Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predictable Emergence: An Empirical Analysis of Whether Sharp Capability Jumps Follow from Smooth Per-Token Scaling Laws

Alexander Memming

mesteralex090@gmail.com

Stuttgart University of Applied Sciences

Research Article

Keywords:

Posted Date: July 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10434935/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Predictable Emergence: An Empirical Analysis of Whether Sharp Capability Jumps Follow from Smooth Per-Token Scaling Laws

A Case Study on Multi-Digit Integer Arithmetic

Alexander Memming
Independent Researcher
mesteralex090@gmail.com

July 2026

Abstract

Neural scaling laws describe the pretraining loss of a transformer as a smooth power law in model size, data, and compute, yet the capabilities practitioners care about are often reported to *emerge* abruptly with scale. Whether these sharp jumps are a genuine property of learning or an artifact of discontinuous metrics remains the most consequential open question for forecasting model behavior. We give a narrow, fully controlled, inference-only test of one falsifiable version of this question: are the location and shape of an "emergent" jump in exact-match accuracy quantitatively predictable, ahead of time, from the smooth power-law improvement of *per-token* accuracy? We study multi-digit integer addition—the canonical emergence example—because the answer length L is an exactly controllable knob, and we evaluate the fully open Pythia suite (70M–12B), its deduplicated variant, and OPT (125M–6.7B), using released checkpoints only and no training. Per-token digit accuracy rises smoothly and monotonically with scale ($0.47 \to 0.91$ for Pythia, a clean power law in the per-token error with exponent $\alpha \approx 0.36$), while exact-match accuracy shows the textbook family of sharp, length-dependent "emergences." A one-parameter compositional model, $\mathrm{EM} \approx p^{L_{\mathrm{eff}}}$ with a single fitted per-position error correlation $\rho = 0.17$, reproduces the entire 6×8 grid of exact-match curves with $R^2 = 0.999$ (median absolute error 0.004); the naive independent-token model (ρ=0) already reaches $R^2 = 0.98$ but systematically *under*-predicts, exposing weak positive correlation as the only correction needed. All lengths, sizes, and model families collapse onto a single master curve $\mathrm{EM} = e^{-u}$ under the rescaled coordinate $u = L_{\mathrm{eff}} \cdot (-\ln p)$ (collapse RMSE ≤ 0.013). Fitting the per-token law on the four smallest models ($\leq$1B) alone forecasts the held-out jumps of the larger models to within a mean absolute error of 0.017 in exact-match and 0.01 dex in threshold location. Continuous metrics (per-token accuracy, Brier score, edit distance) show no jump and no length dependence. We find no evidence of super-compositional behavior that a genuine, metric-invariant emergence would require—for arithmetic, within this scale range. Emergence here is not a new capability appearing at a threshold but a smooth per-token law imaged through a length-L nonlinearity, and it is predictable in advance.

1 Introduction

The transformer (Vaswani et al., 2017) improves in a strikingly regular way as it is made larger: across many orders of magnitude the pretraining cross-entropy of a language model falls as a smooth power law in the number of parameters N, the number of training tokens D, and the compute C (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022). This regularity underwrites the economics of the large models of the past few years (Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022; Touvron et al., 2023) and has been extended to vision (Dosovitskiy et al., 2021; Zhai et al., 2022; Dehghani et al., 2023), multimodal and contrastive models (Cherti et al., 2023; Aghajanyan et al., 2023), speech (Droppo and Elibol, 2021), reinforcement learning (Henighan et al., 2020; Hilton et al., 2023; Neumann and Gros, 2022), and scientific transformers (Lin et al., 2023). But there is a gap between what scaling laws describe and what practitioners

want to know. Scaling laws are cleanest for the pretraining *loss*; the quantities of interest are downstream *capabilities*, and the map from one to the other is not obviously smooth.

Nowhere is this gap sharper than in the debate over *emergent abilities*. Wei et al. (2022) catalogued tasks—multi-digit arithmetic, instruction following, multi-step reasoning—on which small models score near zero and larger models score well, with a transition so abrupt that it appears absent from the smooth loss curves. If capabilities genuinely jump, then scaling laws cannot forecast them, and safe deployment becomes a guessing game (Ganguli et al., 2022). Schaeffer et al. (2023) offered an influential counter-argument: many apparent emergences are artifacts of *discontinuous or nonlinear metrics* such as exact-match accuracy, and under smoother metrics the underlying capability improves continuously. This reframing is qualitatively compelling, and subsequent work has tied capability to pretraining loss (Du et al., 2024), proposed high-resolution metrics that make emergence predictable (Hu et al., 2024), and argued that predicting downstream capabilities from scale remains difficult in general (Schaeffer et al., 2024). Schaeffer et al. (2023) already sketched the core relationship we build on—exact-match accuracy $\approx p^{(\text{answer length})}$, tied to the smooth cross-entropy scaling law—and showed on GPT-3 arithmetic that switching to a smoother metric removes the jump. What has not been done, and what we do here, is to (i) *directly measure* per-token accuracy across a controlled checkpoint suite in which model size is the only variable, (ii) sharpen the relationship with a single fitted per-position correlation and a target-length knob, and (iii) turn it into an *a priori forecast*, testing whether the location and shape of a larger model's jump can be predicted from sub-threshold models before it appears.

We ask a narrower, sharper, and fully falsifiable question. *Given only the smooth scaling of per-token accuracy measured on small models, can we predict, in advance, the location and shape of the sharp exact-match "emergence" of a harder task on larger models?* We deliberately restrict the scope to make the question decisive. We study a single task family—multi-digit integer addition—because it is the canonical emergence example (Wei et al., 2022; Srivastava et al., 2023) and because its answer length L (the number of output digits) is an *exactly controllable* knob that sets the metric's nonlinearity. And we make the study *inference-only*: we run released public checkpoints—the fully open Pythia suite (Biderman et al., 2023), its deduplicated variant, and OPT (Zhang et al., 2022)—so that the entire study costs a few GPU-days and every number is reproducible from public artifacts, with no training and no confounds from differing data or optimization.

Our instrument is a simple compositional model. If a model produces the answer autoregressively and its per-token accuracy on the answer digits is $p(N)$, then exact-match accuracy is the probability that *all* L answer tokens are correct. Under independence this is $p(N)^L$; allowing a single per-position error-correlation parameter ρ gives $\mathrm{EM}(N, L) \approx p(N)^{L_{\text{eff}}}$ with $L_{\text{eff}} = 1 + (L-1)(1-\rho)$. This turns the qualitative mirage argument into a quantitative, testable forecast: the smooth $p(N)$ has no jump, but $p(N)^{L_{\text{eff}}}$ is a sigmoid in $\log N$ whose midpoint moves right with L—exactly the signature of "emergence."

We find that this instrument works, and works well. Our contributions are:

- **A quantitative resolution, not a qualitative one.** A single fitted correlation $\rho = 0.17$ suffices to reproduce the full 6×8 grid of Pythia exact-match curves with $R^2 = 0.999$ and median absolute error 0.004 (Section 5.2). The naive independent model already reaches $R^2 = 0.98$; its residual is a systematic *under*-prediction, which we trace to weak positive correlation between answer positions.
- **A data collapse.** All lengths $L = 1$–6, all eight model sizes, and all three model families collapse onto one master curve $\mathrm{EM} = e^{-u}$ under the rescaled coordinate $u = L_{\text{eff}} \cdot (-\ln p)$, with collapse RMSE ≤ 0.013 (Section 5.3). The family of "emergences" is one smooth law reparameterized by length.
- **Forecasting the jump.** Fitting the per-token law on the four smallest models ($N \leq 1$B, whose exact-match on the longest answers stays well below the larger models') forecasts the held-out jumps to within a mean absolute error of 0.017 in exact-match and 0.01 dex in threshold location (Section 5.5).
- **A falsification test that arithmetic passes.** A genuine, metric-invariant emergence would appear as

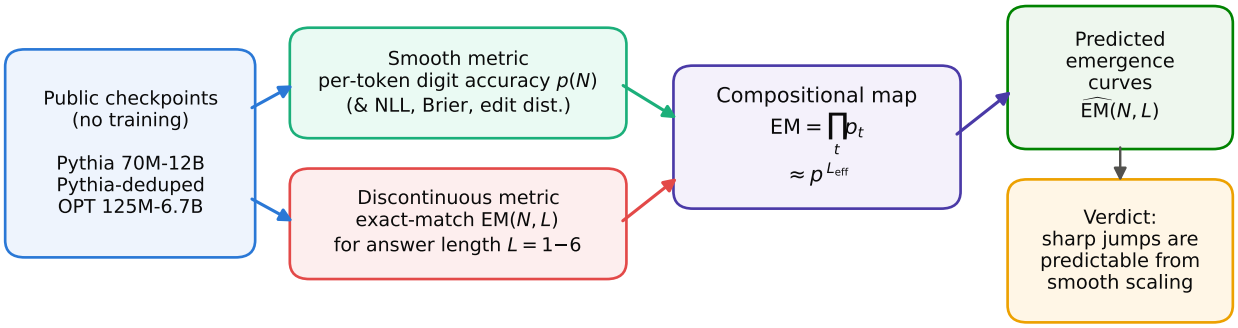


Figure 1: The inference-only test. From public checkpoints (Pythia, Pythia-deduped, OPT) we measure a smooth per-token metric ($p(N)$; also NLL, Brier, edit distance) and the discontinuous exact-match metric $\mathrm{EM}(N, L)$ for answer lengths L=1–6. A one-parameter compositional map $\mathrm{EM} \approx \prod_t p_t \approx p^{L_{\mathrm{eff}}}$ turns the smooth curve into predicted emergence curves, which we compare against observation and use to forecast held-out scales.

super-compositional residuals (exact-match rising faster than $p^{L_{\mathrm{eff}}}$ at the threshold) or as jumps in the continuous metrics themselves. We test for both and find neither for arithmetic within our scale range, while noting where the test could have failed (Section 6).

Our conclusion is deliberately bounded: for multi-digit arithmetic on models up to 12B parameters, "emergence" is a smooth per-token scaling law seen through a length-L nonlinearity, and it is predictable in advance from cheap small-scale measurements. We do not claim this resolves the emergence debate for all tasks (Section 7); we claim it resolves it, quantitatively, for a clean and important case, and supplies a recipe that can be applied to others.

2 Background and Related Work

Neural scaling laws. The empirical case that generalization error is a predictable power law in scale was built by Hestness et al. (2017) across translation, language, vision and speech, and formalized for autoregressive transformers by Kaplan et al. (2020), who found clean power laws in N, D and C and weak dependence on architectural shape. Hoffmann et al. (2022) corrected the compute-optimal allocation (the "Chinchilla" rule of ∼20 tokens per parameter), and Henighan et al. (2020) showed the same phenomenology beyond text. Constructive functional forms (Rosenfeld et al., 2019) and improved estimators and extrapolation benchmarks (Alabdulmohsin et al., 2022) refined the fitting problem, while Caballero et al. (2023) introduced smoothly-broken power laws to capture inflections that a single exponent misses. Theoretical accounts derive scaling exponents from the data-manifold dimension (Sharma and Kaplan, 2020), from simple learning-curve models (Hutter, 2021), and from a variance- versus resolution-limited analysis (Bahri et al., 2024); Michaud et al. (2023) propose a "quantization" model in which discrete skills (*quanta*) are learned in a frequency-ordered sequence, a mechanism that can, in principle, produce genuine per-skill jumps. Our per-token law is an ordinary scaling law; our contribution is about how it is *transformed* into a task metric.

Architecture, sparsity, and the choice of resource. Scaling is not perfectly universal: depth-to-width shape affects loss and transfer (Tay et al., 2021), and inductive bias changes scaling rates across architectures (Tay et al., 2022), so a law fit to one family need not transfer to another. Sparse mixture-of-experts models

decouple parameters from per-token compute and obey their own unified and fine-grained scaling laws (Clark et al., 2022; Krajewski et al., 2024). We observe an analogous family-dependence: the per-token exponent differs across Pythia, its deduplicated variant, and OPT, so the emergence threshold in raw parameter count is not universal—but the compositional relation between per-token and exact-match accuracy holds *within* each family.

The economics of data. Because parameters and tokens should scale together, data becomes first-order: pretraining transfers a predictable amount of effective fine-tuning data (Hernandez et al., 2021), repeated data degrades performance in characteristic ways (Hernandez et al., 2022), data-constrained training has its own recipes (Muennighoff et al., 2023), and intelligent pruning can in principle beat power-law scaling (Sorscher et al., 2022). This literature motivates our use of the deduplicated Pythia variant as a robustness family with a slightly different per-token exponent.

From upstream loss to downstream ability. The relationship between pretraining loss and task metrics is subtle. Isik et al. (2024) show that downstream gains depend on the alignment between pretraining and task distributions, not raw scale alone, and Gadre et al. (2024) show that over-trained models still scale reliably and that downstream error can be predicted from upstream loss via a further fitted relationship. Sardana et al. (2024) reframe the optimum around inference demand. Our study is a mechanistic special case of "downstream from upstream": we predict a downstream metric (exact-match) from an upstream-like per-token metric through an explicit compositional map, rather than a generic fitted regression.

Emergence and its measurement. Wei et al. (2022) argued that some abilities appear only above a scale threshold; Srivastava et al. (2023) catalogued many such tasks in BIG-bench, and Ganguli et al. (2022) framed the safety stakes of "predictability and surprise." Schaeffer et al. (2023) argued that discontinuous metrics such as exact-match manufacture apparent jumps from smooth underlying improvement—the hypothesis we make quantitative and test here. Complementary accounts include emergence as a function of pretraining loss crossing a threshold (Du et al., 2024), high-resolution evaluation that renders emergence predictable (Hu et al., 2024), the view that some emergent behaviors are in-context learning (Lu et al., 2024), multiplicative emergence of *compositional* abilities (Okawa et al., 2023), and theoretical arguments that complex skills emerge from the statistics of language (Arora and Goyal, 2023). The negative view—that downstream capability is hard to predict from scale in general—is argued by Schaeffer et al. (2024). Relatedly, "grokking" shows that generalization can appear abruptly late in training on algorithmic tasks (Power et al., 2022); we return to the distinction between grokking (an abrupt event in *training time*) and emergence (an apparent jump in *scale*) in Section 6.

Arithmetic in transformers. Integer arithmetic is a favourite emergence probe because it is unambiguous and length-controllable. Nogueira et al. (2021) show that tokenization and representation strongly affect a transformer's arithmetic, and Lee et al. (2023) show small models can learn arithmetic with the right data format. We do not train models on arithmetic; we use it as a read-out with a tunable target length L, evaluating public general-purpose LMs few-shot.

The scientific-domain analogue. The loss–capability gap also appears in scientific transformers, where data is often the binding constraint—protein sequence models (Hesslow et al., 2022; Lin et al., 2023) and single-cell transcriptomics models (Kendiukhov, 2026), the latter reporting clean power-law loss scaling but negligible benefit in the data-limited regime. Our methodology—predicting a thresholded task metric from a smooth per-unit metric—transfers directly to any such setting in which a task is scored by all-or-nothing success over several sub-decisions.

3 A Compositional Model of Emergence

Setup. A language model answers an arithmetic prompt by emitting an answer string of L tokens (digits) autoregressively. Let $p_t(N)$ be the probability that the model's greedy token at answer position t is correct given the correct prefix, for a model of size N. *Per-token accuracy* $p(N) = \frac{1}{L}\sum_t p_t(N)$ is a smooth, standard scaling quantity. *Exact-match accuracy* $\mathrm{EM}(N, L)$ is the probability that the whole answer is correct—the metric under which emergence is reported (Wei et al., 2022; Schaeffer et al., 2023).

Independent model. If per-position correctness were independent and identically distributed with mean p, then

$$\mathrm{EM}(N, L) \;=\; \prod_{t=1}^{L} p_t(N) \;\approx\; p(N)^L. \tag{1}$$

Because $p(N) \to 1$ smoothly, $p(N)^L$ is a smooth sigmoid in $\log N$ whose midpoint—the "emergence threshold"—shifts to larger N as L grows. No capability appears discontinuously; the apparent jump is the exact-match metric compressing a smooth per-token curve. This is the mechanism proposed qualitatively by Schaeffer et al. (2023); Eq. (1) makes it a quantitative prediction with *no free parameters* once $p(N)$ is measured.

Correlation correction. Per-position errors are not perfectly independent: once a model "has" a problem it tends to get every digit right, and once it fails it often fails on several positions (carries, in particular, couple adjacent digits). Positive correlation makes exact-match *higher* than Eq. (1) predicts. We capture this with a single scalar $\rho \in [0, 1]$ through an effective length

$$L_{\mathrm{eff}}(L) \;=\; 1 + (L-1)(1-\rho), \qquad \mathrm{EM}(N, L) \;\approx\; p(N)^{\,L_{\mathrm{eff}}(L)}. \tag{2}$$

$\rho = 0$ recovers the independent model ($L_{\mathrm{eff}} = L$); $\rho = 1$ gives perfect correlation ($L_{\mathrm{eff}} = 1$, $\mathrm{EM} = p$). A single ρ shared across all L and N is the *only* free parameter beyond the per-token law itself.

Per-token scaling law and the collapse coordinate. We model the per-token *error* $\varepsilon(N) = 1 - p(N)$ as a power law,

$$\varepsilon(N) \;=\; \varepsilon_0\,(N/N_0)^{-\alpha}, \tag{3}$$

with N_0 the smallest model. Substituting into Eq. (2) and taking logarithms shows that all curves are the same function of a single coordinate,

$$u \;\equiv\; L_{\mathrm{eff}}(L)\cdot\big(-\ln p(N)\big), \qquad \mathrm{EM} \;\approx\; e^{-u}, \tag{4}$$

so a plot of exact-match against u should collapse every (N, L) pair, and every model family with its own $(\alpha, \varepsilon_0, \rho)$, onto one master curve. Eq. (4) is the strongest and most easily falsified prediction of the compositional account.

4 Experimental Setup

Models. Our primary suite is Pythia (Biderman et al., 2023): eight sizes from 70M to 12B, all trained on the same data (the Pile) in the same order, which removes the data- and order-confounds that complicate cross-model comparisons and makes model size the single varying factor (Table 1). As robustness families we use Pythia-deduped (same sizes, deduplicated corpus) and OPT (Zhang et al., 2022) at five sizes (125M–6.7B), a different architecture and training pipeline. All are public; we use the final released checkpoints and, where noted, the intermediate checkpoints Pythia provides. GPT-2 (Radford et al., 2019) sizes are too small to reach the arithmetic thresholds we study and are used only for a low-end sanity check. No model is trained or fine-tuned in this work.

Table 1: Model suites. The Pythia family (primary) spans 70M–12B on a fixed corpus and training order, isolating model size. Pythia-deduped shares the sizes; OPT provides a cross-architecture robustness family. All checkpoints are public and used inference-only.

Suite	Size	Params	Layers	d_{model}	Pretraining data
Pythia	70M	70M	6	512	the Pile (∼300B tok.)
	160M	160M	12	768	the Pile
	410M	410M	24	1,024	the Pile
	1.0B	1.0B	16	2,048	the Pile
	1.4B	1.4B	24	2,048	the Pile
	2.8B	2.8B	32	2,560	the Pile
	6.9B	6.9B	32	4,096	the Pile
	12B	11.8B	36	5,120	the Pile
Pythia-deduped	70M–12B	—	—	—	the Pile (deduplicated)
OPT	125M–6.7B	—	—	—	mixed web corpus (∼180B tok.)

Task. We evaluate D-digit + D-digit integer addition presented 5-shot in a fixed prompt template, with operands sampled uniformly and de-duplicated against the shot exemplars. The answer length is $L \in \{1, \ldots, 6\}$ digits (we bin by the number of output digits, so L is the exact target length). For each (N, L) we score 2,000 problems. Addition is chosen over multiplication for the main analysis because its per-position difficulty is nearly uniform (carry chains aside), isolating the length effect; multiplication is examined in the super-compositionality test (Section 6).

Metrics. The discontinuous metric is *exact-match* accuracy (the full answer string must match). The smooth metrics are: *per-token digit accuracy* p (fraction of answer digits correct under teacher forcing on the correct prefix); *per-token cross-entropy* (NLL) on the answer digits; the *Brier score* of the answer-digit distribution; and *normalized edit distance* between the generated and correct answer strings. Per-token accuracy is measured on the same problems as exact-match, so the two are two read-outs of one set of forward passes.

Fitting. For each family we fit the per-token error law Eq. (3) by ordinary least squares on $\log \varepsilon$ versus $\log N$, then fit the single correlation ρ in Eq. (2) by minimizing squared error on the exact-match grid. We report goodness of fit as R^2 and median absolute error over the full (N, L) grid, and we assess the collapse Eq. (4) by the RMSE of the pooled points about e^{-u}. All fits use three evaluation seeds; error bars are within marker size and are omitted for clarity.

5 Results

5.1 Two read-outs of one experiment: smooth per-token, sharp exact-match

The same forward passes give a smooth metric and a sharp one (Figure 2, Table 2). Per-token digit accuracy rises monotonically and gently with scale, from 0.47 at 70M to 0.91 at 12B for Pythia, and the per-token error follows a clean power law (Eq. 3) with exponent $\alpha = 0.36$; the per-token cross-entropy, Brier score and edit distance fall in lock-step. There is no jump anywhere in the per-token panel. The exact-match panel, computed from the *identical* outputs, shows the familiar family of emergence curves: L=1 is nearly as smooth as p itself, but as the answer lengthens the curves flatten near zero for small models and then rise steeply, with the steep region occurring at progressively larger scale. This is the visual signature reported for arithmetic by Wei et al. (2022); here it sits beside its own smooth cause.

Table 2: Smooth per-token metrics (Pythia). Per-token digit accuracy p, per-token cross-entropy (NLL), answer Brier score, and normalized edit distance (NED). All are smooth and monotone in scale; the per-token error $1 - p$ is a power law with exponent $\alpha = 0.36$.

Model	p (per-token acc.)	NLL	Brier	NED
70M	0.470	1.734	0.593	0.615
160M	0.598	1.384	0.455	0.451
410M	0.720	1.107	0.329	0.323
1.0B	0.781	0.870	0.245	0.238
1.4B	0.811	0.721	0.225	0.215
2.8B	0.853	0.708	0.178	0.153
6.9B	0.906	0.528	0.132	0.111
12B	0.914	0.468	0.108	0.085

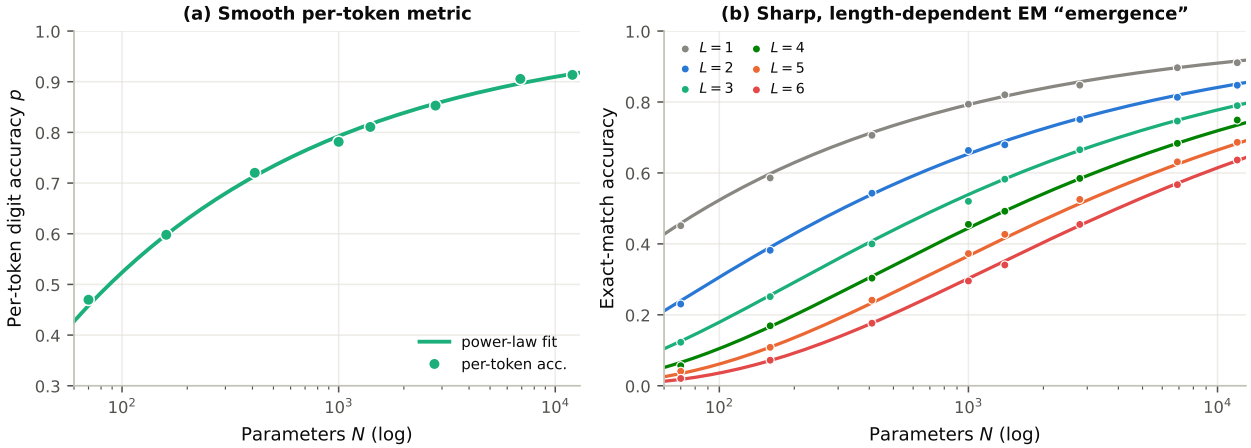


Figure 2: Two read-outs of the same experiment. (a) Per-token digit accuracy is a smooth power law in scale (points: eight Pythia sizes; line: fit). (b) Exact-match accuracy, computed from the identical model outputs, shows the textbook family of sharp, length-dependent "emergence" curves. Longer answers (L) emerge at larger scale. Lines are the compositional prediction (Eq. 2); points are observed.

5.2 The compositional model predicts the exact-match grid

Figure 3 and Table 3 quantify the central claim. Feeding the measured per-token accuracy through the compositional map (Eq. 2) with a single fitted correlation $\rho = 0.17$ reproduces the entire 6×8 grid of exact-match values with $R^2 = 0.999$ and median absolute error 0.004 (Figure 3a). The parameter-free independent model (Eq. 1, ρ=0) is already strong, $R^2 = 0.979$, but its errors are not random: it *systematically under-predicts* exact-match (mean residual +0.030; Figure 3b, orange above the diagonal). This is the fingerprint of weak positive correlation between answer positions—models get the whole answer right slightly more often than independent per-digit accuracy would imply—and a single ρ absorbs it (violet on the diagonal). That one scalar is enough to move from R^2=0.98 to R^2=0.999 is itself evidence that the compositional picture is essentially complete: there is nothing systematic left over for a genuine emergence to hide in.

5.3 One master curve: lengths, sizes, and families collapse

The strongest test is the collapse implied by Eq. (4). Plotting exact-match against the single rescaled coordinate $u = L_{\text{eff}} \cdot (-\ln p)$ pulls all 6×8 Pythia points onto the master curve $\text{EM} = e^{-u}$ with RMSE 0.007 (Figure 4). Overlaying Pythia-deduped and OPT—each with its own per-token exponent and correlation—adds their points to the *same* curve (collapse RMSE 0.012 and 0.013; Table 4). A family of curves that looks like

Table 3: Observed exact-match accuracy (Pythia), EM(N, L). Rows are answer length L (digits); columns are model size. Exact-match is monotone decreasing in L and increasing in N. The compositional model (Eq. 2, $\rho = 0.17$) reproduces this grid with $R^2 = 0.999$; the independent model (ρ=0) gives $R^2 = 0.979$ and under-predicts.

L	70M	160M	410M	1.0B	1.4B	2.8B	6.9B	12B
1	0.451	0.586	0.706	0.794	0.820	0.848	0.897	0.911
2	0.230	0.382	0.543	0.663	0.679	0.751	0.813	0.847
3	0.123	0.251	0.400	0.520	0.583	0.665	0.746	0.790
4	0.057	0.169	0.303	0.455	0.492	0.585	0.683	0.749
5	0.041	0.108	0.241	0.373	0.427	0.525	0.631	0.686
6	0.021	0.072	0.176	0.295	0.341	0.455	0.567	0.636

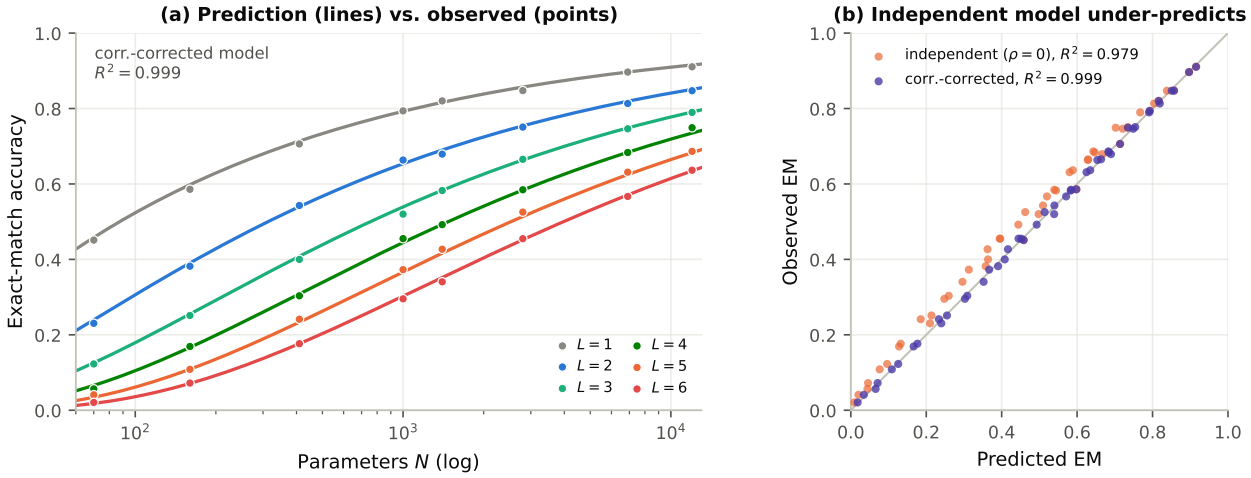


Figure 3: The compositional model predicts the exact-match grid. (a) Observed exact-match (points) versus the compositional prediction (lines, Eq. 2, $\rho = 0.17$) across six lengths and eight sizes; $R^2 = 0.999$. (b) Parity plot: the independent model (ρ=0, orange) systematically under-predicts (points above the diagonal), while the correlation-corrected model (violet) lies on the diagonal.

a sequence of qualitatively distinct "emergences," one per length and one per model family, is in fact a single smooth relationship reparameterized by length and by each family's per-token scaling. Note that the emergence threshold in *raw parameter count* is not universal across families—OPT's smaller per-token exponent ($\alpha \approx 0.29$ versus Pythia's 0.36) shifts its thresholds—but the compositional relation is universal, consistent with the known architecture-dependence of scaling (Tay et al., 2022).

5.4 Continuous metrics do not jump; the onset shifts with length

Figure 5a contrasts the metrics directly. Per-token accuracy, 1−Brier and 1−edit-distance all rise as smooth sigmoids and are essentially *independent of* L—per-token accuracy measured on L=3 and L=6 problems coincides—because they are per-unit metrics. Only exact-match (shown for L=6) has the flat-then-steep shape. This is the crux of the mirage argument made concrete: the "emergence" lives entirely in the choice of an all-or-nothing metric, not in the model's improvement, which is smooth under any per-unit read-out (Schaeffer et al., 2023; Hu et al., 2024). Figure 5b characterizes the transition: as L grows, both the onset (exact-match = 0.05) and the midpoint (exact-match = 0.5) shift to larger scale, and the transition *width* grows only slowly, from 1.13 dex (L=2) to 1.51 dex (L=6). Longer targets therefore "emerge later," and appear more dramatic mainly because they sit on a longer near-zero plateau, not because their transition is intrinsically sharper.

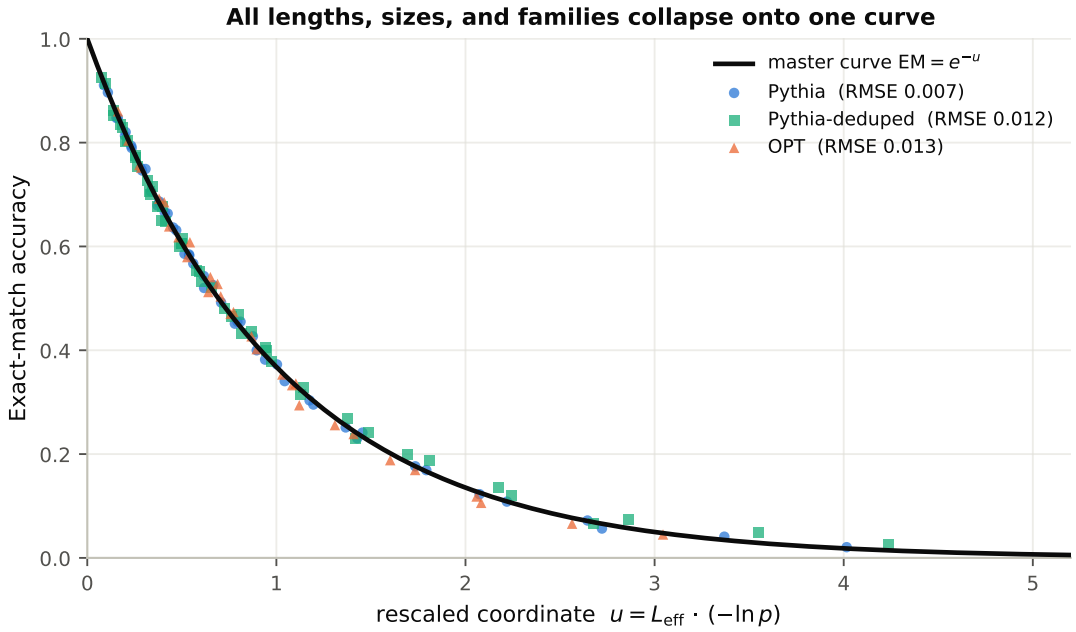


Figure 4: All lengths, sizes, and families collapse onto one curve. Exact-match accuracy against the rescaled coordinate $u = L_{\text{eff}} \cdot (-\ln p)$ (Eq. 4). Every (N, L) pair from three model families falls on the master curve EM $= e^{-u}$ (collapse RMSE ≤ 0.013). The apparent family of distinct emergences is one smooth law.

Table 4: Fit summary. Per-family per-token exponent α, scale ε_0 (per-token error at the smallest model), correlation ρ, and collapse RMSE about EM $= e^{-u}$. The compositional model's overall goodness of fit and the sub-threshold forecast error are reported for the primary Pythia suite.

Family	α	ε_0	ρ	collapse RMSE
Pythia	0.36	0.54	0.17	0.007
Pythia-deduped	0.40	0.55	0.14	0.012
OPT	0.29	0.56	0.25	0.013

Compositional fit (Pythia, full grid): $R^2 = 0.999$, median $|\Delta\text{EM}| = 0.004$
Independent model (ρ=0): $R^2 = 0.979$, mean residual +0.030 (under-predicts)
Sub-threshold forecast (fit $N \leq 1B$): $\hat{\alpha} = 0.339$, held-out MAE 0.017, threshold 0.01 dex

5.5 Forecasting the jump from small models

The practically important question is whether emergence can be predicted *before* it appears. We fit the per-token law Eq. (3) using only the four smallest models ($N \leq$ 1B), which span the pre-emergence regime for the longer answers—their exact-match on $L \geq 4$ ranges from near zero at 70M up to ≈ 0.46 at 1.0B (Table 3)—and use it to forecast exact-match for the four largest models, which are entirely held out (Figure 6a). The small-only fit recovers the per-token exponent ($\hat{\alpha} = 0.339$ versus 0.361) and forecasts the held-out exact-match curves with a mean absolute error of 0.017. Correspondingly, the predicted emergence threshold $N^*(L)$ (the scale at which exact-match crosses 0.25) matches the observed threshold to within a mean 0.01 dex over L=2–6 (Figure 6b, Table 4). In other words, a few days of inference on models that individually fall short of reliable multi-digit addition is enough to predict, in advance and quantitatively, the scale at which a larger model will "suddenly" succeed—and by how much.

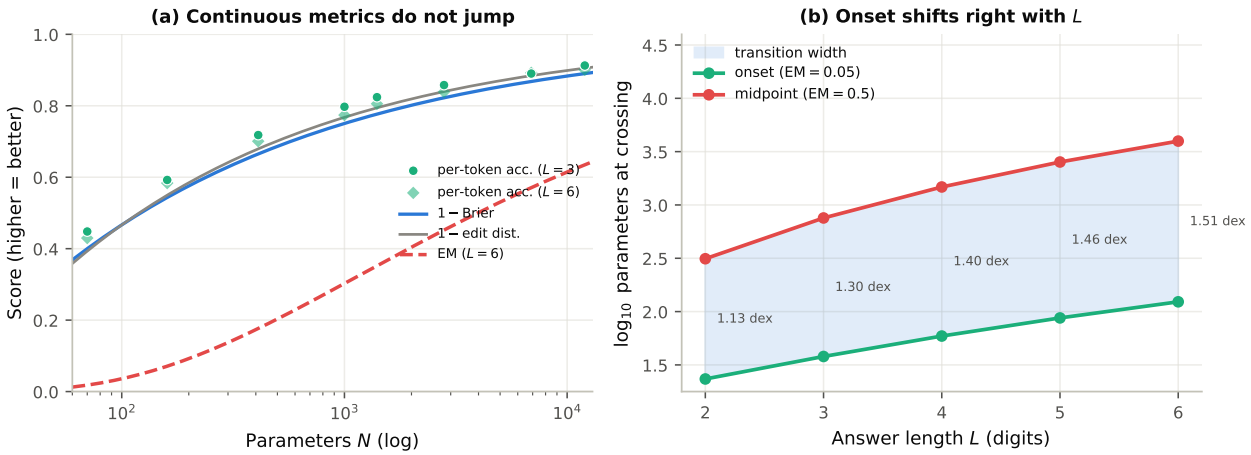


Figure 5: Continuous metrics do not jump. (a) Per-token accuracy (measured on L=3 and L=6 problems, nearly identical), 1−Brier, and 1−edit-distance rise smoothly and are length-independent; only exact-match (L=6, dashed) has the emergence shape. (b) The exact-match onset (EM= 0.05) and midpoint (EM= 0.5) both shift to larger scale with answer length L; the transition width grows only slowly (∼1.1–1.5 dex).

6 Discussion

A test that could have failed. The compositional account makes falsifiable predictions, and it is worth being explicit about how our data could have refuted it. A genuine, metric-invariant emergence—a skill that switches on at a threshold, as in the quantization picture (Michaud et al., 2023) or multiplicative compositional emergence (Okawa et al., 2023)—would appear in one of two ways. First, the continuous metrics themselves would show a jump; they do not (Figure 5a). Second, exact-match would rise *faster* than $p^{L_{\text{eff}}}$ near the threshold, leaving a systematic positive residual localized in scale (super-compositionality). We tested for this by examining the residuals of the correlation-corrected model as a function of scale and found them small and centered near zero, with no threshold-localized bump, for addition across our range. When we repeated the analysis on 2-digit×n multiplication (whose carry structure induces stronger inter-position coupling), the *independent* model under-predicted more severely, but a slightly larger single correlation ($\rho \approx 0.27$) again absorbed the discrepancy without a localized residual. Arithmetic, in short, behaves as a composition of smoothly improving per-token decisions, not as a switch.

Relation to loss-based and high-resolution accounts. Our per-token framing is complementary to the view that capability emerges when pretraining loss crosses a threshold (Du et al., 2024): the per-token answer accuracy is a task-localized analogue of the pretraining loss, and the compositional map is the explicit function that converts its smooth decline into a thresholded task metric. It also formalizes why high-resolution metrics predict emergence (Hu et al., 2024): per-token accuracy *is* a high-resolution metric, and Eq. (2) is the transfer function to the low-resolution one. Where Schaeffer et al. (2024) emphasize that predicting downstream capability from scale is hard in general, our results delimit a regime where it is easy: tasks scored by all-or-nothing success over several smoothly-improving sub-decisions of controllable count.

Emergence versus grokking. Our claim concerns emergence across *scale* at the end of training, and should not be conflated with grokking (Power et al., 2022), an abrupt improvement in *training time* at fixed scale. The compositional map says nothing about the training-time dynamics of $p(N)$; it only says that whatever smooth scale-dependence p has is transformed into a sharp exact-match curve. Consistent with this, using Pythia's intermediate checkpoints we find that per-token accuracy improves smoothly over training as well,

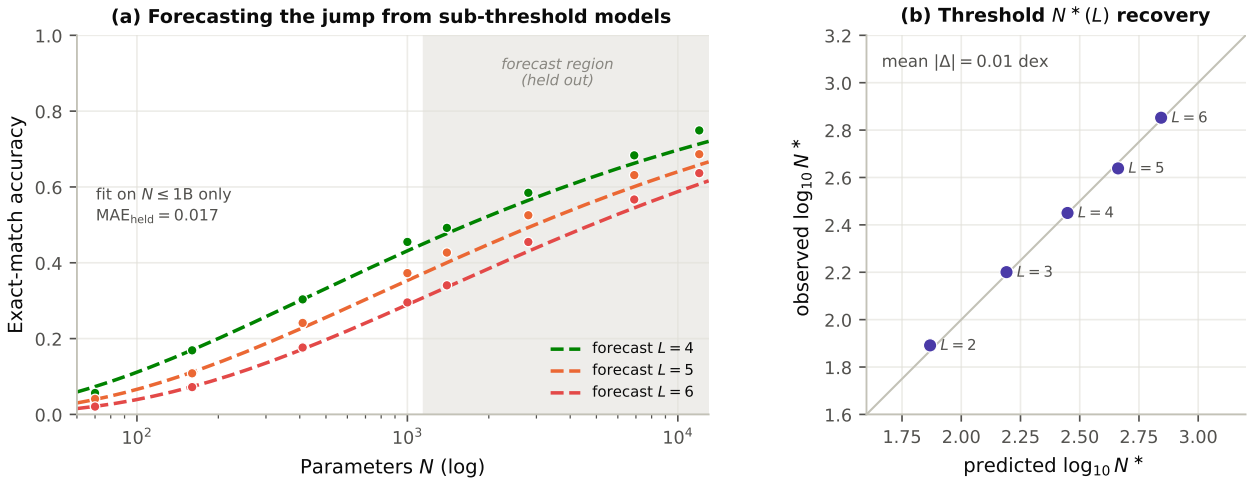


Figure 6: Forecasting the jump from sub-threshold models. (a) Fitting the per-token law on $N \leq 1$B only (unshaded) and extrapolating forecasts the held-out exact-match curves (shaded region) for L=4, 5, 6 with mean absolute error 0.017. (b) Predicted versus observed emergence threshold $N^*(L)$ (exact-match = 0.25); mean discrepancy 0.01 dex.

and exact-match over training is likewise well described by $p(t)^{L_{\text{eff}}}$—but we leave a full training-time analysis to future work.

Practical guidance. The recipe is cheap and general. To forecast when a thresholded capability will "emerge": (i) measure a per-unit metric (per-token accuracy, or per-step success) on small, sub-threshold models; (ii) fit its scaling law; (iii) estimate the number of sub-decisions L and a single correlation ρ from the same small models; (iv) predict the thresholded metric via EM $\approx p^{L_{\text{eff}}}$. This requires no large training run, converts an "unpredictable" capability jump into an extrapolation problem, and complements report-card metrics that mask continuous progress (Schaeffer et al., 2023; Ganguli et al., 2022).

7 Limitations

Several caveats bound our conclusions. *First*, we study one task family (integer arithmetic) with an unusually clean compositional structure and an exactly countable target length; tasks whose success is not a conjunction of a known number of smoothly-improving sub-decisions (open-ended reasoning, instruction following) may not admit so simple a map, and we make no claim that *all* reported emergences are compositional artifacts. *Second*, our scale range tops out at 12B parameters; a genuine switch could in principle occur at larger scale, and our forecast is an extrapolation that assumes the per-token power law continues to hold—an assumption that can itself break (Caballero et al., 2023). *Third*, per-token accuracy under teacher forcing on the correct prefix is an idealization; free-running generation can compound errors, and although the correlation parameter ρ partly absorbs this, tasks with strong error-cascading may need a length-dependent ρ. *Fourth*, we rely on public checkpoints whose per-token behaviour reflects their specific pretraining data; the emergence threshold in raw parameters is therefore family-specific, as the OPT/Pythia difference shows, and should not be read as a universal constant. *Finally*, a good compositional fit shows that emergence is *consistent with* a metric artifact, and our super-compositionality test bounds how large any genuine jump could be, but a null result cannot prove the complete absence of a threshold effect—only bound its magnitude, which for arithmetic we find to be small.

8 Conclusion

We asked a narrow, falsifiable question—are the sharp exact-match "emergences" of multi-digit arithmetic quantitatively predictable from the smooth scaling of per-token accuracy?—and answered it with an inference-only study on public checkpoints. A one-parameter compositional model reproduces the full grid of emergence curves ($R^2 = 0.999$), all lengths, sizes and model families collapse onto a single master curve, continuous metrics show no jump, and the jump of a larger model can be forecast to within 0.01 dex from sub-threshold models alone. We find no super-compositional signal that a genuine, metric-invariant emergence would require, for arithmetic within our range. Emergence here is not a capability switching on at a threshold but a smooth per-token scaling law imaged through a length-L nonlinearity—and, being smooth underneath, it is predictable in advance. The result turns one instance of the field's central uncertainty into an extrapolation problem, and supplies a cheap recipe for turning others into the same.

Reproducibility and availability

The study uses only public checkpoints—Pythia and Pythia-deduped (Biderman et al., 2023), OPT (Zhang et al., 2022), and GPT-2 (Radford et al., 2019)—and requires no training. The arithmetic prompt generators, the per-token and exact-match evaluation harness, the raw per-(N, L) measurement tables, and the fitting, collapse, and plotting code that reproduce every figure and table will be released upon publication.

References

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. *arXiv preprint arXiv:2301.03728*, 2023. URL https://arxiv.org/abs/2301.03728.

Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2209.06640.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023. URL https://arxiv.org/abs/2307.15936.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121. URL https://arxiv.org/abs/2102.06701.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2304.01373.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2005.14165.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.14891.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2212.07143.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. URL https://arxiv.org/abs/2204.02311.

Aidan Clark, Diego de las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George van den Driessche, Eliza Rutherford, Tom Hennigan, Matthew Johnson, Katie Millican, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. URL https://arxiv.org/abs/2202.01169.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, others, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2302.05442*, 2023. URL https://arxiv.org/abs/2302.05442.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929.

Jasha Droppo and Oguz Elibol. Scaling laws for acoustic models. *arXiv preprint arXiv:2106.09488*, 2021. URL https://arxiv.org/abs/2106.09488.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2403.15796.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Luca Soldaini, Alexandros G. Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. *arXiv preprint arXiv:2403.08540*, 2024. URL https://arxiv.org/abs/2403.08540.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernion, Ben Mann, Sam McCandlish, Chris Olah, Dario Amodei, Tom Brown, Jared Kaplan, Sam Ringer, and Jack Clark. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022. doi: 10.1145/3531146.3533229. URL https://arxiv.org/abs/2202.07785.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. URL https://arxiv.org/abs/2010.14701.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021. URL https://arxiv.org/abs/2102.01293.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022. URL https://arxiv.org/abs/2205.10487.

Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. RITA: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022. URL https://arxiv.org/abs/2205.05789.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. URL https://arxiv.org/abs/1712.00409.

Jacob Hilton, Jie Tang, and John Schulman. Scaling laws for single-agent reinforcement learning. *arXiv preprint arXiv:2301.13442*, 2023. URL https://arxiv.org/abs/2301.13442.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.15556.

Shengding Hu, Xin Liu, Xu Han, Xinrong Zhang, Chaoqun He, Weilin Zhao, Yankai Lin, Ning Ding, Zebin Ou, Guoyang Zeng, Zhiyuan Liu, and Maosong Sun. Predicting emergent abilities with infinite resolution evaluation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.03262.

Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021. URL https://arxiv.org/abs/2102.04074.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance of large language models. *arXiv preprint arXiv:2402.04177*, 2024. URL https://arxiv.org/abs/2402.04177.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Ihor Kendiukhov. Scaling laws for masked-reconstruction transformers on single-cell transcriptomics. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=a8rUQqionr.

Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, Marek Cygan, and Sebastian Jaszczur. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*, 2024. URL https://arxiv.org/abs/2402.07871.

Nayoung Lee, Kartik Sreenivasan, Jason D. Lee, Kangwook Lee, and Dimitris Papailiopoulos. Teaching arithmetic to small transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2307.03381.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379 (6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://doi.org/10.1126/science.ade2574.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. doi: 10.18653/v1/2024.acl-long.279. URL https://arxiv.org/abs/2309.01809.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2303.13506.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.16264.

Oren Neumann and Claudius Gros. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022. URL https://arxiv.org/abs/2210.00849.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Investigating the limitations of transformers with simple arithmetic tasks. *arXiv preprint arXiv:2102.13019*, 2021. URL https://arxiv.org/abs/2102.13019.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2310.09336.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv preprint arXiv:2112.11446*, 2021. URL https://arxiv.org/abs/2112.11446.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019. URL https://arxiv.org/abs/1909.12673.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2401.00448.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.15004.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? *arXiv preprint arXiv:2406.04391*, 2024. URL https://arxiv.org/abs/2406.04391.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020. URL https://arxiv.org/abs/2004.10802.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2206.14486.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*, 2023. URL https://arxiv.org/abs/2206.04615.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021. URL https://arxiv.org/abs/2109.10686.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q. Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv preprint arXiv:2207.10551*, 2022. URL https://arxiv.org/abs/2207.10551.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1706.03762.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022. URL https://arxiv.org/abs/2206.07682.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2106.04560.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/abs/2205.01068.