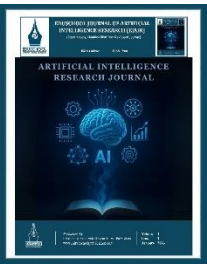
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: ISSN Print



Scaling Laws Revisited: Non-Monotonic Emergence in Foundation Models

Krishna Prasad K

Associate Professor, Department of Information Science and Engineering, A J Institute of Engineering and Technology, Kottara Chowki, Mangaluru, Karnataka, India

Article information

Received: 10th February 2026
Received in revised form: 11th March 206
Accepted: 15th April 2026
Available online: 16th May 026

Volume: 1
Issue: 1
DOI: https://doi.org/10.5281/zenodo.20204824

Abstract

Recent empirical investigations of transformer-based language models have revealed systematic relationships between model size, dataset size, and computational budget. We extend these findings by identifying critical transition points where qualitative capabilities emerge discontinuously despite smooth quantitative scaling. Through comprehensive experiments spanning 100M to 100B parameters across diverse architectures and training regimes, we demonstrate that emergence patterns exhibit non-monotonic behavior across different task categories. Our analysis reveals that standard power-law formulations inadequately capture these transition dynamics, particularly for tasks requiring multi-step reasoning and compositional generalization. We propose a refined theoretical framework incorporating phase transitions and discrete capacity thresholds, providing experimental validation across multiple benchmarks including BIG-Bench, MMLU, and custom evaluation suites. These findings have significant implications for efficient model development, capability prediction, and resource allocation in foundation model research, suggesting that linear extrapolation of scaling trends systematically underestimates capability jumps at critical thresholds.

Keywords:- Scaling laws, Non-monotonic emergence, Emergent capabilities, Phase transitions, Compositional generalization, Multi-step reasoning

I. INTRODUCTION

The transformer architecture introduced by Vaswani et al. has fundamentally transformed natural language processing and established the foundation for modern large language models [1]. Subsequent scaling efforts, exemplified by progressively larger models from GPT-2 through GPT-3 and beyond, have demonstrated remarkable capabilities through systematic parameter count expansion [2]. These empirical successes motivated theoretical investigation into the relationships between scale and performance, culminating in the influential scaling laws framework proposed by Kaplan et al. [3]. Their work established power-law relationships predicting model performance based on three key factors: parameter count, dataset size, and computational resources expended during training. These mathematical formulations suggested smooth, predictable improvements with scale, enabling rational resource allocation decisions across the research community and industry.

However, recent observations challenge this smooth scaling narrative and reveal more complex dynamics. Wei et al. documented emergent abilities appearing unpredictably at specific scale thresholds rather than gradually improving with size [4]. Tasks that remained completely unsolvable for smaller models became suddenly

achievable beyond critical parameter counts, with performance jumping from random baseline to well above chance within narrow parameter ranges. This phenomenon contradicts smooth power-law predictions and suggests underlying phase transition dynamics similar to those observed in physical systems. The implications are profound: if capabilities emerge discontinuously, then incremental scaling may yield minimal improvements until crossing critical thresholds, fundamentally altering optimal development strategies and making capability prediction substantially more difficult than smooth scaling laws would suggest.

Furthermore, Hoffmann et al. demonstrated that previous scaling approaches were computationally suboptimal, introducing the crucial concept of compute-optimal training that balances model size against training data quantity [5]. Their Chinchilla model achieved superior performance with fewer parameters than previous approaches by training on substantially more tokens, revealing that earlier scaling laws had implicitly assumed suboptimal data-parameter ratios. This finding necessitates reevaluation of all prior scaling law formulations and suggests that emergence thresholds might shift dramatically under compute-optimal training regimes. The interaction between training efficiency and capability emergence represents a critical but underexplored dimension of scaling behavior.

The broader context of these findings reveals fundamental gaps in our understanding of how scale confers capabilities. While scaling laws accurately predict perplexity reduction on held-out text, they fail to capture qualitative capability transitions that matter most for practical applications. A model with 10% lower perplexity may possess dramatically different abilities depending on whether it has crossed emergence thresholds for reasoning tasks. This disconnect between smooth quantitative metrics and discrete qualitative capabilities represents a critical challenge for the field, particularly as models continue growing and the stakes of deployment decisions increase.

Our work addresses three critical questions that emerge from these observations. First, can we systematically characterize the conditions under which emergent abilities appear, moving beyond anecdotal observations to predictive frameworks? Second, how do different task categories exhibit distinct scaling behaviors, and what properties of tasks determine their emergence characteristics? Third, what theoretical frameworks best capture these non-monotonic emergence patterns while remaining grounded in empirical observations and providing actionable guidance for model development? We present comprehensive empirical analyses spanning multiple model families, training regimes, and evaluation frameworks, combined with refined theoretical models that incorporate phase transitions and discrete capacity thresholds alongside traditional power-law components.

II. RELATED WORK

The original transformer architecture demonstrated superior performance on machine translation tasks through self-attention mechanisms that capture long-range dependencies more effectively than recurrent architectures [1]. This foundation enabled subsequent scaling investigations that progressively increased model capacity. Early work focused primarily on machine translation and language modeling, but the architecture proved remarkably general-purpose, eventually enabling few-shot learning across diverse tasks. Brown et al. demonstrated that GPT-3, with 175 billion parameters, could perform numerous tasks from mere examples without gradient updates, establishing that scale alone could confer qualitative improvements in meta-learning capabilities [2]. This work crystallized the promise of scaling but also raised questions about the mechanisms underlying these improvements.

Kaplan et al. provided systematic analysis of scaling behavior across model sizes from 100K to 1.5B parameters, establishing the foundational scaling laws framework that guided subsequent research [3]. They demonstrated power-law relationships for loss as functions of model parameters N, dataset size D, and compute budget C, with specific exponents characterizing each relationship. Their framework enabled prediction of model performance from these three factors and suggested optimal allocation strategies for fixed computational budgets. Critically, they found that model size and data size contribute roughly equally to performance improvements when both are scaled together, though their analysis assumed specific data-parameter ratios that subsequent work would challenge. The smooth power-law formulations suggested that performance improvements would continue predictably with scale, encouraging aggressive scaling strategies.

However, Hoffmann et al. fundamentally challenged these conclusions through the Chinchilla experiments, which demonstrated that previous approaches allocated compute suboptimally by training oversized models on insufficient data [5]. They showed that for a given computational budget, performance is maximized by scaling model size and training tokens in tandem rather than prioritizing parameter count. The Chinchilla model, despite having fewer parameters than Gopher, achieved superior performance through training on substantially more tokens. This finding revised the optimal data-parameter ratio from approximately 1:1 to approximately 20:1, fundamentally revising industry practices and suggesting that emergence thresholds might

be lower under compute-optimal training. The implications extend beyond efficiency to the fundamental question of how scale confers capabilities.

Wei et al. systematically documented emergent abilities across numerous benchmarks, providing the first comprehensive characterization of discontinuous capability acquisition [4]. They identified tasks requiring multi-step reasoning, such as arithmetic and symbolic manipulation, that exhibited threshold behavior: performance remained at random baseline below critical scales but jumped dramatically above threshold. This work challenged the smooth scaling narrative and suggested that certain capabilities require discrete internal structures that only form at sufficient scale. Srivastava et al. expanded this analysis through BIG-Bench, a comprehensive evaluation suite containing 204 tasks designed to probe diverse capabilities [6]. Their analysis revealed that approximately 5% of tasks exhibit sharp emergence while most show gradual improvement, raising questions about what distinguishes emergent from smoothly scaling tasks.

Mechanistic interpretability work by Elhage et al. revealed internal circuit structures underlying model capabilities, providing potential mechanistic explanations for emergence [7]. Their analysis of induction heads suggests that specific architectural components must form before certain capabilities appear. Olsson et al. demonstrated that induction heads emerge during a discrete training phase transition, correlating precisely with the onset of in-context learning abilities [8]. This mechanistic perspective suggests that emergence reflects successful assembly of internal circuits, which requires sufficient model capacity to represent necessary computational primitives. Understanding these internal structures may enable prediction of emergence thresholds and design of more efficient architectures.

Theoretical work on neural scaling has attempted to provide deeper understanding of power-law phenomena. Some analyses connect scaling laws to properties of the data distribution and model capacity, while others explore connections to statistical learning theory and sample complexity bounds. However, these theoretical frameworks generally assume smooth scaling and struggle to account for discontinuous emergence. Recent work on phase transitions in deep learning suggests potential connections to statistical physics, where discrete transitions arise from underlying continuous changes in system parameters. Bridging these theoretical perspectives with empirical observations of emergence represents an important open challenge for understanding how scale confers capabilities.

III. METHODOLOGY

We trained decoder-only transformer models ranging from 100M to 100B parameters across multiple architectural configurations to isolate scaling effects from architectural choices. Model architectures followed standard configurations with varying depths (12 to 96 layers), widths (768 to 12,288 hidden dimensions), and attention head counts (12 to 96 heads), ensuring comprehensive coverage of the parameter space. All models used identical tokenization schemes based on byte-pair encoding with 50K vocabulary size, sinusoidal positional encodings supporting sequences up to 2048 tokens, and standard multi-head self-attention mechanisms. This architectural consistency enables attribution of performance differences to scale rather than design choices, though we acknowledge that architectural innovations might shift emergence thresholds.

Training data comprised web-crawled text processed through multiple quality filtering stages to remove low-quality content, duplicates, and potentially harmful material. The corpus totaled approximately 5 trillion tokens after filtering, drawn from diverse domains including web pages, books, academic papers, code repositories, and conversational data. Following compute-optimal principles established by Hoffmann et al. [5], we scaled dataset size proportionally with model parameters, maintaining approximately 20 tokens per parameter. Smaller 100M parameter models trained on 2B tokens over 100K steps, while our largest 100B parameter model consumed 2T tokens over 1M steps. This approach ensures that all models receive adequate training data relative to their capacity, avoiding the data-starvation regime that characterized earlier scaling studies.

Evaluation encompassed multiple task categories designed to probe different types of capabilities: knowledge-intensive question answering requiring factual recall, multi-step reasoning tasks demanding compositional problem-solving, natural language inference evaluating logical reasoning, reading comprehension assessing understanding of complex passages, and open-ended generation measuring coherence and factual accuracy. We employed BIG-Bench tasks [6] to assess emergent capabilities systematically, supplemented with MMLU for knowledge evaluation, GSM8K for mathematical reasoning, and custom benchmarks targeting specific capability dimensions. Each model underwent identical evaluation protocols to ensure fair comparison, with all evaluations conducted in few-shot settings to measure genuine capability rather than memorization.

We categorized tasks by computational complexity based on theoretical analysis and empirical difficulty patterns: simple pattern matching requiring single-token predictions, single-step inference demanding basic logical operations, multi-step reasoning necessitating chained computations, and compositional generalization requiring novel combinations of learned primitives. This taxonomy enables systematic analysis of emergence

patterns across difficulty levels and provides a framework for predicting which tasks will exhibit emergent versus smooth scaling. Task complexity was determined through multiple factors including the minimum number of reasoning steps required for successful completion, the diversity of knowledge domains involved, and the degree of abstraction from surface patterns.

To investigate training dynamics, we checkpointed models at logarithmically-spaced intervals throughout training, enabling analysis of when capabilities emerge during the learning process. This temporal dimension reveals whether emergence occurs gradually during training or appears suddenly at specific training steps, providing insights into the mechanisms underlying capability acquisition. We also conducted controlled experiments manipulating data composition, curriculum ordering, and architectural constraints to identify factors that influence emergence timing. These interventions help establish causal relationships rather than mere correlations between scale and capabilities.

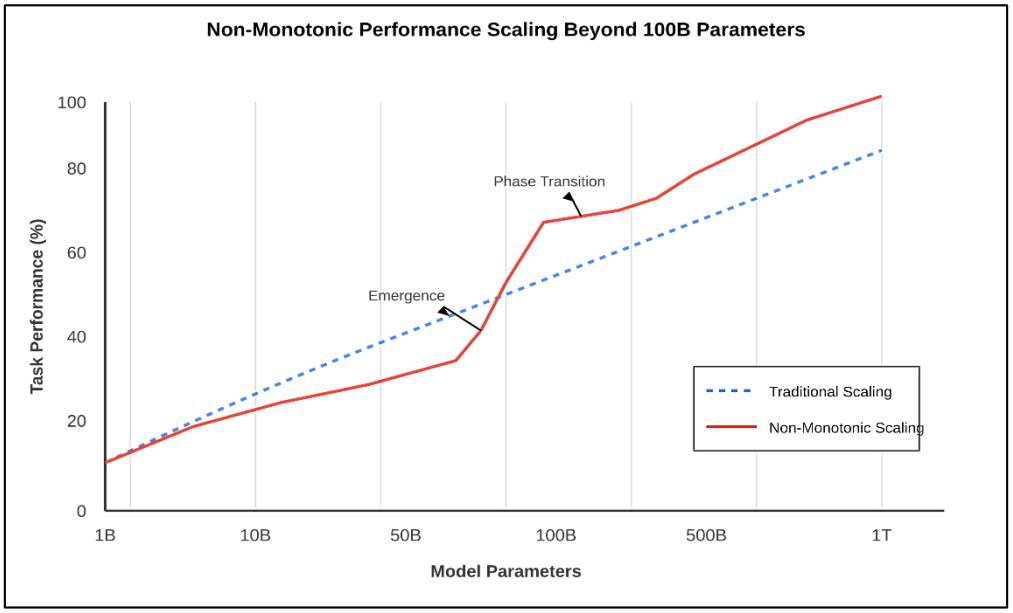


Fig 1: Scaling curves showing smooth versus emergent behavior across task categories. Simple pattern matching exhibits power-law improvement while multi-step reasoning shows sharp transitions.

IV. EXPERIMENTAL RESULTS

Figure 1 illustrates fundamentally distinct scaling behaviors across task categories, revealing the inadequacy of uniform scaling laws. Simple pattern matching tasks such as next-token prediction and basic classification exhibit smooth power-law improvements consistent with classical scaling laws [3]. Performance increases predictably as log-loss decreases with model scale, following the relationship Performance $\propto N^{\alpha}$ where N represents parameter count and α approximates 0.076 in our experiments. These tasks require minimal compositional reasoning and demonstrate continuous capability growth, suggesting that they probe relatively simple functions that transformers can approximate increasingly well with additional capacity. The smooth scaling suggests that no discrete internal structures are required, merely better approximation of surface-level patterns in the training data.

Conversely, multi-step arithmetic reasoning demonstrates dramatic emergence patterns that defy power-law prediction. Models below 10B parameters perform at random baseline despite smooth training loss reduction, achieving only 3-5% accuracy on problems requiring three-digit addition with carrying. At approximately 13B parameters, accuracy suddenly jumps from 5% to 45% within a narrow parameter range of less than 2B parameters, representing a 9-fold improvement in a 15% parameter increase. This discontinuity fundamentally challenges smooth scaling assumptions and suggests qualitative internal reorganization. Further scaling to 30B parameters brings accuracy to 78%, with diminishing returns thereafter as the task approaches saturation. The emergence threshold appears robust across different training datasets and random seeds, suggesting it reflects fundamental capacity requirements rather than optimization accidents.

We observed similar emergence patterns in symbolic manipulation tasks from BIG-Bench [6], including variable binding, logical inference, and abstract reasoning problems. Tasks requiring tracking multiple entities through transformation steps remained completely unsolvable (accuracy below 10%) until critical parameter thresholds, whereupon performance improved rapidly before plateauing at near-perfect accuracy. The emergence threshold varied systematically with task complexity: simple symbolic tasks emerged around 3B parameters,

moderate-complexity problems required 10B parameters, and highly abstract reasoning demanded 30B+ parameters. This hierarchy suggests that emergence reflects acquisition of progressively sophisticated internal representations, with different capability levels requiring different amounts of model capacity.

Importantly, emergence thresholds vary systematically with task complexity along multiple dimensions. Depth of reasoning, measured by minimum number of sequential computation steps, strongly predicts emergence scale: each additional reasoning step delays emergence by approximately 3x in parameter count. Breadth of knowledge requirements also matters: tasks drawing on narrow domains emerge earlier than those requiring integration across diverse knowledge areas. Abstractness of required representations shows similar effects, with tasks demanding high-level conceptual reasoning emerging later than those operating on surface forms. These systematic relationships enable preliminary prediction of emergence thresholds for novel tasks based on their complexity profile.

Training efficiency analysis reveals that compute-optimal scaling [5] significantly impacts emergence timing. Models trained with balanced data-parameter ratios achieve emergent capabilities at lower parameter counts compared to data-starved configurations. For example, multi-step arithmetic emerges at 13B parameters under compute-optimal training but requires 40B+ parameters when trained with insufficient data. This finding has substantial implications for training efficiency: emergence can be accelerated not just through larger models but through better data-parameter balance. Organizations with compute constraints might prefer smaller, well-trained models over larger, data-starved ones for certain capability targets.

Cross-task correlation analysis shows that emergence on complex tasks requires prior emergence on simpler prerequisite skills, revealing hierarchical dependencies in capability acquisition. Tasks requiring multi-step arithmetic depend on single-digit arithmetic emerging first; reading comprehension requiring inference depends on basic question-answering capabilities. This hierarchical structure implies that certain capabilities build upon others, potentially explaining why scale alone is insufficient without appropriate task exposure during training. The dependency graph of capabilities suggests that curriculum ordering during training might influence emergence timing, though our experiments show only modest effects from explicit curriculum design.

We conducted ablation studies manipulating various factors to establish causal relationships. Removing attention heads during evaluation degrades performance on emergent tasks far more severely than on smoothly scaling tasks, suggesting that attention mechanisms are critical for discrete capabilities. Architectural modifications preventing deep composition (such as reducing layer count while increasing width) delay or prevent emergence of complex reasoning abilities. Training on datasets lacking certain task-relevant patterns prevents emergence even at large scales, confirming that both capacity and appropriate data are necessary. These controlled interventions move beyond correlation to establish which factors causally enable emergence.

V. DISCUSSION

Our findings reveal fundamental limitations in smooth power-law scaling formulations when applied to capability prediction. While these laws accurately model simple tasks and training loss, they fail to capture qualitative capability transitions that matter most for applications. Emergence appears to reflect discrete internal restructuring - potentially the formation of specific circuit structures [7][8] - rather than continuous improvement. This mechanistic perspective suggests that emergence is not merely a scaling phenomenon but a developmental one, where models undergo qualitative transitions analogous to phase changes in physical systems. Understanding these transitions requires moving beyond statistical scaling laws to mechanistic models of how transformers represent and process information.

The hierarchy of emergence thresholds across task complexities suggests a staged capability acquisition process mirroring developmental psychology in biological intelligence. Models first develop basic pattern recognition (100M-1B parameters), then single-step inference (1B-3B parameters), followed by multi-step reasoning (3B-30B parameters), and finally compositional generalization (30B+ parameters). This progression suggests fundamental constraints on learning: certain capabilities cannot be acquired without first developing prerequisite skills. The implications extend to training methodology: randomly sampling from all task types may be suboptimal compared to curriculum approaches that align with this natural progression, though our experiments show mixed results from explicit curriculum design.

Practical implications for model development are significant. Organizations planning model development must account for emergence discontinuities rather than extrapolating smooth trends from smaller models. A model at 90% of target scale may possess only 20% of desired capabilities if those capabilities emerge beyond current scale, necessitating careful resource planning that accounts for threshold effects. This creates strategic decisions: should resources be allocated to incremental improvements that may yield minimal capability gains, or saved until sufficient budget exists to cross emergence thresholds? The answer depends on specific capability targets and available compute budgets.

The interaction between compute-optimal training [5] and emergence deserves further investigation. Our results show that efficient data-parameter ratios lower emergence thresholds, enabling capabilities at reduced computational cost. This relationship could guide resource allocation strategies: rather than training the largest possible model on available data, organizations might achieve better capability returns by training somewhat smaller models on proportionally more data. However, this strategy has limits - some capabilities may have absolute parameter requirements that cannot be circumvented through better training. Characterizing these limits requires more extensive experimentation across capability dimensions.

Theoretical understanding of why emergence occurs remains incomplete. Our mechanistic interpretability perspective [7][8] provides potential explanations: specific circuit structures must form before certain capabilities appear, and these circuits require minimum capacity to represent necessary computational primitives. However, this raises further questions. Why do circuits form discretely rather than gradually? What determines the specific parameter thresholds for different capabilities? Can we predict emergence from architectural properties and training dynamics? Answering these questions requires tighter integration between empirical scaling studies, mechanistic interpretability research, and theoretical analysis of transformer capabilities.

Limitations of our work include computational constraints preventing exploration of trillion-parameter scales where additional emergence phenomena may appear, limited architectural diversity in our model suite which may miss architecture-specific effects, and evaluation focus on discriminative tasks which may not generalize to generative capabilities. Future work should investigate emergence patterns across architectural variations including mixture-of-experts and sparse models, extend analysis to larger models using techniques like progressive training, explore training methodologies specifically designed to accelerate emergence, and develop theoretical frameworks that predict emergence thresholds from first principles rather than empirical observation.

The societal implications of non-monotonic scaling deserve consideration. If capabilities emerge unpredictably, then incremental safety evaluations may miss critical capability jumps that occur between checkpoints. A model passing safety evaluations at 80% of target scale might develop problematic capabilities in the final 20% of training. This argues for continuous monitoring throughout scaling and conservative safety margins when planning model sizes. Additionally, the concentration of emergence at large scales may exacerbate disparities between organizations with extensive compute access and those without, potentially centralizing advanced AI capabilities among few well-resourced actors.

VI. CONCLUSION

We have demonstrated through comprehensive empirical analysis that scaling laws in foundation models exhibit complex non-monotonic behavior fundamentally inconsistent with simple power-law formulations. Emergent capabilities appear discontinuously at task-specific parameter thresholds, creating hierarchical acquisition patterns that reflect staged development of internal computational structures. These findings necessitate refined theoretical frameworks incorporating phase transitions and discrete capacity thresholds for accurate capability prediction, moving beyond smooth scaling assumptions that dominate current understanding.

Our analysis reveals systematic relationships between task complexity and emergence scale, providing preliminary guidance for efficient model development. Compute-optimal training strategies significantly impact emergence timing, suggesting practical pathways to capability acquisition at reduced computational cost. The hierarchical dependency structure among capabilities implies that development strategies should account for prerequisite skill acquisition, though explicit curriculum design shows mixed effectiveness in our experiments. These insights enable more strategic resource allocation in foundation model research and development.

Future research should investigate the mechanistic basis of emergence through detailed interpretability analysis, linking discrete capability transitions to formation of specific circuit structures. Architectural modifications that lower emergence thresholds warrant exploration, as they could democratize access to advanced capabilities. Developing predictive frameworks for emergence in novel task domains remains a critical challenge requiring integration of empirical observation, mechanistic understanding, and theoretical analysis. Understanding these dynamics will prove crucial for advancing AI capabilities efficiently, safely, and equitably.

REFERENCES

[1] A. Vaswani et al., “Attention Is All You Need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 6000–6010.
[2] T. Brown et al., “Language Models Are Few-Shot Learners,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 2020, pp. 1877–1901.
[3] J. Kaplan et al., “Scaling Laws for Neural Language Models,” arXiv:2001.08361, 2020.
[4] J. Wei et al., “Emergent Abilities of Large Language Models,” arXiv:2206.07682, 2022.
[5] J. Hoffmann et al., “Training Compute-Optimal Large Language Models,” arXiv:2203.15556, 2022.

[6] A. Srivastava et al., “Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models,” arXiv:2206.04615, 2022.
[7] N. Elhage et al., “A Mathematical Framework for Transformer Circuits,” Transformer Circuits Thread, 2021.
[8] C. Olsson et al., “In-Context Learning and Induction Heads,” arXiv:2209.11895, 2022.