A Model for Scaling Laws of General Intelligence

Ari Brill*[1]

[1]Principles of Intelligence (PIBBSS)

July 1, 2025; Revised September 16, 2025

Abstract

Deep neural networks trained on vast datasets achieve strong performance on diverse tasks. These models exhibit empirical neural scaling laws, under which prediction error steadily improves with larger model scale. The cause of improvement is unclear, as strong general performance could result from acquiring general-purpose capabilities or specialized knowledge across many domains. To address this question theoretically, we study model scaling laws for a capacity-constrained predictor that optimally instantiates task-specific or general-purpose latent circuits. For a data distribution consisting of power-law-distributed tasks, each represented by a low-dimensional data manifold, general capabilities emerge abruptly at a threshold model scale and decline in relative importance thereafter. Data diversity and model expressivity increase general capabilities in distinct ways.

1 Introduction

A longstanding aim of artificial intelligence research has been to create artificial general intelligence (AGI), an artificial system capable of strong or superhuman performance on a broad range of novel tasks with no restriction as to domain, context, or objective (Legg and Hutter, 2007; Goertzel, 2014; Chollet, 2019; Bubeck et al., 2023). Recently, a successful approach for creating generally capable AI systems has been pretraining large language models (LLMs) on vast and diverse data corpora (Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023; Gemini Team et al., 2023). However, the world's complexity ensures that no pretraining corpus can include all situations a generally capable AI system would need to handle. A pretrained AI system's benefits and risks may depend crucially on whether it achieves strong performance primarily by accumulating domain-specific knowledge for each pretraining task, or by learning a task-agnostic core of general intelligence.

Powerful AI agents with general-purpose capabilities could act effectively when given any task, situation, or goal, unlocking extraordinary economic value, but also posing inherent dangers. For example, the general-purpose capabilities of situational awareness, reasoning, and self-preservation could lead a strong AI system with misaligned goals to deceive its human supervisors (Carlsmith, 2023; Hubinger et al., 2019; Greenblatt et al., 2024; Hubinger et al., 2024). Understanding when general capabilities arise is therefore of great importance in AI safety.

*Email: aryeh.brill@gmail.com

Neural scaling laws are essential for understanding LLMs' success. As the size of the model or training dataset increases, prediction error steadily decreases, in accordance with empirical power laws (Hestness et al., 2017; Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022). Multiple theoretical models have been proposed to explain these observed scaling laws.

One theoretical viewpoint treats a neural network as a nonparametric function approximator that, when given access to additional degrees of freedom (DOF), resolves a low-dimensional data manifold at increasingly fine resolution, yielding power-law scaling with an exponent inversely proportional to the intrinsic data dimension (Sharma and Kaplan, 2022; Bahri et al., 2024). The DOF can be interpreted either as model features or training data points, giving the same exponents for model and data scaling. Furthermore, Bahri et al. (2024) showed that manifold approximation also can be understood in terms of kernel regression with power-law random features (Maloney et al., 2022; Bordelon et al., 2024; Paquette et al., 2024).

However, it is unclear how well the wide variety of tasks a generally capable AI system must perform are described by a data distribution consisting of a single manifold. An alternative approach instead models the data distribution as a set of tasks with a power-law frequency distribution (Feldman, 2020; Feldman and Zhang, 2020; Hutter, 2021; Michaud et al., 2023; Cabannes et al., 2023; Fonseca et al., 2024; Brill, 2024; Liu et al., 2025; Pan et al., 2025). A capacity-constrained model optimally memorizes the correct behavior for the most important tasks in order, yielding a power-law loss curve recapitulating the task distribution. One way to unify power-law-distributed data with manifold approximation scaling was proposed by Brill (2024, 2025), who considered a data model based on percolation on a hypercubic lattice. In this model, the resulting clusters have both a power-law size distribution and a low-dimensional fractal representation in data space.

Scaling laws imply that an AI system's pretraining loss decreases smoothly with increasing model scale. However, this summary metric may combine multiple distinct causes. A priori, both task-specific features and general-purpose capabilities could reduce the loss equally well. Furthermore, predicting an AI system's capabilities from its pretraining loss at a particular scale is not trivial. Models may exhibit apparently emergent abilities, such as in-context learning or mathematical reasoning, that arise discontinuously after reaching a particular model scale (Brown et al., 2020; Ganguli et al., 2022; Srivastava et al., 2022; Wei et al., 2022); but see (Schaeffer et al., 2023). Furthermore, the relation between scale and capabilities remains important for understanding post-trained systems built on a pretrained model. Post-training may primarily elicit a pretrained model's latent capabilities rather than inducing new ones, as suggested both because most computation and data is used in pretraining and by empirical evidence (Zhou et al., 2023; Jain et al., 2023).

Given that an AI system has a capability, it's natural to search for the mechanisms that implement it. A major goal of mechanistic interpretability research is to decompose a neural network's internal operations into circuits, each implementing an interpretable algorithm involving interactions among latent features (Olah et al., 2020; Wang et al., 2022; Conmy et al., 2023; Dunefsky et al., 2024; Marks et al., 2024; Braun et al., 2025; Ameisen et al., 2025; Lindsey et al., 2025). From a theoretical perspective, Vaintrob (2025) proposed a toy model of neural networks as a collection of circuits. In this picture, each potentially learnable circuit is parameterized by a measure of its size or complexity, and a measure of its independent contribution to the overall accuracy.

Multiple factors might arguably push an AI system to learn general capabilities rather than specialized ones (cf. Hubinger et al., 2019). Intuitively, an AI system faced with diverse tasks might more efficiently learn one expensive general-purpose circuit over numerous special-purpose ones. Another factor with a less obvious effect is the model's expressivity, or ability to fit complex functions. A priori, increased expressivity could make both specialized and general capabilities more efficient, so its relative effect is unclear. A quantitative description is needed to say more about how data diversity and model expressivity affect general capabilities.

In this work, we investigate a mathematical model of an abstract capacity-constrained AI system that optimally balances specialized and general capabilities. To minimize prediction error on a data distribution consisting of one or more tasks supported on distinct low-dimensional data manifolds, the AI system draws from a latent population of circuits that provide either task-specific features or general capabilities. We compute scaling laws with respect to model size for loss and for several measures of general capabilities. A single task yields limited general capabilities. Power-law-distributed tasks yield nontrivial general capabilities that emerge abruptly at a threshold model scale and then decline in relative importance compared to task-specific features. Data diversity and model expressivity enhance general capabilities in distinct ways.

2 Model

2.1 Setup

We consider a stylized AI system with N units of model capacity, with model capacity being some nonnegative, additive scalar measure of the AI system's size[1]. We assume the AI system is a machine learning model trained to convergence in a regression setting. The AI system is capacity-constrained, with access to unlimited amounts of computation and data for training, and fixed, sufficient amounts of computation and contextual data for inference.

We model the data distribution as a set of one or more distinct data manifolds of equal dimension. Each manifold corresponds to a separate task, defined by a continuous target function supported on that manifold. We assume that the manifolds may have different sizes, determining the relative importance of each task. We index tasks by their rank $k \in \mathbb{N}$ in order of descending manifold size. Each task's baseline loss is proportional to its corresponding manifold's size and can be written as a function of k. We consider two data distribution models:

1. A single task corresponding to a D-dimensional manifold.
2. A power-law task distribution parameterized by rank-frequency distribution $L_k \propto k^{-(1+\alpha)}$, where $\alpha > 0$, with each task corresponding to a D-dimensional manifold.

For each task, we assume the AI system nonparametrically approximates the required function using n DOF or effective features. That task's loss contribution then scales as $n^{-c/D}$ for mean squared error or cross-entropy loss (Sharma and Kaplan, 2022; Bahri et al., 2024). The constant c measures model expressivity, with $c \geq 2$ for a piecewise constant function approximator and $c \geq 4$ for a piecewise linear function approximator. We assume any target function is a generic Lipschitz continuous function, so that the inequality for c saturates as an equality.

2.2 Circuit distribution

To convert model capacity into DOF, the AI system internally implements one or more circuits. We assume that there exists a latent population of potential circuits, from which the AI system

[1] For example, one might consider model capacity in terms of description length or neural network parameters.

instantiates circuits optimally[2]. We characterize each circuit by a model capacity cost, N_{circ}, and the number of expected DOF it provides for each task, $n_{\text{circ}}(k)$. We define a circuit's efficiency for task k to be the ratio $\epsilon_{\text{circ}}(k) \equiv n_{\text{circ}}(k)/N_{\text{circ}}$. The latent circuit population is assumed to be unbounded, so that the model is not constrained by available circuits to learn. If many circuits contribute to each task, it is reasonable to approximate n as a continuous quantity, and we generally do so throughout this work. When on occasion it matters that DOF are discrete, we require that $n_{\text{circ}} \geq 1$ for all circuits. We assume for simplicity that $N_{\text{circ}} \ll N$ for all relevant circuits.

We consider two classes of potential circuits. First, a *feature circuit* computes a feature relevant for approximating one task's target function. In this usage, a "feature" might be, for example, a basis function, cluster prototype, or memorized data point, and need not be human-interpretable. All features are modeled as identically generic, with the same capacity cost N_{F} and efficiency ϵ_{F}. The number of DOF provided for task k' by a feature circuit specialized for task k is then $n_{\text{F}}(k') = \epsilon_F N_{\text{F}} \mathbf{1}_k$. From now on, we choose units such that $N_{\text{F}} = \epsilon_{\text{F}} = 1$ and write n_k to denote the total number of DOF provided by feature circuits for task k.

Next, we consider a class of *general circuits*. These circuits implement capabilities of use on any logically consistent and physically plausible task, even if all contingent facts about the world were different. For example, general circuits might represent[3]

- logical inference and probabilistic reasoning;
- in-context learning (Brown et al., 2020);
- mesa-optimization (Hubinger et al., 2019);
- situational awareness (Carlsmith, 2023; Berglund et al., 2023; Laine et al., 2024), including self-knowledge (Betley et al., 2025);
- and/or core knowledge priors such as objectness and elementary physics; agentness and goal-directedness; natural numbers and elementary arithmetic; and elementary geometry and topology (Spelke and Kinzler, 2007; Chollet, 2019).

It seems likely that general capabilities could vary widely in complexity and utility. Since there is no reason to assume a preferred efficiency scale for general circuits, we approximate the population's marginal efficiency as a power law, $\mathrm{d}\epsilon_{\text{G}} = \mathrm{d}n_{\text{G}}/\mathrm{d}N_{\text{G}} = \epsilon_0 n_{\text{G}}^{-\gamma}$. The efficiency prefactor $\epsilon_0 > 0$ and exponent $\gamma > 0$ are free parameters. Since $\epsilon_{\text{F}} = 1$ is fixed, ϵ_0 determines the efficiency of general circuits compared to feature circuits. We assume that a general circuit's value is (in expectation) equal for all tasks, so that n_{G} is independent of k. This gives

$$N_{\text{G}} = \int \epsilon_0^{-1} n_{\text{G}}^{\gamma} \, \mathrm{d}n_{\text{G}} = m n_{\text{G}}^{1+\gamma}, \tag{1}$$

where $m \equiv (\epsilon_0(1+\gamma))^{-1}$.

In Fig. 1, we illustrate the assumed latent circuit population graphically as an "efficiency spectrum", inspired by Vaintrob (2025). In the graph, n_{circ} is plotted against N_{circ}, so that the slope

[2] The latent population could be thought of as the elements of the platonic set of all relevant circuits, or more concretely, as the set of circuits learned by a model with capacity M in the limit $M \to \infty$.

[3] Circuits representing syntax (Pan et al., 2025) or modality-specific surface features (Brill, 2025) might also be considered general. Our definition excludes them, because they would not be useful under a different contingent input representation. That said, these could be interpreted as general capabilities with no change to the formal model.

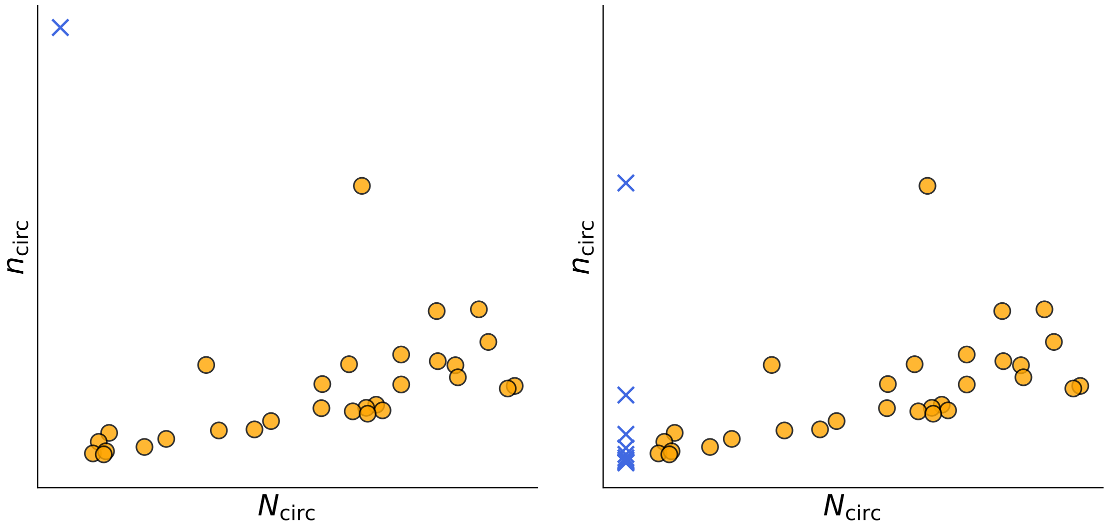

Figure 1: Cartoon diagrams representing efficiency spectra of the latent circuit population, with slope $\epsilon_{\text{circ}} = n_{\text{circ}}/N_{\text{circ}}$. Blue crosses show feature circuits and orange circles show general circuits. Left: Efficiency spectrum for a single task. All feature circuits overlap. Right: Efficiency spectrum for a power-law task distribution, with $\alpha = 1$. Feature circuits overlap within each task, with n_{circ} proportional to that task's frequency. For both sides, the general circuits are drawn from an efficiency distribution with $\epsilon_0 = 0.1$ and $\gamma = 0.5$. They appear identical on both sides, as they have the same value for all tasks.

between a point and the origin equals the corresponding circuit's efficiency. As we will see below, DOF contributed by different circuits may combine nonlinearly to reduce the loss. If these nonlinear interactions were neglected, the optimal capacity-constrained model would implement circuits in descending efficiency order. Therefore, the efficiency spectrum can be interpreted as a rough guideline as to each circuit's importance.

We want to measure to what extent the AI system uses general capabilities. We define two metrics measuring distinct but related quantities. First, to measure the fraction of model capacity devoted to general circuits, we define the *capacity fraction*,

$$\frac{N_{\text{G}}}{N} = \frac{N_{\text{G}}}{\sum_k n_k + N_{\text{G}}}. \tag{2}$$

Second, to measure the expected fraction of DOF attributable to general circuits, we define the *DOF fraction*,

$$\left\langle \frac{n_{\text{G}}}{n} \right\rangle = \mathbb{E}_{\text{task}} \left[\frac{n_{\text{G}}}{n_{\text{F}}(\text{task}) + n_{\text{G}}} \right]. \tag{3}$$

For the power law setting, the DOF fraction is computed as

$$\left\langle \frac{n_{\text{G}}}{n} \right\rangle = \frac{1}{\sum_k^\infty k^{-(1+\alpha)}} \sum_k^\infty \frac{n_{\text{G}}}{n_k + n_{\text{G}}} k^{-(1+\alpha)} = \frac{1}{\zeta(1+\alpha)} \sum_k^\infty \frac{n_{\text{G}}}{n_k + n_{\text{G}}} k^{-(1+\alpha)}, \tag{4}$$

where $\zeta(s)$ is the Riemann zeta function.

2.3 Single task

For a single-task data distribution, the loss is (up to an overall constant factor),

$$L = \max(1,\ n_{\mathrm{F}} + n_{\mathrm{G}})^{-c/D}, \tag{5}$$

where the max function sets the loss to that of a random predictor if both n_{F} and n_{G} are 0. The capacity constraint is,

$$N = n_{\mathrm{F}} + m n_{\mathrm{G}}^{1+\gamma}. \tag{6}$$

Intuitively, if $\epsilon_0 < 1$, feature circuits dominate, giving an optimal value of $n_{\mathrm{G}} = 0$. If $\epsilon_0 > 1$, general circuits dominate until their marginal efficiency decreases enough that feature circuits take over. To see this quantitatively, we examine the only three possible cases. First, only feature circuits could be learned, so that $n_{\mathrm{G}} = 0$, giving $n_{\mathrm{F}} = N$ and $L = N^{-c/D}$. Next, only general circuits could be learned, so that $n_{\mathrm{F}} = 0$, giving $n_{\mathrm{G}} = (N/m)^{1/(1+\gamma)}$ and $L = (N/m)^{-c/D/(1+\gamma)}$. Finally, both feature circuits and general circuits could be learned. To solve for n_{F} and n_{G}, we use the method of Lagrange multipliers. The Lagrangian is

$$\mathcal{L}_1 = (n_{\mathrm{F}} + n_{\mathrm{G}})^{-c/D} - \lambda(n_{\mathrm{F}} + m n_{\mathrm{G}}^{1+\gamma} - N). \tag{7}$$

In Appendix A, we solve for the optimal values of n_{G} and n_{F}. The solution is

$$n_{\mathrm{G}} = [m(1+\gamma)]^{-1/\gamma} = \epsilon_0^{1/\gamma}, \tag{8}$$

$$n_{\mathrm{F}} = N - (1+\gamma)^{-1}\epsilon_0^{1/\gamma}. \tag{9}$$

It follows that

$$L = \left(N + \frac{\gamma}{1+\gamma}\epsilon_0^{1/\gamma}\right)^{-c/D}. \tag{10}$$

Eq. 8, along with the constraints $1 \leq n_{\mathrm{G}} \leq N$, implies that feature circuits and general circuits can only coexist when $\epsilon_0 \geq 1$ and $N \gtrsim (1+\gamma) m \epsilon_0^{(1+\gamma)/\gamma}$. Of allowed cases, whichever one minimizes the loss determines the overall behavior. If $\epsilon_0 < 1$, then $\langle n_{\mathrm{G}}/n \rangle = 0$ exactly. If $\epsilon_0 \geq 1$, then for large N, $n_{\mathrm{G}} = \epsilon_0^{1/\gamma}$ is a constant and $\langle n_{\mathrm{G}}/n \rangle \to 0$.

2.4 Power-law task distribution

Next, we study a power-law task distribution. The loss is

$$L = \sum_{k=1}^{\infty} L_k = \sum_{k=1}^{\infty} \max(1,\ n_k + n_{\mathrm{G}})^{-c/D} \frac{1}{\zeta(1+\alpha)} k^{-(1+\alpha)}, \tag{11}$$

and the capacity constraint is

$$N = \sum_{k=1}^{\infty} n_k + m n_{\mathrm{G}}^{1+\gamma}. \tag{12}$$

Since $k^{-(\alpha+1)}$ is monotonically decreasing, the optimal allocation of the n_k must be monotonically non-increasing. Since the n_k are nonnegative and discrete, there exists a break at some rank k_{br} such that $n_k > 0$ for all $k < k_{\mathrm{br}}$ and $n_k = 0$ for all $k > k_{\mathrm{br}}$. At equilibrium, the marginal loss reduction due to features learned for any two tasks of rank $k_i,\ k_j < k_{\mathrm{br}}$ must be equal, yielding

$$\frac{n_{k_i} + n_{\mathrm{G}}}{n_{k_j} + n_{\mathrm{G}}} = \left(\frac{k_i}{k_j}\right)^{-\frac{1+\alpha}{1+c/D}}. \tag{13}$$

Eq. 13 is solved by the ansatz, introducing variables $a > 0$ and $b < 1$,

$$n_k = \begin{cases} ak^{b-1} - n_{\mathrm{G}}, & k < k_{\mathrm{br}} \\ 0, & k \geq k_{\mathrm{br}}. \end{cases} \tag{14}$$

It follows from Eq. 13 and Eq. 14 that

$$b = 1 - \frac{1+\alpha}{1+c/D} = \frac{c/D - \alpha}{1+c/D}. \tag{15}$$

To find the optimal values of a, k_{br} and n_{G}, we again use the method of Lagrange multipliers. The Lagrangian is

$$\begin{aligned}\mathcal{L}(a, k_{\mathrm{br}}, n_{\mathrm{G}}, \lambda) = &\sum_{k=1}^{k_{\mathrm{br}}} \max\left(1,\ ak^{b-1}\right)^{-c/D} \frac{1}{\zeta(1+\alpha)} k^{-(1+\alpha)} + \sum_{k=k_{\mathrm{br}}}^{\infty} \max(1,\ n_{\mathrm{G}})^{-c/D} \frac{1}{\zeta(1+\alpha)} k^{-(1+\alpha)} \\ &- \lambda \left(\sum_{k=1}^{k_{\mathrm{br}}} \left(ak^{b-1} - n_{\mathrm{G}}\right) + mn_{\mathrm{G}}^{1+\gamma} - N \right).\end{aligned} \tag{16}$$

As before, there are three possible cases: the optimal model instantiates only feature circuits (Case F); only general circuits (Case G); or both (Case FG). As before, the case that minimizes the loss determines the overall behavior, and the overall loss is given by $\min(L_{\mathrm{F}}, L_{\mathrm{G}}, L_{\mathrm{FG}})$.

2.4.1 Case F: feature circuits only

This case is equivalent to the model scaling setting previously studied by Brill (2024). However, the derivation presented here employs fewer approximations than Brill (2024), yielding a more exact loss curve prefactor. With feature circuits only, the Lagrangian is

$$\mathcal{L}_{\mathrm{F}}(a, k_{\mathrm{br}}, \lambda) = \sum_{k=1}^{k_{\mathrm{br}}} \left(ak^{b-1}\right)^{-c/D} \frac{1}{\zeta(1+\alpha)} k^{-(1+\alpha)} + \sum_{k=k_{\mathrm{br}}}^{\infty} \frac{1}{\zeta(1+\alpha)} k^{-(1+\alpha)} - \lambda \left(\sum_{k=1}^{k_{\mathrm{br}}} ak^{b-1} - N \right). \tag{17}$$

In Appendix B, we solve for the optimal a and k_{br} using Lagrange multipliers. We obtain

$$a = Ck_{\mathrm{br}}^{1-b}, \tag{18}$$

where $C = (1 + c/D)^{1/(c/D)}$, and

$$\frac{1}{b} k_{\mathrm{br}} \left(1 - k_{\mathrm{br}}^{-b}\right) = \frac{N}{C}. \tag{19}$$

Eq. 19 has no closed-form solution. We obtain the approximations for k_{br} in limiting cases and for large N, where $W(x)$ denotes the Lambert W function,

$$k_{\mathrm{br}} \approx \begin{cases} N/C & 0 < b < 1,\ |b| \sim 1, \\ W(N/C)^{-1}(N/C) & |b| \ll 1, \\ (|b|N/C)^{1/(1+|b|)} & b < 0,\ |b| \gg 1. \end{cases} \tag{20}$$

The expression for the loss is then,

$$L_{\mathrm{F}} = \left[1 + \left(\frac{\alpha}{1 + c/D}\right) \frac{N/C}{k_{\mathrm{br}}}\right] k_{\mathrm{br}}^{-\alpha} \tag{21}$$

$$\propto \begin{cases} N^{-\alpha} & 0 < b < 1,\ |b| \sim 1, \\ W(N/C)^{1+\alpha} N^{-\alpha} & |b| \ll 1, \\ N^{-c/D} & b < 0,\ |b| \gg 1, \end{cases} \tag{22}$$

The full expressions including prefactors are given in Eq. 45.

2.4.2 Case G: general circuits only

With only general circuits, the Lagrangian simplifies to

$$\begin{aligned} \mathcal{L}_{\mathrm{G}}(n_{\mathrm{G}}, \lambda) &= \sum_{k=1}^{\infty} n_{\mathrm{G}}^{-c/D} \frac{1}{\zeta(1+\alpha)} k^{-(1+\alpha)} - \lambda \left(m n_{\mathrm{G}}^{1+\gamma} - N\right) \\ &= n_{\mathrm{G}}^{-c/D} - \lambda \left(m n_{\mathrm{G}}^{1+\gamma} - N\right). \end{aligned} \tag{23}$$

It follows that $n_{\mathrm{G}} = (N/m)^{1/(1+\gamma)}$ and $L_{\mathrm{G}} = (N/m)^{-c/D/(1+\gamma)}$, as for a single task.

2.4.3 Case FG: both feature and general circuits

If both feature circuits and general circuits are instantiated, the Lagrangian simplifies to

$$\mathcal{L}_{\mathrm{FG}}(a, k_{\mathrm{br}}, n_{\mathrm{G}}, \lambda) \approx \left(\alpha a^{-c/D} - \lambda a\right) \frac{k_{\mathrm{br}}^{b} - 1}{b} + n_{\mathrm{G}}^{-c/D} k_{\mathrm{br}}^{-\alpha} - \lambda \left(m n_{\mathrm{G}}^{1+\gamma} - (k_{\mathrm{br}} - 1)\, n_{\mathrm{G}} - N\right). \tag{24}$$

In Appendix C, we solve for the optimal a, k_{br}, and n_{G}. Again, there is no closed-form solution, and we obtain

$$a = n_{\mathrm{G}} k_{\mathrm{br}}^{1-b}, \tag{25}$$

$$k_{\mathrm{br}} = \frac{\alpha}{1+\alpha}\left(1 + m(1+\gamma)n_{\mathrm{G}}^{\gamma}\right), \tag{26}$$

$$N = n_{\mathrm{G}}\left[\frac{k_{\mathrm{br}}\left(1 - k_{\mathrm{br}}^{-b}\right)}{b} - (k_{\mathrm{br}} - 1) + m n_{\mathrm{G}}^{\gamma}\right], \tag{27}$$

We use the same approximations employed in Sec. 2.4.1 and consider large N to obtain approximate expressions for n_{G}. We obtain

$$n_{\mathrm{G}} \propto \begin{cases} (N/m)^{1/(1+\gamma)} & 0 < b < 1,\ |b| \sim 1, \\ W(\mathrm{const} \cdot N) \cdot (N/m)^{1/(1+\gamma)} & |b| \ll 1, \\ (N/m)^{1/(1+\gamma-b\gamma)} & b < 0,\ |b| \gg 1. \end{cases} \tag{28}$$

The full expressions including prefactors are given in Appendix C, in Eq. 59, Eq. 61, and Eq. 64. The resulting expression for the loss is

$$L_{\mathrm{FG}} \approx \alpha\left(\frac{1-b}{R}\right)^{1+\alpha}\left(\gamma + \frac{N}{m n_{\mathrm{G}}^{1+\gamma}}\right)\left(m n_{\mathrm{G}}^{1+\gamma}\right)^{-\alpha} n_{\mathrm{G}}^{-(c/D-\alpha)} \tag{29}$$

$$\propto \begin{cases} (N/m)^{-\frac{c/D-\alpha}{1+\gamma}} N^{-\alpha} & 0 < b < 1,\ |b| \sim 1, \\ (\gamma + W(\mathrm{const} \cdot N))\, W(\mathrm{const} \cdot N)^{\frac{c/D-\alpha}{1+\gamma}+\alpha} \cdot (N/m)^{-\frac{c/D-\alpha}{1+\gamma}} N^{-\alpha} & |b| \ll 1, \\ N^{-c/D} & b < 0,\ |b| \gg 1. \end{cases} \tag{30}$$

The full expressions including prefactors are given in Appendix C, in Eq. 65, Eq. 66, and Eq. 67.

3 Experiments

We performed numerical experiments to gain richer insight into the model's properties and to verify our analytical approximations. The experiments examined the setting of power-law-distributed data defined by the loss function given by Eq. 11 and the capacity constraint given by Eq. 12. The loss function was numerically optimized with gradient descent using PyTorch (Paszke, 2019). For all experiments, the parameters were a vector of n_k values, truncated to a length of 5000, and a scalar value of n_{G}, with the values of c/D, α, ϵ_0, γ, and N specified as hyperparameters. These hyperparameters were varied across the experiments so that scaling curves in N could be studied for a range of qualitatively different b values.

To enforce the capacity constraint given by Eq. 12, an auxiliary term was appended to the loss,

$$L_{\mathrm{aux}} = \delta|L|\left|\frac{\hat{N} - N}{N}\right|, \tag{31}$$

where δ is a hyperparameter setting the constraint strength, $|L|$ is the loss magnitude at the current step (computed with a stop gradient), and $\hat{N}$ is the value of N computed from the parameters. For all experiments, $\delta = 5.0$ was used.

To enforce the bounds $n_{\mathrm{G}} > 0$ and $n_k > 0$ for all k during optimization, each parameter was wrapped with a soft maximum constraint of the form $\mathrm{Softplus}(n)$. Similarly, to clamp $n_k + n_{\mathrm{G}}$ to a minimum of 1, a soft constraint of the form $\mathrm{Softplus}(n-1)+1$ was applied. When reporting final metrics, these soft constraints were replaced with hard maximum constraints.

For all experiments, training was performed for 2×10^4 steps using the AdamW optimizer with a learning rate of 1.0, no weight decay, and a cosine annealing learning rate decay schedule with $T_{\max}$ equal to the number of steps (Kingma and Ba, 2014; Loshchilov and Hutter, 2017, 2016). If not otherwise mentioned, PyTorch default hyperparameters were used.

Fig. 2 and Fig. 3 show the experimental results. The top panels of Fig. 2 show model scaling curves for n_{G} and L, overlaid with the corresponding analytical approximations for appropriate limits. The analytical curves agree well with the numerical results. The bottom panels of Fig. 2 show how the capacity fraction and DOF fraction scale with N. Fig. 3 shows how these metrics change when keeping N fixed and instead varying either c/D or α.

4 Discussion

Although the presented model is quite simple, it predicts a number of interesting qualitative properties that may have parallels in realistic AI systems. We study the model scaling laws that result from allocating model capacity among neural circuits to optimally predict power-law-distributed data. Scaling laws for power-law-distributed atomic tasks were initially studied by Hutter (2021) and extended by Michaud et al. (2023). Motivated by percolation theory, Brill (2024) proposed the data model that we consider in this work, in which tasks are represented by low-dimensional data manifolds. Pan et al. (2025) further incorporated a universal syntax component when studying data scaling laws for power-law-distributed knowledge clusters, but omitted syntax when studying model scaling. The model scaling laws predicted by these works are similar to those presented in Sec. 2.4.1 for feature circuits only. The model presented here extends these prior works by making novel predictions about the scaling laws of general capabilities. The key differentiating assumption is that a capacity-constrained AI system can learn not only task-specific features but also general circuits that have a heavy-tailed marginal efficiency distribution.

4.1 General capabilities require training task diversity

For a data distribution consisting of a single task, the amount of general capabilities learned has a constant upper bound, which may be none (Sec. 2.3). In particular, for any general capabilities to be learned, the most efficient general circuits must actually be more efficient than feature circuits, with $\epsilon_0 > \epsilon_{\mathrm{F}}$. By contrast, a power-law task distribution can support general capabilities (Sec. 2.4). In the power-law task setting, the absolute level of general capabilities increases indefinitely with model scale, as indicated by the scaling of n_{G} in Fig. 2 (top left). This property quantifies the hypothesis that training on large and diverse datasets leads an AI system to develop neural circuits implementing capabilities key to general intelligence (Olah et al., 2020; Bubeck et al., 2023).

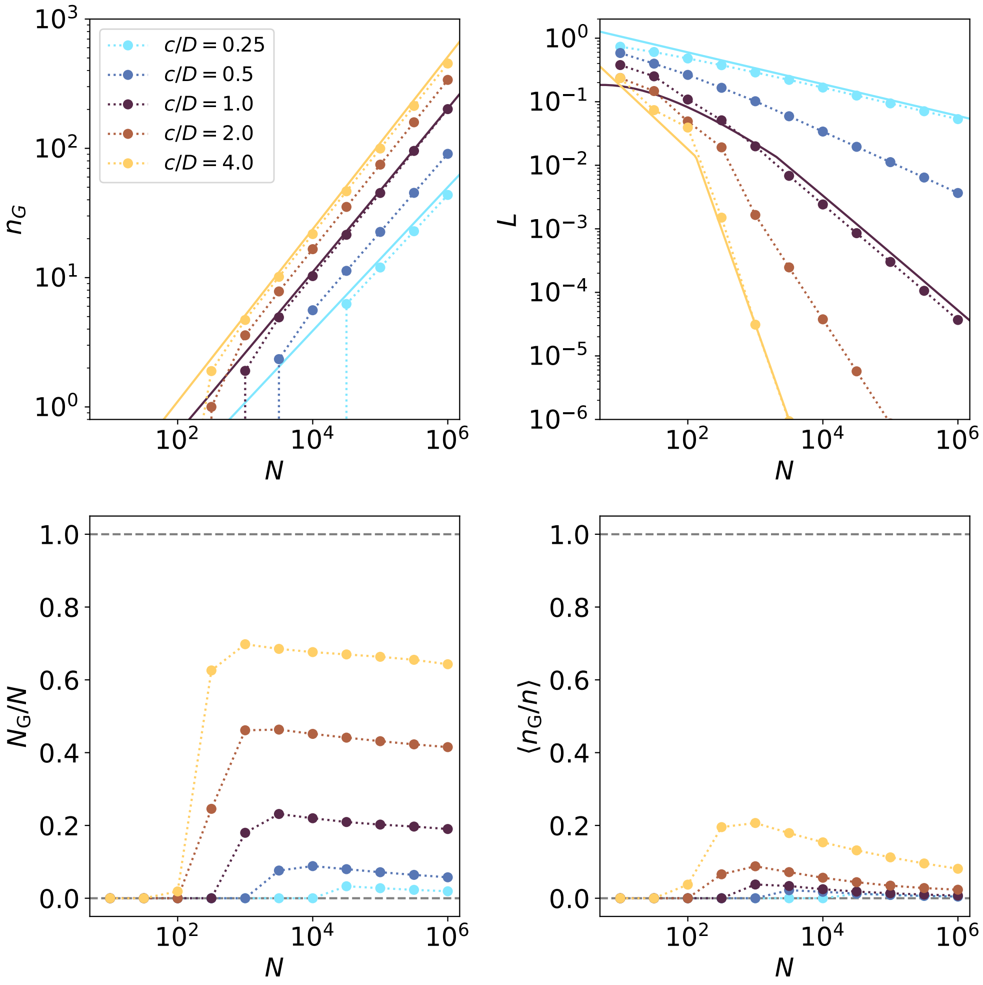

Figure 2: Model scaling curves with power-law distributed data. Points connected by dotted lines show numerically optimized results, and solid lines show analytical approximations. Top left: DOF from general circuits vs. N. Top right: loss vs. N. Bottom left: capacity fraction vs. N. Bottom right: DOF fraction vs. N. All results computed using $\alpha = 1$, $\epsilon_0 = 0.01$, and $\gamma = 0.5$.

4.2 General capabilities emerge abruptly

General capabilities are discontinuous with model scale. As shown in Fig. 2, below a threshold model scale at which the first general circuit is learned, $n_\mathrm{G} = 0$ and the model has no general capabilities. Above that threshold scale, the fraction of the model devoted to general capabilities jumps up abruptly to a finite value, with $n_\mathrm{G} > 1$. This abrupt transition corresponds to a shift from a prediction strategy purely based on memorized features, to one also incorporating general capabilities. A features-only strategy efficiently allocates model capacity to the most important tasks. At small model scales, this effect dominates. However, as discussed in the next section, general capabilities provide a faster loss decrease as the model size increases. An abrupt transition therefore occurs at the scale at which it first becomes optimal to learn general capabilities.

This abrupt transition may relate to the emergent abilities observed in LLMs (Brown et al., 2020; Wei et al., 2022). Many reported examples of emergent abilities appear to involve general-purpose

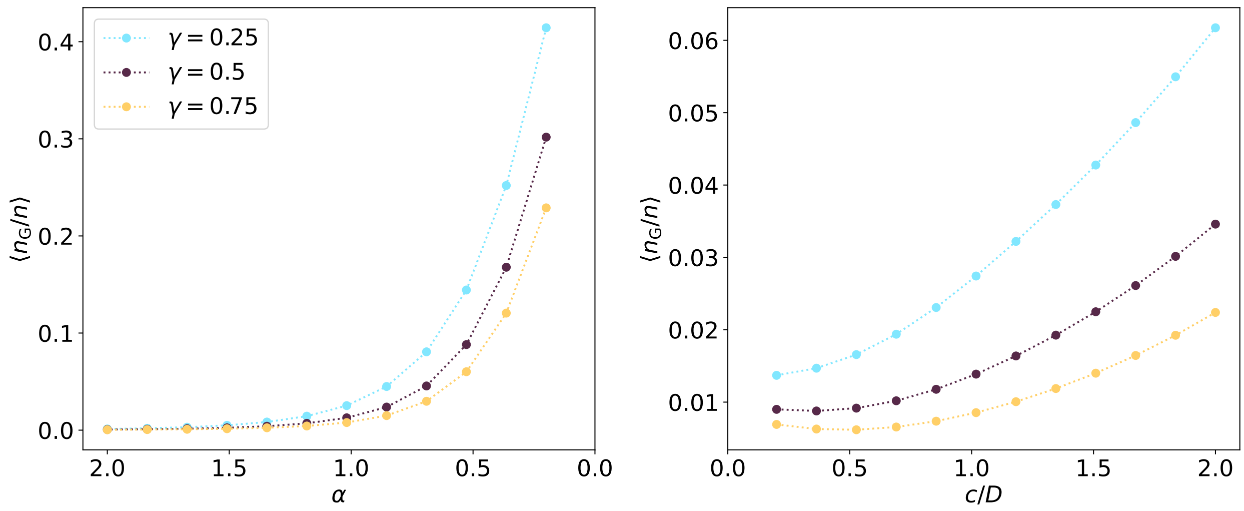

Figure 3: Data diversity and model expressivity increase general capabilities. Left: $\langle n_{\rm G}/n \rangle$ vs. α for $c/D = 1$ and several γ values. Right: $\langle n_{\rm G}/n \rangle$ vs. c/D for $\alpha = 1$ and several γ values. Results computed numerically using $\epsilon_0 = 0.01$ and $N = 1 \times 10^5$.

capabilities, such as in-context few-shot learning. The abrupt appearance of general capabilities above a threshold model scale could plausibly lead to discontinuous performance improvements on related downstream tasks. However, this interpretation needs several caveats. First, the presented model describes how general capabilities vary with model scale, but emergent abilities of LLMs might also depend on training computation, training dataset size, and data composition (Wei et al., 2022). Second, an AI system's intrinsic capabilities may have a complex relationship with measured performance on benchmark tasks. In particular, nonlinear or discontinuous evaluation metrics can confound findings of emergent abilities (Schaeffer et al., 2023). Third, because the toy model approximates the general-circuit population as continuous, it cannot make quantitative predictions about emergence connected to learning discrete general circuits after the first one.

4.3 General capabilities unlock steeper scaling laws

As is visible in Fig. 2 (top right), the emergence of general capabilities is accompanied by a break in the loss curve where the loss begins to decrease faster. The observed break corresponds to the turnover from learning only feature circuits (Sec. 2.4.1) to learning both feature circuits and general circuits (Sec. 2.4.3). A steeper scaling law occurs because general circuits can decrease the loss on multiple tasks in parallel. The appearance of a strong break requires $c/D > \alpha$.

The scaling exponent depends strongly on the parameters c/D and α through the exponent b, which controls the distribution of allocated DOF (Eq. 15). A value of $b \approx 1$ maximizes general capabilities and yields a steep scaling law, while a large negative b minimizes general capabilities and yields a shallow scaling law. For a given b value, Eq. 30 shows that γ effectively interpolates the scaling exponent between the bounds set by c/D and α. The efficiency prefactor ϵ_0 is largely unimportant for determining the scaling exponent but plays a role in setting the threshold at which general capabilities emerge and steeper scaling laws begin.

Assuming that the presented toy model usefully describes real AI systems, theory may suggest realistic parameter values. Kaplan et al. (2020) predicts $c \geq 4$ for neural networks using the ReLU activation function. Brill (2024) predicts values of $D = 4$ and $\alpha = 1$ using a data model based on percolation theory. In general, it seems natural for realistic values of c, D, and α all to be of order unity, so that $b \approx 0$. We therefore conjecture that real AI systems are best modeled by the transitional regime in which non-negligible general capabilities can only just be learned.

4.4 General capabilities decrease proportionally with model scale

After general capabilities emerge, their relative importance diminishes as the model scale increases further. This pattern occurs similarly for the capacity fraction and DOF fraction (Fig. 2, bottom). It arises because the marginal efficiency of the remaining general circuit population decreases as more of them are learned, and feature circuits are relied on proportionally more. As a result, the AI system's reliance on general capabilities decouples from its loss. With increasing scale, less of the loss improvement is attributable to general capabilities. Notably, this occurs even though we assume a training procedure with access to an arbitrary amount of training computation and data. An AI system's performance on novel real-world tasks may be associated with its level of general capabilities learned during pretraining and subsequently elicited through post-training. If so, this suggests that scaling up model size would improve real-world performance less efficiently than pretraining loss scaling laws may appear to imply.

4.5 Data diversity strongly enhances general capabilities

Data diversity has a strong effect on general capabilities. As shown in Fig. 3 (left), the DOF fraction increases sharply as α approaches 0. This limit corresponds to the task distribution becoming more uniform. This toy model therefore suggests that methods to artificially enhance data diversity, such as pruning examples of common tasks or oversampling examples of rare ones, could disproportionately increase general capability learning.

4.6 Model expressivity moderately enhances general capabilities

Model expressivity also enhances general capabilities, but much more weakly. This behavior is shown in Fig 3 (right), and can be interpreted as follows. If c is made larger (holding fixed the data dimension D), the model can achieve lower loss while using the same model capacity. Model capacity can then be more quickly allocated to more tasks' circuits. In effect, model expressivity indirectly increases general capabilities by increasing the accessible task diversity.

4.7 Limitations

This work studies a highly simplified and abstract toy model of AI systems. The analysis assumes that an AI system is purely capacity-constrained, but constraints such as training data, training computation, contextual data at inference time, or inference computation also could be important. In addition, the assumption of an optimal model is only useful if practical training procedures can approximate one. This requires that complications such as circuit learnability and path-dependence in training are negligible. While the simplified notions of model capacity and degrees of freedom applied in this work suffice to predict qualitative phenomena, making specific quantitative predictions would require connecting them to empirically measurable quantities such as neural network parameters. Finally, it may be difficult to empirically constrain the parameters ϵ_0 and γ that govern the putative latent circuit population, independently from analyzing a trained AI system.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

E. Ameisen, J. Lindsey, A. Pearce, W. Gurnee, N. L. Turner, B. Chen, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. Ben Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL `https://transformer-circuits.pub/2025/attribution-graphs/methods.html`.

Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

L. Berglund, A. C. Stickland, M. Balesni, M. Kaufmann, M. Tong, T. Korbak, D. Kokotajlo, and O. Evans. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*, 2023.

J. Betley, X. Bao, M. Soto, A. Sztyber-Betley, J. Chua, and O. Evans. Tell me about yourself: Llms are aware of their learned behaviors. *arXiv preprint arXiv:2501.11120*, 2025.

B. Bordelon, A. Atanasov, and C. Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024.

D. Braun, L. Bushnaq, S. Heimersheim, J. Mendel, and L. Sharkey. Interpretability in parameter space: Minimizing mechanistic description length with attribution-based parameter decomposition. *arXiv preprint arXiv:2501.14926*, 2025.

A. Brill. Neural scaling laws rooted in the data distribution. *arXiv preprint arXiv:2412.07942*, 2024.

A. Brill. Representation learning on a random lattice. *arXiv preprint arXiv:2504.20197*, 2025.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

V. Cabannes, E. Dohmatob, and A. Bietti. Scaling laws for associative memories. *arXiv preprint arXiv:2310.02984*, 2023.

J. Carlsmith. Scheming ais: Will ais fake alignment during training in order to get power? *arXiv preprint arXiv:2311.08379*, 2023.

F. Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

J. Dunefsky, P. Chlenski, and N. Nanda. Transcoders find interpretable llm feature circuits. *arXiv preprint arXiv:2406.11944*, 2024.

V. Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

V. Feldman and C. Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

N. Fonseca, S. H. Lee, C. Mingard, A. Louis, et al. An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem. *Advances in Neural Information Processing Systems*, 37:39632–39693, 2024.

D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.

Gemini Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

B. Goertzel. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, 5(1):1–48, Dec. 2014. doi: 10.2478/jagi-2014-0001.

R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

M. Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

S. Jain, R. Kirk, E. S. Lubana, R. P. Dick, H. Tanaka, E. Grefenstette, T. Rocktäschel, and D. S. Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*, 2023.

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

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

R. Laine, B. Chughtai, J. Betley, K. Hariharan, M. Balesni, J. Scheurer, M. Hobbhahn, A. Meinke, and O. Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms. *Advances in Neural Information Processing Systems*, 37:64010–64118, 2024.

S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17:391–444, 2007.

J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`.

Z. Liu, Y. Liu, E. J. Michaud, J. Gore, and M. Tegmark. Physics of skill learning. *arXiv preprint arXiv:2501.12391*, 2025.

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A. Maloney, D. A. Roberts, and J. Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

S. Marks, C. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

E. Michaud, Z. Liu, U. Girit, and M. Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023.

C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Z. Pan, S. Wang, and J. Li. Understanding llm behaviors via compression: Data generation, knowledge acquisition and scaling laws. *arXiv preprint arXiv:2504.09597*, 2025.

E. Paquette, C. Paquette, L. Xiao, and J. Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

A. Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36:55565–55581, 2023.

U. Sharma and J. Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022. URL `http://jmlr.org/papers/v23/20-1111.html`.

E. S. Spelke and K. D. Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

D. Vaintrob. Efficiency spectra and "bucket of circuits" cartoons. `https://www.lesswrong.com/posts/fhK3dKW9RQhr9Ymf7/efficiency-spectra-and-bucket-of-circuits-cartoons`, 2025. LessWrong.

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

A Single task derivation

For a single-task data distribution, the Lagrangian is

$$\mathcal{L}_1 = (n_{\mathrm{F}} + n_{\mathrm{G}})^{-c/D} - \lambda(n_{\mathrm{F}} + m n_{\mathrm{G}}^{1+\gamma} - N), \tag{7}$$

with the optimal values of n_{F} and n_{G} given by the equations,

$$\frac{\partial \mathcal{L}_1}{\partial n_\mathrm{F}} = 0 = -\frac{c}{D}(n_\mathrm{F} + n_\mathrm{G})^{-(1+c/D)} - \lambda, \tag{32}$$

$$\frac{\partial \mathcal{L}_1}{\partial n_\mathrm{G}} = 0 = -\frac{c}{D}(n_\mathrm{F} + n_\mathrm{G})^{-(1+c/D)} - \lambda m(1+\gamma)n_\mathrm{G}^{\gamma}, \tag{33}$$

$$\frac{\partial \mathcal{L}_1}{\partial \lambda} = 0 = N - n_\mathrm{F} - m n_\mathrm{G}^{1+\gamma}. \tag{34}$$

From Eq. 34, $n_\mathrm{F} = N - m n_\mathrm{G}^{1+\gamma}$, and from Eq. 32, $\lambda = -(c/D)(n_\mathrm{F} + n_\mathrm{G})^{-(1+c/D)}$. Substituting into Eq. 33 gives

$$0 = -\frac{c}{D}(N - m n_\mathrm{G}^{1+\gamma} + n_\mathrm{G})^{-(1+c/D)} \left[1 - m(1+\gamma)n_\mathrm{G}^{\gamma}\right], \tag{35}$$

which has the solution

$$n_\mathrm{G} = [m(1+\gamma)]^{-1/\gamma} = \epsilon_0^{1/\gamma}, \tag{8}$$

$$n_\mathrm{F} = N - (1+\gamma)^{-1}\epsilon_0^{1/\gamma}. \tag{9}$$

It follows that

$$L = \left(N + \frac{\gamma}{1+\gamma}\epsilon_0^{1/\gamma}\right)^{-c/D}. \tag{10}$$

Eq. 8, along with the constraints $1 \leq n_\mathrm{G} \leq N$, implies that feature circuits and general circuits can only coexist if $\epsilon_0 \geq 1$ and $N \gtrsim (1+\gamma)m\epsilon_0^{(1+\gamma)/\gamma}$. Of allowed cases, whichever one minimizes the loss determines the overall behavior. Combining the three cases yields

$$n_\mathrm{G} = \begin{cases} 0 & \epsilon_0 < 1, \\ (N/m)^{1/(1+\gamma)} & \epsilon_0 \geq 1,\ N \leq (1+\gamma)m\epsilon_0^{(1+\gamma)/\gamma}, \\ \epsilon_0^{1/\gamma} & \epsilon_0 \geq 1,\ N > (1+\gamma)m\epsilon_0^{(1+\gamma)/\gamma}. \end{cases} \tag{36}$$

If $\epsilon_0 < 1$, then $\langle n_\mathrm{G}/n \rangle = 0$ exactly. If $\epsilon_0 \geq 1$, then for large N, $\langle n_\mathrm{G}/n \rangle = \epsilon_0^{1/\gamma}\left(N + \frac{\gamma}{1+\gamma}\epsilon_0^{1/\gamma}\right)^{-1}$, so as N increases, $\langle n_\mathrm{G}/n \rangle$ approaches 0.

B Power-law task distribution derivation: Case F

B.1 Lagrange equations

With feature circuits only, the Lagrangian is,

$$\mathcal{L}_\mathrm{F}(a, k_\mathrm{br}, \lambda) = \sum_{k=1}^{k_\mathrm{br}} \left(ak^{b-1}\right)^{-c/D} \frac{1}{\zeta(1+\alpha)} k^{-(1+\alpha)} + \sum_{k=k_\mathrm{br}}^{\infty} \frac{1}{\zeta(1+\alpha)} k^{-(1+\alpha)} - \lambda\left(\sum_{k=1}^{k_\mathrm{br}} ak^{b-1} - N\right) \tag{17}$$

This simplifies to

$$\begin{aligned}\mathcal{L}_F &= \left(\frac{a^{-c/D}}{\zeta(1+\alpha)} - \lambda a\right)\sum_{k=1}^{k_{\mathrm{br}}} k^{b-1} + \frac{1}{\zeta(1+\alpha)}\sum_{k=k_{\mathrm{br}}}^{\infty} k^{-(1+\alpha)} + \lambda N \\ &\approx \left(\alpha a^{-c/D} - \lambda a\right)\frac{k_{\mathrm{br}}^b - 1}{b} + k_{\mathrm{br}}^{-\alpha} + \lambda N, \end{aligned} \tag{37}$$

where we used the identity,

$$\frac{c}{D}(1-b) = \alpha + b.$$

The optimal values of a and k_{br} are determined by the equations,

$$\frac{\partial \mathcal{L}_{\mathrm{F}}}{\partial a} = 0 = \left(-\alpha\frac{c}{D}a^{-(1+c/D)} - \lambda\right)\frac{k_{\mathrm{br}}^b - 1}{b}, \tag{38}$$

$$\frac{\partial \mathcal{L}_{\mathrm{F}}}{\partial k_{\mathrm{br}}} = 0 = \left(\alpha a^{-c/D} - \lambda a\right)k_{\mathrm{br}}^{b-1} - \alpha k_{\mathrm{br}}^{-(1+\alpha)}, \tag{39}$$

$$\frac{\partial \mathcal{L}_{\mathrm{F}}}{\partial \lambda} = 0 = N - a\frac{k_{\mathrm{br}}^b - 1}{b}. \tag{40}$$

Solving Eq. 38 and Eq. 39 for λ and equating the resulting expressions gives

$$-\alpha\frac{c}{D}a^{-(1+c/D)} = \alpha\left(a^{-(1+c/D)} - \frac{1}{a}k_{\mathrm{br}}^{-\frac{c}{D}(1-b)}\right), \tag{41}$$

which reduces to

$$a = Ck_{\mathrm{br}}^{1-b}, \tag{18}$$

where $C = (1+c/D)^{1/(c/D)}$. Solving Eq. 40 for a and substituting into Eq. 18, we obtain

$$\frac{1}{b}k_{\mathrm{br}}\left(1 - k_{\mathrm{br}}^{-b}\right) = \frac{N}{C}. \tag{19}$$

B.2 Limiting cases

Eq. 19 has no closed-form solution, but we can find approximations in limiting cases. Table 1 lists the most suitable limiting case for each combination of magnitudes of c/D and α. For brevity, we denote $K \equiv k_{\mathrm{br}}\left(1 - k_{\mathrm{br}}^{-b}\right)/b$.

1. $0 < b < 1,\ |b| \sim 1$.

   Expanding to first order in the small quantity $(1-b)$, we obtain

$$K \approx (k_{\mathrm{br}} - 1) + (1-b)\left(k_{\mathrm{br}} - \ln k_{\mathrm{br}} - 1\right). \tag{42}$$

2. $|b| \ll 1$.

   Expanding to leading order in b, we obtain

$$K \approx k_{\rm br} \ln k_{\rm br} \left(1 - \frac{b}{2} \ln k_{\rm br}\right). \tag{43}$$

   Neglecting the correction term, Eq. 43 can be equivalently written as $k_{\rm br} = K/W(K) = \exp W(K)$, where $W(x)$ is the Lambert W function.

3. $b < 0,\ |b| \gg 1$.

   Assuming $k_{\rm br}$ is not too small, we can make the approximation

$$K \approx -\frac{1}{b} k_{\rm br}^{1-b} = \frac{1}{|b|} k_{\rm br}^{1+|b|}. \tag{44}$$

Using these approximations, we have, for large N,

$$k_{\rm br} \approx \begin{cases} N/C & 0 < b < 1,\ |b| \sim 1, \\ W(N/C)^{-1}(N/C) & |b| \ll 1, \\ (|b|N/C)^{1/(1+|b|)} & b < 0,\ |b| \gg 1. \end{cases} \tag{20}$$

The loss is then, using the above approximations and for large N,

$$\begin{aligned} L_{\rm F} &= \alpha a^{-c/D} \frac{k_{\rm br}^b - 1}{b} + k_{\rm br}^{-\alpha} \\ &= \left[1 + \left(\frac{\alpha}{1 + c/D}\right) \frac{N/C}{k_{\rm br}}\right] k_{\rm br}^{-\alpha} \qquad (21) \\ &\approx \begin{cases} C^\alpha \left[1 + \frac{\alpha}{1+c/D}\right] N^{-\alpha} & 0 < b < 1,\ |b| \sim 1, \\ C^\alpha \left[1 + \frac{\alpha}{1+c/D} \cdot f(N)\right] N^{-\alpha} & |b| \ll 1, \\ \frac{\alpha}{|b|}(|b|N)^{-c/D} & b < 0,\ |b| \gg 1, \end{cases} \qquad (45) \end{aligned}$$

where

$$f(N) = W(N/C)^{1+\alpha} - \frac{1 + c/D}{\alpha}.$$

The prefactor in Eq. 45 differs from the one in the approximate model scaling law derived by Brill (2024). Eq. 45 is more accurate, as the model scaling law derived by Brill (2024) made an additional approximation equivalent to $C \approx 1$. Also, the prefactor here incorporates a factor of α from the normalization.

Regime	c/D	α	$\lvert b\rvert$	$b \to$	$C \to$
I	$\ll 1$	$\ll 1$	$\ll 1$	$c/D - \alpha$	e
II		~ 1	~ 1	$-\alpha$	
III		$\gg 1$	$\gg 1$		
IV	~ 1	$\ll 1$	~ 1	$(c/D)/(1+c/D)$	2
V		~ 1	$\ll 1$	$(c/D-\alpha)/(1+c/D)$	
VI		$\gg 1$	$\gg 1$	$-\alpha/(1+c/D)$	
VII	$\gg 1$	$\ll c/D$	~ 1	$1-(1+\alpha)/(1+c/D)$	1
VIII		$\sim c/D$	$\ll 1$	$1-\alpha/(c/D)$	
IX		$\gg c/D$	$\gg 1$		

Table 1: Parameter regimes in terms of variables c/D and α, listing the closest limiting approximation for $|b|$ and corresponding approximations for b and C.

C Power-law task distribution derivation: Case FG

C.1 Lagrange equations

If both feature circuits and general circuits are instantiated, the Lagrangian simplifies to

$$\mathcal{L}_{\mathrm{FG}}(a, k_{\mathrm{br}}, n_{\mathrm{G}}, \lambda) \approx \left(\alpha a^{-c/D} - \lambda a\right) \frac{k_{\mathrm{br}}^{b} - 1}{b} + n_{\mathrm{G}}^{-c/D} k_{\mathrm{br}}^{-\alpha} - \lambda \left(m n_{\mathrm{G}}^{1+\gamma} - \left(k_{\mathrm{br}} - 1\right) n_{\mathrm{G}} - N\right), \quad (24)$$

and the optimal values of a, k_{br}, and n_{G} are determined by the equations

$$\frac{\partial \mathcal{L}_{\mathrm{FG}}}{\partial a} = 0 = \left(-\alpha \frac{c}{D} a^{-(1+c/D)} - \lambda\right) \frac{k_{\mathrm{br}}^{b} - 1}{b}, \quad (46)$$

$$\frac{\partial \mathcal{L}_{\mathrm{FG}}}{\partial k_{\mathrm{br}}} = 0 = \left(\alpha a^{-c/D} - \lambda a\right) k_{\mathrm{br}}^{b-1} - \alpha n_{\mathrm{G}}^{-c/D} k_{\mathrm{br}}^{-(1+\alpha)} + \lambda n_{\mathrm{G}}, \quad (47)$$

$$\frac{\partial \mathcal{L}_{\mathrm{FG}}}{\partial n_{\mathrm{G}}} = 0 = -\frac{c}{D} n_{\mathrm{G}}^{-(1+c/D)} k_{\mathrm{br}}^{-\alpha} - \lambda m (1+\gamma) n_{\mathrm{G}}^{\gamma} + \lambda (k_{\mathrm{br}} - 1), \quad (48)$$

$$\frac{\partial \mathcal{L}_{\mathrm{FG}}}{\partial \lambda} = 0 = N - a \frac{k_{\mathrm{br}}^{b} - 1}{b} - m n_{\mathrm{G}}^{1+\gamma} + \left(k_{\mathrm{br}} - 1\right) n_{\mathrm{G}}. \quad (49)$$

Each of Eq. 46, Eq. 47, and Eq. 48 can be solved for λ, giving the expressions

$$\lambda = -\alpha \frac{c}{D} a^{-(1+c/D)}, \quad (50)$$

$$\lambda = \alpha \left(a^{-(1+c/D)} - \frac{1}{a} \left(n_{\mathrm{G}} k_{\mathrm{br}}^{1-b}\right)^{-c/D}\right) \left(1 - \frac{1}{a} n_{\mathrm{G}} k_{\mathrm{br}}^{1-b}\right)^{-1}, \quad (51)$$

$$\lambda = -\frac{c}{D} n_{\mathrm{G}}^{-(1+c/D)} k_{\mathrm{br}}^{-\alpha} \left[m(1+\gamma) n_{\mathrm{G}}^{\gamma} - (k_{\mathrm{br}} - 1)\right]^{-1}. \quad (52)$$

Equating Eq. 50 and Eq. 52 yields the relation

$$\frac{a}{n_{\mathrm{G}} k_{\mathrm{br}}^{1-b}} = \Gamma, \tag{53}$$

where

$$\Gamma = \left(\frac{\alpha}{k_{\mathrm{br}}} \left[m(1+\gamma) n_{\mathrm{G}}^{\gamma} - (k_{\mathrm{br}} - 1) \right] \right)^{\frac{1}{1+c/D}}. \tag{54}$$

Equating Eq. 50 and Eq. 51 yields

$$0 = \left(\frac{a}{n_{\mathrm{G}} k_{\mathrm{br}}^{b-1}} \right)^{1+c/D} - \left(1 + \frac{c}{D}\right) \left(\frac{a}{n_{\mathrm{G}} k_{\mathrm{br}}^{b-1}} \right) + \frac{c}{D}, \tag{55}$$

and the further substitution of Eq. 53 results in the equation

$$0 = \Gamma^{1+c/D} - \left(1 + \frac{c}{D}\right) \Gamma + \frac{c}{D}. \tag{56}$$

By inspection, $\Gamma = 1$ is a solution. Furthermore, because the derivative $(1 + c/D)(\Gamma^{c/D} - 1)$ is positive for all $\Gamma > 1$ and negative for all $0 < \Gamma < 1$, this solution is unique. It follows that

$$a = n_{\mathrm{G}} k_{\mathrm{br}}^{1-b}, \tag{25}$$

$$k_{\mathrm{br}} = \frac{\alpha}{1+\alpha} \left(1 + m(1+\gamma) n_{\mathrm{G}}^{\gamma}\right). \tag{26}$$

Finally, Eq. 49, along with Eq. 25, yields

$$N = n_{\mathrm{G}} \left[\frac{k_{\mathrm{br}} \left(1 - k_{\mathrm{br}}^{-b}\right)}{b} - (k_{\mathrm{br}} - 1) + m n_{\mathrm{G}}^{\gamma} \right]. \tag{27}$$

C.2 Limiting cases

As with Eq. 19, Eq. 27 does not have a simple closed-form solution for k_{br}. We approach it by examining exact or approximate solutions for specific cases. We can simplify Eq. 27 by considering approximations for limiting cases and keeping only dominant terms for large N.

1. $0 < b < 1, \ |b| \sim 1$.

   Using the approximation Eq. 42, we have

$$\begin{aligned} N &\approx n_{\mathrm{G}} \left[(1-b)(k_{\mathrm{br}} - \ln k_{\mathrm{br}} - 1) + m n_{\mathrm{G}}^{\gamma} \right] \\ &\approx (1+R) m n_{\mathrm{G}}^{1+\gamma}, \end{aligned} \tag{57}$$

   where

$$R \equiv \frac{\alpha(1+\gamma)}{1+c/D}, \tag{58}$$

and the expression for n_{G} is

$$n_{\mathrm{G}} = (1+R)^{-1/(1+\gamma)} \left(\frac{N}{m}\right)^{1/(1+\gamma)}. \tag{59}$$

2. $|b| \ll 1$.

   Using Eq. 43, we have

$$\begin{aligned} N &\approx n_{\mathrm{G}} \left[k_{\mathrm{br}} \ln k_{\mathrm{br}} - (k_{\mathrm{br}} - 1) + m n_{\mathrm{G}}^{\gamma}\right] \\ &\approx \left[\frac{\alpha(1+\gamma)}{1+\alpha} \ln\left(\frac{\alpha(1+\gamma)}{1+\alpha} m n_{\mathrm{G}}^{\gamma}\right) - \frac{\alpha(1+\gamma)}{1+\alpha} + 1\right] m n_{\mathrm{G}}^{1+\gamma} \\ &\approx \left[1 + R \ln\left(R m n_{\mathrm{G}}^{\gamma}/e\right)\right] m n_{\mathrm{G}}^{1+\gamma}, \end{aligned} \tag{60}$$

   and the expression for n_{G} is

$$n_{\mathrm{G}} = (1 + R \cdot f(N))^{-1/(1+\gamma)} \left(\frac{N}{m}\right)^{1/(1+\gamma)}, \tag{61}$$

   where

$$f(N) = \frac{\gamma}{1+\gamma} W\left(\frac{(1+\gamma)N}{R\gamma m} \cdot \exp\left(\frac{(1+\gamma)(1+R\ln(Rm/e))}{R\gamma}\right)\right) - \frac{1}{R}. \tag{62}$$

3. $b < 0,\ |b| \gg 1$.

   Using Eq. 44, we have

$$\begin{aligned} N &\approx n_{\mathrm{G}} \left[\frac{1}{|b|} k_{\mathrm{br}}^{1-b} - (k_{\mathrm{br}} - 1) + m n_{\mathrm{G}}^{\gamma}\right] \\ &\approx \frac{m^{-b}}{|b|} \left(\frac{R}{1-b}\right)^{1-b} m n_{\mathrm{G}}^{1+\gamma-b\gamma}, \end{aligned} \tag{63}$$

   so that the expression for n_{G} is,

$$n_{\mathrm{G}} = \left(\frac{|b|}{m^{-b}}\right)^{1/(1+\gamma-b\gamma))} \left(\frac{1-b}{R}\right)^{(1-b)/(1+\gamma-b\gamma)} \left(\frac{N}{m}\right)^{1/(1+\gamma-b\gamma)}. \tag{64}$$

C.3 Loss curve

The loss is then, keeping only dominant terms for large N,

$$\begin{aligned}
L_{\mathrm{FG}} &= \alpha a^{-c/D} \frac{k_{\mathrm{br}}^{b} - 1}{b} + n_{\mathrm{G}}^{-c/D} k_{\mathrm{br}}^{-\alpha} \\
&= \left[1 + \frac{\alpha}{k_{\mathrm{br}}} \left(\frac{N}{n_{\mathrm{G}}} + (k_{\mathrm{br}} - 1) - m n_{\mathrm{G}}^{\gamma} \right)\right] n_{\mathrm{G}}^{-c/D} k_{\mathrm{br}}^{-\alpha} \\
&\approx \left[\frac{\alpha}{1+\alpha} m(1+\gamma)\right]^{-\alpha} \left(\frac{1+\alpha}{1+\gamma}\right) \left(\gamma + \frac{N}{m n_{\mathrm{G}}^{1+\gamma}}\right) n_{\mathrm{G}}^{-(\alpha\gamma + c/D)} \\
&= \alpha \left(\frac{1-b}{R}\right)^{1+\alpha} \left(\gamma + \frac{N}{m n_{\mathrm{G}}^{1+\gamma}}\right) \left(m n_{\mathrm{G}}^{1+\gamma}\right)^{-\alpha} n_{\mathrm{G}}^{-(c/D - \alpha)} \qquad (29)
\end{aligned}$$

For the case $0 < b < 1,\ |b| \sim 1$, we use Eq. 29 and Eq. 59 to obtain

$$L \approx \alpha \left(\frac{1-b}{R}\right)^{1+\alpha} (1+\gamma+R)(1+R)^{\frac{c/D-\alpha}{1+\gamma}+\alpha} \left(\frac{N}{m}\right)^{-\frac{c/D-\alpha}{1+\gamma}} N^{-\alpha}. \qquad (65)$$

For the case $|b| \ll 1$, we use Eq. 29 and Eq. 61 to obtain

$$L \approx \alpha \left(\frac{1-b}{R}\right)^{1+\alpha} (1+\gamma+R\cdot f(N))(1+R\cdot f(N))^{\frac{c/D-\alpha}{1+\gamma}+\alpha} \left(\frac{N}{m}\right)^{-\frac{c/D-\alpha}{1+\gamma}} N^{-\alpha}, \qquad (66)$$

with $f(N)$ given by Eq. 62.

For the case $b < 0,\ |b| \gg 1$, we use Eq. 29 and Eq. 64 to obtain, after some algebra,

$$L \approx \frac{\alpha}{|b|} (|b|N)^{-c/D}. \qquad (67)$$

Eq. 67 is the same as for Case F (Eq. 45), showing that our approximation in this limit is too coarse to characterize the subdominant contribution from general circuits.

C.4 Threshold model scale

Putting the above results together, the overall loss is given by

$$L = \min\left(L_{\mathrm{F}}, L_{\mathrm{G}}, L_{\mathrm{FG}}\right). \qquad (68)$$

.

At a threshold model scale N_{th}, the class or classes of instantiated circuits may abruptly transition if the case giving the minimum loss changes. A critical transition can only occur if the loss curves of two cases intersect at a model scale $N_{\mathrm{th}} > 1$. When $0 < b < 1,\ |b| \sim 1$, equating Eq. 45 and Eq. 65 yields

$$N_{\mathrm{th}} = m(1+R)^{1+R/b} \left[R^{-1}(1-b)^2(1+\alpha)^{(1-\alpha)/\alpha}\right]^{R/b}. \qquad (69)$$

In the limit $|b| \ll 1$, equating Eq. 45 and Eq. 66 yields

$$N_{\mathrm{th}} = m(1 + R \cdot f_{\mathrm{FG}}(N_c))^{1+R/b} \left(\frac{1 + \gamma + R \cdot f_{\mathrm{FG}}(N_c)}{1 + \gamma + R \cdot f_{\mathrm{F}}(N_c)} \right)^{R/b} \left[R^{-1}(1-b)^2(1+\alpha)^{(1-\alpha)/\alpha} \right]^{R/b}, \quad (70)$$

which could be solved numerically for N_{th}. When $b < 0,\ |b| \gg 1$, Eq. 45 and Eq. 67 are identical, indicating that our approximations in this limit are too coarse to compute N_{th}.

In the numerical experiments reported in Sec. 3, the analytical loss curves are combined using Eq. 68, rather than using Eq. 69 or Eq. 70.