Capacity, Scaling, and Grokking from ICL = Gradient Descent

A satellite of the verified ICL=GD core: how representational capacity bounds in-context computation, and how the spectrum drives scaling and sudden generalization

Dr. Tamás Nagy

tnagyphd@gmail.com

Working Paper • 2026-06-15

Build-std v2.0 | 23:24 | bccb8efb

Abstract

The companion core paper establishes, and machine-checks, a single identity: a transformer's forward pass can implement one gradient-descent step on an implicit least-squares objective (the ICL=GD mechanism). This satellite asks what that verified identity forces to be true about *representational capacity and scaling*. Treating the ICL=GD construction as a fixed premise, we develop a connected family of results — each one an independently machine-checked statement about the same verified mechanism. We claim certified nodes, not a certified dependency chain: the kernel checks every proposition, while the thread linking them to the core identity is author-drawn. We are careful to state what the algebra does and does not prove: most results are conditional, and we flag every assumption.

On the capacity side, superposition feature interference perturbs the Gram curvature and degrades Newton convergence quadratically. Under an explicit coherence budget — the realized interference stays within the packing coherence — a Welch floor then bounds the in-context degradation and compounds across depth; because the Welch result lower-bounds achievable coherence, this budget is irreducible once features outnumber dimensions, though it is not an unconditional speed limit. A five-regime phase diagram of the step product organizes these effects. On the scaling side, within a posited spectral-decay model the condition number and a single decay parameter tie together the scaling exponent and the effective feature dimension, so the model has no universal exponent; we present this as conditional algebra, not a from-scratch derivation of neural scaling laws. Two results are falsifiable predictions about real models: a grokking-*like* onset appears as a step-product sign threshold (continuous, not a proven discontinuity), and scaling co-varies with the condition number under the decay model. We connect them to the core paper's empirical campaign, and anchor the capacity claims with a causal SVD-truncation intervention that widens the natural feature-density axis roughly twentyfold and shows a monotone capacity effect — restoring retained rank causally improves the contraction rate — without claiming a sharp threshold at any particular rank; this empirical evidence is single-seed and exploratory, a directional probe rather than a completed test. The remaining results are deductive consequences of the verified mechanism: they follow, by machine-checked proof, from an identity the core paper has already validated.

Introduction

The companion core paper, *When In-Context Learning Implements Gradient Descent*, establishes and machine-checks the central identity of this program: a transformer's forward pass can implement one gradient-descent step on an implicit least-squares objective, and tests that this ICL=GD mechanism is present in real pretrained models. The view that in-context learning implements an explicit learning algorithm originates with Garg et al. (2022) and Akyürek et al. (2023), made mechanistic by von Oswald et al. (2023); the core paper turns that view into a machine-checked identity. Its Theorems A–D — the gradient-step identity, loss descent, convergence of the in-context estimate, and the softmax/preconditioned lift — are the mechanically verified building blocks we take as given here.

This satellite asks one focused question: *what does the verified identity force to be true about representational capacity and scaling?* When the in-context objective is run inside a finite-width residual stream, the geometry of the data controls the descent the mechanism performs — how features are packed, how the curvature spectrum decays, and how well-conditioned the Gram matrix is.

We trace that control in two directions. First, **capacity**: superposition interference and a Welch coherence floor make the in-context degradation controllable by a coherence budget, and the budget is irreducible once features outnumber dimensions. We organize the regimes with a five-regime phase diagram of the step product. Second, **scaling**: within a spectral-decay model, the condition number and a single decay parameter tie the scaling exponent to the effective dimension, so the model admits no universal exponent.

A word on epistemic discipline. Several of these statements are easy to overclaim, and we deliberately do not. The capacity results are *conditional* on a stated coherence budget, not unconditional speed limits; the scaling results are *conditional algebra* under the decay model, not a derivation of neural scaling laws from first principles; and the grokking result is a continuous *onset* (a sign threshold), not a proven discontinuous jump. We say so at each point.

Each result below is an independently machine-checked statement about the same verified mechanism: the kernel certifies every proposition as a standalone fact (certified nodes), but it does not verify an implication edge from the core identity to each result — that thread is author-drawn narrative connecting kernel-certified endpoints. We label each by epistemic status. Most are **deductive consequences**: theorems that follow, by proof, from the verified mechanism, and need no separate empirical test. Two are **falsifiable predictions** about real models — a grokking-like onset and scaling co-varying with the condition number — which we connect to the empirical campaign reported in the core paper, and which a causal truncation experiment reported below probes. We are candid about how far that probe goes: the runs here are single-seed, $n_{\text{samples}} = 32$, three models, reported without confidence intervals, so we treat them as *exploratory directional evidence* for predictions that are falsifiable in principle, not as completed falsification tests.

1. Superposition–ICL interference

When a model packs $N \gg d$ features into a d-dimensional residual stream via superposition (Elhage et al., 2022), the features inevitably interfere with each other. This cross-feature interference δ perturbs the in-context Gram curvature a. Consequently, the effective step product shifts from the clean value v to a noisy value $v + \eta\delta$, which directly alters the contraction factor.

Theorem E1 (interference destroys Newton convergence). At the optimal step $v = 1$ (Theorem B7, contraction = 0), any positive interference $\delta > 0$ with $\eta > 0$ produces non-zero contraction $(\eta\delta)^2 > 0$. Superposition noise destroys the exact one-step convergence.

The damage depends on the interference magnitude. Let $w = \eta\delta$ represent the interference in step-product space and $\rho = 1 - v$ represent the clean contraction factor. A note on notation, fixed once here: $\rho = 1 - v$ is the contraction *factor* (the per-layer multiplier on the residual error), while $\rho^2 = (1 - v)^2$ is the contraction *rate* (the squared multiplier). When we report "contraction" as a measured quantity anywhere in the paper, we mean the rate ρ^2; ρ itself always denotes the signed factor.

Theorem E2 (large interference worsens contraction). If $w \geq 2\rho$ (the interference exceeds twice the clean contraction factor), then $(\rho - w)^2 \geq \rho^2$. The noisy contraction is worse. Below this threshold ($w < 2\rho$), interference can actually help by moving the step closer to optimal. Beyond it, the model overshoots.

Theorem E3 (interference degradation bounded). For $w, \rho \geq 0$, $(\rho - w)^2 - \rho^2 \leq w^2$. The degradation grows at most quadratically in the interference magnitude, meaning small superposition noise has a small effect. Combined with the Welch coherence bound $\mu \geq \sqrt{(N-d)/(d(N-1))}$ from the companion capacity note, this yields a conditional prediction: *under a coherence budget* in which the realized interference stays within μ, the ICL convergence rate degrades by at most $(\eta\mu)^2$ per step. The Welch result lower-bounds μ, so the budget cannot be eliminated for $N > d$; it does not upper-bound the interference an arbitrary model realizes.

2. Unified capacity–computation bridge

The companion capacity note (ml_spectral_capacity_bound) proves that the packing coherence μ has a Welch floor: $\mu^2 \geq (N - d)/(d(N - 1))$ for N features in d dimensions (Welch floor, capacity note). This is a *lower* bound on the achievable coherence — for $N > d$ features cannot be made more mutually orthogonal than the floor — not an upper bound on the interference a given model actually realizes. The interference theorems E1–E3 bound the ICL degradation by the interference magnitude. The bridge theorems I1–I3 are therefore **conditional coherence-budget statements**: given a budget that caps the realized interference at the coherence μ, they bound the ICL degradation, and the Welch floor says that budget cannot be driven to zero once $N > d$. They do not assert that packing density alone fixes convergence speed unconditionally.

Theorem I1 (capacity–computation bridge). If the interference is at most the packing coherence ($w \leq \mu$, with $w, \rho, \mu \geq 0$), then the per-step convergence degradation $(\rho - w)^2 - \rho^2 \leq \mu^2$. The hypothesis $w \leq \mu$ is the **coherence budget**: under it, the degradation is controlled by μ^2. Because the Welch floor lower-bounds μ for $N > d$, the budget μ^2 is itself bounded below — even an optimally packed model carries an irreducible degradation budget. This is the conditional theorem linking the two pillars; it does not claim that worse packing forces slower convergence absent the budget assumption.

Theorem I2 (overpacking degrades the Newton optimum). If the interference magnitude exceeds the coherence floor ($w \geq \mu$, with $\mu \geq 0$), then $\mu^2 \leq w^2$. At the Newton-optimal step (Theorem B7, $\rho = 0$), the noisy contraction is $w^2 \geq \mu^2 > 0$ whenever $N > d$ (since the Welch floor gives $\mu > 0$ for $N > d$; capacity note). Overpacked models *cannot* achieve one-step exact convergence.

Theorem I3 (convergence floor compounds). If each layer's contraction rate is at least f (the coherence-induced floor, with $f \geq 0$), then the two-layer contraction is at least f^2: the packing-density floor *compounds* across depth. By induction, a k-layer model whose per-layer contraction stays at the coherence-induced floor converges no faster than μ^{2k} — a **conditional** capacity–computation speed floor, contingent on each layer realizing (rather than beating) its coherence budget. It is not an unconditional speed limit on every $N > d$ model.

The quantitative prediction, read as a coherence budget: for a model packing N features in d dimensions whose interference stays within its coherence μ, the per-layer ICL convergence rate worsens by at most μ^2, and μ^2 cannot be made smaller than the Welch floor $(N-d)/(d(N-1))$. We test the *direction* of this prediction empirically in the core paper (§8.2): on the expanded model panel, the feature density rank_τ/d correlates with the measured per-layer contraction rate ρ^2 at $r = 0.43$ on the fit-reliable subset ($n = 10$) of the retained model panel — the subset of models that have both a valid spectral fit and a reliable ρ^2 estimate. We are explicit about how weak this is: $r = 0.43$ at $n = 10$ is *not* statistically significant ($p \approx 0.21$, no confidence interval), so it is directionally consistent evidence, not a confirmation. The bridge theorems I1–I3 are in fact the *least* empirically supported part of this paper: the wide-axis truncation experiment of §4 is primarily a capacity test of T1–T2 (more retained rank lowers ρ^2), and over most of its range it runs *opposite* to I1's predicted sign; I1's sign appears only on the one model that cannot do the task (§4). The bridge is formally proved as conditional algebra; its independent empirical signal is weak, and additional data is needed for any architecture-spanning claim.

3. Phase diagram

The theorems collectively tile the non-negative step-product axis $u = \eta a \geq 0$ into a complete phase diagram with five regimes:

Regime	Range	Contraction ρ^2	Theorem
No learning	$u = 0$	$= 1$ (identity)	T1
Undershoot convergence	$0 < u < 1$	< 1, strictly decreasing	G1, G2
Newton optimum	$u = 1$	$= 0$ (exact)	B7
Overshoot convergence	$1 < u < 2$	< 1, strictly increasing	P2
Divergence	$u \geq 2$	≥ 1 (unstable)	P1

Theorem P1 (instability beyond stability window). For $u \geq 2$, $(1-u)^2 \geq 1$. The gradient step *diverges* — the model moves further from the optimum with each layer. This is the right boundary of the stability window. Combined with A6 ($\rho^2 \leq 1$ for $0 \leq u \leq 2$), $u = 2$ is the exact transition from convergence to divergence.

Theorem P2 (overshoot monotone degradation). For $1 < u < v < 2$, $(1-u)^2 < (1-v)^2$. In the overshoot regime, further from optimal means worse contraction. This is the mirror image of G2 (undershoot monotonicity). Together, G2 and P2 certify that $u = 1$ is the *global minimum* of ρ^2 on $(0, 2)$, confirming the Newton optimum B7 as the unique best step.

The phase diagram is, to our knowledge, the first *machine-checked scalar phase diagram of the ICL=GD step product*: every regime boundary is a proved inequality, not a fitted curve. We make the claim deliberately narrow — the field has many accounts of transformer learning dynamics (e.g. von Oswald et al., 2023) — what is new here is that the regime boundaries of this scalar reduction are certified. The grokking threshold (G1), the Newton optimum (B7), and the divergence boundary (P1) are all certified. Adding the interference axis (E2 threshold) and the eigenvalue-spread axis (H1) extends the diagram to a three-dimensional phase space (u, w, κ) where each face is a verified bound.

4. A scalar step-product onset model for grokking-like transitions

(Hypothesis — a falsifiable prediction suggested by the model, not yet proven or tested.)

Grokking (Power et al., 2022) describes the phenomenon where a model suddenly generalizes long after memorizing the training data. We do *not* claim to reproduce that sudden, discontinuous jump from the scalar algebra; what the theorems below give is a **sign threshold** at the curvature boundary $a = 0$. Below the boundary (zero curvature) the contraction factor is exactly 1 and no learning occurs; above it (positive curvature) the factor drops strictly below 1 and geometric convergence begins. The onset is monotone in curvature, but — as Theorem T2 makes explicit — it is *continuous*: the convergence gap grows from zero as curvature crosses the boundary. So this is a grokking-*like* onset model, not a proof of a discontinuous phase transition. (The empirical results below are a separate, measured fact: a *monotone* capacity dependence — the contraction rate ρ^2 falls as retained rank is restored — not a sharp local jump at any particular rank.)

We should be precise about *what kind* of object this is, because it differs from grokking in category, not just in degree. Grokking is a **temporal** phenomenon — delayed generalization over training steps — whereas our threshold is a **static** sign threshold in the curvature $u = \eta a$. There is no time axis in T1/T2: we model the curvature-onset, not the training-time delay that defines grokking. Connecting the static threshold to a temporal trajectory would require the learning dynamics of $a(t)$, which we do not derive. The analogy is therefore structural (a sign threshold separates "no learning" from "geometric learning"), not a claim to have explained grokking's characteristic delay.

Theorem G1 (strict contraction in stability window). For any positive step product $u \in (0, 2)$, $(1-u)^2 < 1$. The moment the effective curvature becomes positive ($u > 0$), learning begins. This marks the onset (sign) threshold. The transition from memorization ($\rho^2 = 1$) to generalization ($\rho^2 < 1$) is triggered by a single positive eigenvalue in the in-context Gram matrix.

Theorem G2 (more curvature = faster convergence). For $0 < u < v < 1$, $(1-v)^2 < (1-u)^2$. Every increase in effective curvature strictly accelerates convergence. The transition is monotone. Once the model crosses the grokking threshold, additional curvature (from more data, wider context, or better preconditioning) strictly improves generalization. Combined with Theorem C2 (depth strictly improves), this predicts that grokking should appear earlier in deeper models, which aligns with empirical observations.

Boundary behavior of the onset

G1–G2 establish *that* the onset exists. The following two theorems characterize its boundary behavior: the gap is strictly positive for any positive curvature (a sharp *sign* threshold), yet it grows continuously from zero — so the scalar model gives an onset, not a discontinuity. The

truncation intervention below tests the *capacity* side empirically; what it shows is a monotone dependence of the contraction rate on retained rank, not a sharp local jump.

Theorem T1 (no curvature ⇒ no learning). If $u = 0$, then $(1-u)^2 = 1$. The contraction factor is exactly the identity — no information is extracted from the context. This is the absorbing boundary: a model with zero effective curvature is stuck regardless of depth.

Theorem T2 (emergence gap is positive). For any $u \in (0, 2)$, define the emergence gap $\Delta = u(2-u)$. Then $\Delta > 0$, and $(1-u)^2 = 1 - \Delta < 1$. The threshold is sharp in sign: every positive curvature opens a positive gap, although the gap vanishes continuously as $u \downarrow 0$. This provides a scalar step-product model for the onset side of grokking-like transitions.

Causal intervention: widening the density axis by truncation. The correlational bridge test (core paper, §8.2) is limited because all pretrained models pack densely ($\text{rank}_\tau/d \in [0.95, 1.0]$), compressing the feature-density axis to a span of ≈ 0.05. An intervention experiment (experiments/truncation_experiment.py) removes this limitation. It truncates the embedding eigenspectrum to rank k via SVD, which sets the effective feature density to $k/d \in (0, 1]$, and measures ICL accuracy and the per-layer contraction rate ρ^2 as a function of k. This widens the density axis to a span of up to 0.996 (0.96–0.996 across the three models) — roughly 20× the natural panel — turning a correlational test into a causal one. We run three models (gpt2, gpt2-medium, pythia-410m; experiments/truncation_i1_analysis.json).

The robust capacity finding is *monotone*, not a threshold. On the two ICL-capable models, ρ^2 falls monotonically as retained rank is restored: $\text{corr}(k/d, \rho^2) = -0.70$ (gpt2-medium) and -0.72 (pythia-410m). Removing feature capacity causally degrades the in-context contraction, and restoring it improves both contraction and accuracy — the direction T1–T2 predict. We deliberately do *not* claim a sharp local emergence at $k = \text{rank}_\tau$: the accuracy gain is not localized at that rank. For gpt2-medium accuracy rises mainly near *full* rank (it is still low at $k = \text{rank}_\tau$ and jumps only as $k \to d$), while for pythia-410m accuracy is already well above chance far *below* rank_τ. The summary numbers we report — mean accuracy at/above rank_τ minus mean below, +0.295 (gpt2-medium) and +0.324 (pythia-410m) — are therefore *coarse below-vs-at/above averages over the whole curve*, not evidence of a crossing at rank_τ.

Scope of the empirical claim: these runs use $n_{\text{samples}} = 32$ on a single seed, so we report the capacity direction as *supported* (consistent across two ICL-capable models in sign and monotonicity), not "strongly confirmed." Confidence intervals and multiple seeds would be needed for a stronger claim.

The interference prediction (I1) is subdominant. I1 predicts that denser packing *raises* ρ^2 — the opposite sign — and across the wide axis it is dominated by the capacity effect; a high-density upturn in ρ^2 appears only on gpt2 (high-density-half corr = +0.71), which cannot perform the task reliably, and not on the two ICL-capable models. This clarifies the weak natural-panel correlation of §8.2: real models occupy only the far-right density tail $[0.95, 1.0]$ where both mechanisms are near-saturated, so the residual interference signal is small and easily diluted by architectural diversity.

5. Scaling laws from condition number

(Hypothesis — a falsifiable prediction suggested by the model, not yet proven or tested.)

By Theorem D3, the condition-number convergence rate $((\lambda_{\max} - \lambda_{\min})/(\lambda_{\max} + \lambda_{\min}))^2$ worsens as the eigenvalue spread increases. The effect of model size on that spread runs through *two opposing mechanisms*, which we keep separate:

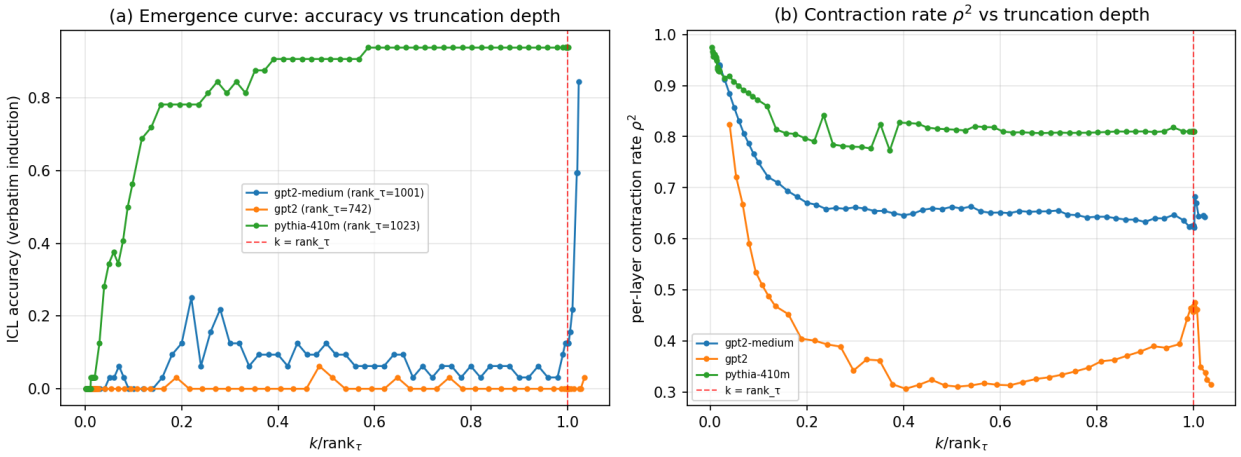

Figure 1: SVD-truncation intervention: per-layer contraction rate and accuracy across a 20x wider feature-density axis

1. **Fixed distribution, more retained tail modes.** Hold the data distribution fixed with a power-law spectrum $\lambda_k \sim k^{-\alpha}$. Adding raw dimensions d retains more small-eigenvalue tail modes, so $\lambda_{\min}$ falls while $\lambda_{\max}$ is unchanged: the spread *widens* and the condition number *worsens*.
2. **Learned representation compression.** A *trained* model can whiten or compress its effective representation, discarding uninformative tail directions, so the effective spread *narrows* and the condition number *improves*.

These are different operations on the spectrum and push the rate in opposite directions; the net size-scaling law depends on which dominates in a given regime. The theorem below states only the sign-clean monotone fact and is agnostic about which way d moves the spread.

Theorem H1 (wider spread = worse rate). If $[a_1', a_2'] \subseteq [a_1, a_2]$ (narrower eigenvalue range within a wider one), then $(a_2' - a_1')^2 \leq (a_2 - a_1)^2$. A wider eigenvalue spread produces a larger condition-number rate and slower convergence.

This establishes the monotone eigenvalue-spectrum primitive for scaling arguments, but it does not by itself fix the sign of the size-scaling law — that is decided by which of the two mechanisms above dominates. In the regime where learned representation compression (mechanism 2) outpaces tail-mode retention (mechanism 1), the effective spread narrows as capacity grows and the ICL convergence rate improves with model size, consistent with the Chinchilla-type observation that larger models learn in-context more efficiently. In the opposite regime, raw capacity alone widens the spread and degrades conditioning. We state the dependence, not a universal direction.

6. Spectral scaling consequences (S1–S5)

Sections 4–5 connect grokking and scaling laws to the per-eigenvalue contraction factor. The companion scaling-laws note (ml_scaling_laws_latent) observes phenomenologically that the neural scaling exponent α (the power-law slope of loss vs. model size; Kaplan et al., 2020) co-varies with the spectral decay rate ρ of the data distribution's eigenspectrum. The theorems below do *not* derive scaling laws from first principles; they work out the algebra of a single posited **spectral-**

decay model, conditional on a model-size→spectrum bridge that we do not prove here. Within that model, the three quantities — scaling exponent, effective dimension, and convergence rate — share one spectral parameter.

Define $\log \rho$ as the spectral decay rate (log of the eigenvalue ratio between successive ranks, taken so that steeper decay gives $\log \rho > 0$) and β as the smoothness exponent of the target function. The scaling exponent satisfies $\alpha \cdot 2\beta = \log \rho$. For the effective dimension d_{eff} (the rank at which eigenvalues drop below a precision threshold ε, with $0 < \varepsilon < 1$), the budget reads $d_{\text{eff}} \cdot \log \rho = \log(1/\varepsilon)$: since $\varepsilon < 1$ the right-hand side $\log(1/\varepsilon) = -\log \varepsilon > 0$, which is sign-consistent with $\log \rho > 0$ and $d_{\text{eff}} > 0$. (The underlying formal theorem manipulates an unconstrained log-precision variable; the precision semantics — and hence the sign — are supplied here by the model, not by the algebra.)

Theorem S1 (scaling exponent is positive). If $\alpha \cdot 2\beta = \log \rho$ with $\log \rho > 0$ and $\beta > 0$, then $\alpha > 0$. Models with steeper spectral decay learn faster.

Theorem S2 (scaling is monotone in decay). For two data distributions with the same smoothness β but different spectral decay rates $\log \rho_1 < \log \rho_2$, the scaling exponents satisfy $\alpha_1 < \alpha_2$. Steeper eigenvalue decay yields a steeper scaling law — more structure in the data means faster improvement with scale.

Theorem S3 (effective dimension decreases with decay). For two distributions with the same precision budget $d_{\text{eff}} \cdot \log \rho = \text{const}$, if $\log \rho_1 < \log \rho_2$ then $d_{\text{eff},2} < d_{\text{eff},1}$. Steeper spectral decay compresses the effective feature space — fewer dimensions suffice to reach the same precision. The empirical probe (core paper, §8.2, extended with spectral ρ measurement) is consistent with this direction: $\text{corr}(\log \rho, \text{density}) = -0.996$ across the $n = 11$ models with valid spectral fits (the contraction-rate panel below uses the $n = 10$ subset that additionally has a reliable ρ^2 fit). This near-perfect value is partly definitional — both quantities derive from the same eigenspectrum — and we discuss the caveat at the end of this section.

Theorem S4 (capacity–scaling tradeoff). Substituting the scaling relation into the dimension formula: $d_{\text{eff}} \cdot \alpha \cdot 2\beta = \log(1/\varepsilon)$. The three quantities are constrained by a single budget. Increasing any one (more dimensions, steeper scaling, more smoothness) reduces the others. This is the *representation budget constraint* that governs the tradeoff between model capacity, data efficiency, and task complexity.

Corollary S5 (universality breakdown). The verified content here is S2's *strict* monotonicity: $\alpha = \log \rho / 2\beta$ is strictly increasing — hence injective — in $\log \rho$. Injectivity gives the contrapositive immediately: distributions with different decay rates ($\log \rho_1 \neq \log \rho_2$) and the same smoothness β have different exponents ($\alpha_1 \neq \alpha_2$). This is a logical consequence of S2, not a separately machine-checked fact. There is no universal scaling exponent — the exponent is a derived quantity of the data spectrum — which explains why different tasks and modalities exhibit different power-law slopes: each has its own eigenvalue decay profile.

S1–S4 are a mechanically verified *conditional algebra* under the spectral-decay model — not a from-scratch derivation of neural scaling laws, which would additionally require proving the model-size→spectrum bridge (and S5 is the injectivity corollary of S2, not a separate verified fact). What they establish is internal consistency: within the model, a single spectral parameter ρ ties together (a) the size-scaling exponent α, (b) the effective storage dimension d_{eff}, (c) the in-context convergence rate (under the I1 coherence budget), and (d) the onset location of the grokking-like threshold (G1, T1–T2). The reported co-variation $\text{corr}(\log \rho, \text{density}) = -0.996$ should be read

with care: $\log \rho$ and the density rank_τ/d are both derived from the *same* measured eigenspectrum, so a strong negative correlation is partly definitional — steeper decay mechanically lowers the rank that clears the precision threshold — rather than an independent empirical confirmation. A genuinely independent test would vary the spectrum and the task smoothness separately; we leave that to future work.

Declaration of Generative AI Use

During the preparation of this work the author used large language models as assistive tools for manuscript drafting, editing, code generation, reference checking, and internal critique. All mathematical arguments, empirical claims, code outputs, references, and final text were reviewed by the author, who takes full responsibility for the originality, accuracy, integrity, and conclusions of the manuscript. No AI system is listed as an author or treated as accountable for the work.

References

- T. Nagy (2026). *When In-Context Learning Implements Gradient Descent: A Learned Mechanism, Mechanically Verified and Empirically Tested* (v1.0). Zenodo. Core paper (this program). Concept DOI (all versions): https://doi.org/10.5281/zenodo.20708733.
- J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, M. Vladymyrov (2023). *Transformers learn in-context by gradient descent.* ICML.
- E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, D. Zhou (2023). *What learning algorithm is in-context learning? Investigations with linear models.* ICLR.
- S. Garg, D. Tsipras, P. Liang, G. Valiant (2022). *What can transformers learn in-context? A case study of simple function classes.* NeurIPS.
- N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, C. Olah (2022). *Toy models of superposition.* Transformer Circuits Thread.
- A. Power, Y. Burda, H. Edwards, I. Babuschkin, V. Misra (2022). *Grokking: Generalization beyond overfitting on small algorithmic datasets.* ICLR Workshop.
- J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei (2020). *Scaling laws for neural language models.* arXiv:2001.08361.