# Problems with Chinchilla Approach 2: Systematic Biases in IsoFLOP Parabola Fits

Eric Czech
Open Athena AI Foundation

**Abstract**

Chinchilla Approach 2 is arguably the most widely adopted method for fitting neural scaling laws, used by leading AI labs and academic groups to plan compute-optimal training configurations. The method fits parabolas to IsoFLOP loss curves and extracts scaling exponents through a sequence of simple polynomial and linear regressions. We show that this parabolic approximation introduces systematic biases in compute-optimal allocation estimates, even on noise-free synthetic data under ideal experimental conditions. Two independent bias sources are identified: surface asymmetry ($\alpha \neq \beta$), which shifts intercept estimates, and off-center sampling, which distorts intercepts or exponents depending on whether the offset is constant or varies with compute budget. These biases compound in practice and grow with sampling grid width. We show that exploiting the partially linear structure of the Chinchilla loss surface, by separating linear coefficients from nonlinear exponents, eliminates these biases entirely. Our realization of this approach, Variable Projection with Non-negative Least Squares (VPNLS), recovers all five surface parameters with machine precision across all conditions tested, while offering comparable data efficiency to Approach 2, high stability with full parametric inference, and no dependence on specialized IsoFLOP experiment designs.

## 1 Introduction

Chinchilla Approach 2 is arguably the most widely adopted method for fitting scaling laws in practice today. Introduced in the original Chinchilla paper [Hoffmann et al., 2022], it has since been used by leading AI labs including DeepMind [Hoffmann et al., 2022, Alabdulmohsin et al., 2023] (its creators), Meta [Grattafiori et al., 2024, Tay et al., 2025], DeepSeek [Bi et al., 2024], Microsoft [Jiang et al., 2025], Amazon [Haldar and Pinto, 2023], Waymo [Chen et al., 2025], and Arc Institute [Nguyen et al., 2024], among others. It is also a workhorse method for academic scaling law studies [Li et al., 2024, Nie et al., 2025, Wagh et al., 2024] and high-profile practitioner tutorials[1] from researchers like Andrej Karpathy.

The method's appeal lies in its stability and data efficiency relative to nonlinear optimization over all loss surface parameters. Rather than fitting all five parameters of the loss surface simultaneously, Approach 2 targets only the two scaling exponents, relying on second-order Taylor approximations that reduce each IsoFLOP curve to a simple parabola. This sacrifices recovery of the full loss surface but makes estimation far more stable and data-efficient, letting practitioners extract the most actionable quantities for compute allocation planning through a sequence of straightforward polynomial and linear fits, without ever touching a nonlinear optimizer.

Despite this broad adoption, the sensitivity of the method's core approximations and its behavior on loss surfaces that are less symmetric than the original Chinchilla form (where parameter and

[1] https://x.com/karpathy/status/2009037707918626874

token scaling exponents are roughly equal) have not, to our knowledge, been studied in detail. Here we revisit the basics of how to apply a simple model like Chinchilla with high precision and stability, to validation loss alone, before considering more advanced extensions. We investigate through noise-free synthetic simulations that isolate systematic biases inherent to the method itself by eliminating all sources of statistical noise.

We show how these biases affect downstream decisions like dataset size selection for final training runs at large compute budgets. We show how extrapolation errors trace back to suboptimal IsoFLOP experiment design, and that pathologies in these designs can be observed in real, high-profile scaling law studies even if they are difficult to quantify precisely. Finally, we propose an alternative fitting method that is simple, stable, and free of these biases while building on the same intuitive computational shortcut: optimizing exponential terms separately from linear terms. We call this approach Variable Projection with Non-negative Least Squares (VPNLS).

This investigation is also motivated by a broader landscape of *analytical* extensions to the Chinchilla loss surface. A growing body of work adds or modifies terms in the original functional form to account for additional training configuration choices such as data repetition [Muennighoff et al., 2023, Goyal et al., 2024], overfitting [Yang et al., 2024], precision [Kumar et al., 2024], MoE sparsity [others, 2025a], data quality [others, 2025c], data mixtures [others, 2025b,d, Goyal et al., 2024], non-embedding parameters [Sardana and Frankle, 2024], and downstream task performance [others, 2024], to name a few. These extensions prescribe explicit functional forms rather than inferring scaling law structure automatically, and they build directly on the Chinchilla model as a foundation. A fitting method that recovers the base surface with higher precision may therefore offer a stronger starting point for these richer settings as well.

## 2 Preliminaries

Placeholder for the preliminaries section covering the loss surface, notation, and fitting methods (Approach 2 and Approach 3).

## 3 The Happy Path: Symmetric Surfaces

Placeholder for the symmetric surface baseline analysis.

## 4 Asymmetric Surfaces: Intercept and Extrapolation Errors

Placeholder for the asymmetric surface analysis showing intercept and extrapolation errors.

## 5 Off-Center Sampling: Exponent and Extrapolation Errors

Placeholder for off-center sampling analysis.

## 6 IsoFLOP Curves in the Wild

Placeholder for the analysis of published IsoFLOP curves.

## 7 Robust Fits: Unbiased Estimation with Linear Separation

Placeholder for the VPNLS method description and comparison.

## 8 Conclusion

Placeholder for the conclusion.

## References

Ibrahim M. Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Getting ViT in shape: Scaling laws for compute-optimal model design. In *NeurIPS*, 2023.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, et al. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Yilun Chen et al. Scaling laws of motion forecasting and planning – technical report. *arXiv preprint arXiv:2506.08228*, 2025.

Sachin Goyal et al. Scaling laws for data filtering – data curation cannot be compute agnostic. In *CVPR*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Siddhant Haldar and Lerrel Pinto. Scaling laws for imitation learning in single-agent games. *Transactions on Machine Learning Research*, 2023.

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. *arXiv preprint arXiv:2203.15556*, 2022.

Ethan Jiang et al. Exploring scaling laws for EHR foundation models. *arXiv preprint arXiv:2505.22964*, 2025.

Tanishq Kumar et al. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024.

Zhengyang Li et al. Scaling laws for diffusion transformers. *arXiv preprint arXiv:2410.08184*, 2024.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, and Thomas Wolf. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, et al. Sequence modeling and design from molecular to genome scale with Evo. *bioRxiv preprint 2024.02.27.582234*, 2024.

Shen Nie et al. Scaling behavior of discrete diffusion language models. *arXiv preprint arXiv:2512.10858*, 2025.

others. Establishing task scaling laws via compute-efficient model ladders. *arXiv preprint arXiv:2412.04403*, 2024.

others. Towards greater leverage: Scaling laws for efficient mixture-of-experts language models. *arXiv preprint arXiv:2507.17702*, 2025a.

others. Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*, 2025b.

others. Scaling laws revisited: Modeling the role of data quality in language model pretraining. *arXiv preprint arXiv:2510.03313*, 2025c.

others. Scaling laws are redundancy laws. *arXiv preprint arXiv:2509.20721*, 2025d.

Nikhil Sardana and Jonathan Frankle. Reconciling Kaplan and Chinchilla scaling laws. *Transactions on Machine Learning Research*, 2024.

Yi Tay et al. Training compute-optimal transformer encoder models. In *EMNLP*, 2025.

Neeraj Wagh et al. Scaling laws for compute optimal biosignal transformers. 2024.

Xinghua Yang et al. MuPT: A generative symbolic music pretrained transformer. *arXiv preprint arXiv:2404.06393*, 2024.