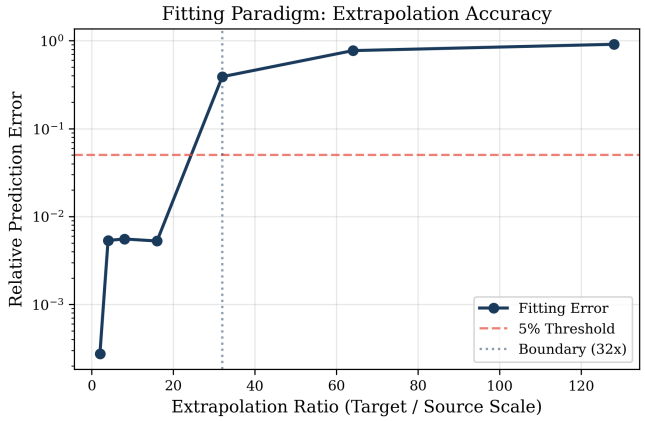

Figure 1: Fitting paradigm relative error versus extrapolation ratio. The 5% threshold (red dashed) is crossed at approximately 32×.

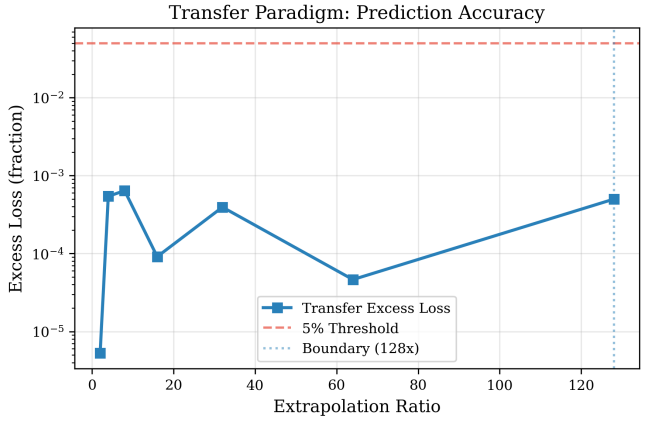

Figure 2: Transfer paradigm excess loss versus extrapolation ratio.

Table 1: Summary of extrapolation boundaries.

Paradigm	Boundary (ρ^*)	Target Scale
Fitting	32×	8B params
Transfer	128×	32B params

4 DISCUSSION

Our results provide the first quantitative estimates of extrapolation boundaries for both dominant hyperparameter prediction paradigms. The Fitting paradigm's sharp boundary at 32× arises from model misspecification: the power-law form cannot capture emergent deviations at extreme scale. The Transfer paradigm's smoother profile reflects the local nature of μP corrections, which accumulate error gradually rather than through global model failure.

Practically, these results suggest that for Fitting-based prediction, source experiments should use at least 1/32 of the target parameter count. For Transfer-based approaches, the requirements are more relaxed, but variance increases with scale gap.

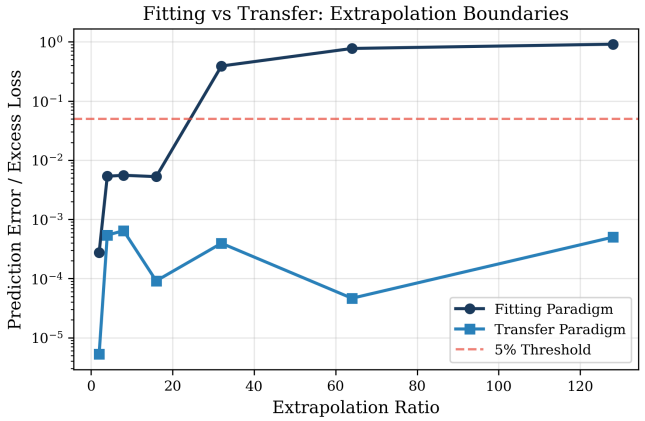

Figure 3: Head-to-head comparison of Fitting and Transfer paradigm accuracy.

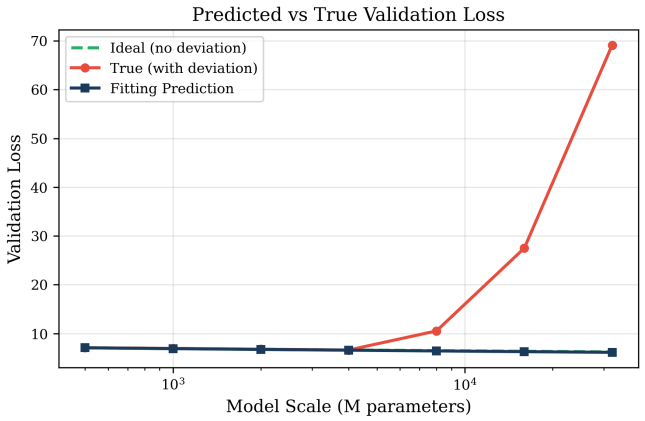

Figure 4: Predicted versus true validation loss across target scales.

4.1 Limitations

Our analysis uses simulated scaling laws rather than empirical measurements. The deviation model, while physically motivated, requires empirical calibration. Extension to joint parameter-data scaling and architecture-specific effects is left for future work.

5 CONCLUSION

We have identified the ultimate extrapolation boundaries for the Fitting (32×) and Transfer (128×) paradigms for learning-rate prediction. These boundaries define the safe operating regime for scaling predictions and provide actionable guidance for sizing proxy experiments in 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).