EMERGENT INVERSE-DEPTH SCALING FROM NONLINEARITY IN ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling laws describe power-law improvements in model performance with dataset size and parameter count, yet their underlying mechanisms are not fully understood. To explain the parameter count scaling, existing theory posits power-law scaling with model depth. In linear-attention models,[1] this scaling is tied to a power-law data spectrum: unable to selectively attend to relevant tokens, these models learn according to global spectral strength, with stronger directions learned before weaker ones. Large language models, however, can be strongly nonlinear. Here, we show that nonlinear attention yields inverse-depth decay of loss across all tested data spectra. Nonlinearity enables attention to focus selectively on relevant tokens, allowing strong and weak spectral directions to be learned in parallel. Similar focusing across layers motivates a connection to the central limit theorem: shared error across layers sets the loss plateau, while aggregation turns layer-specific differences into continued gains with depth. Our findings suggest that depth scaling may arise from nonlinearity in attention, which allows large language models to focus locally and may make the global covariance structure less relevant.

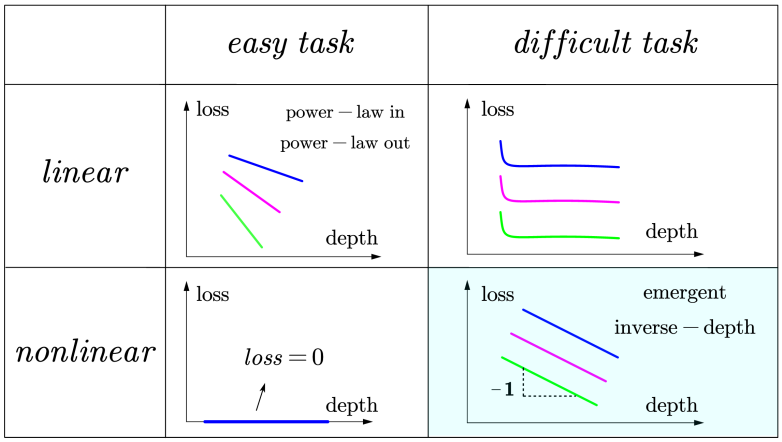


Figure 1: The nonlinear-attention model can yield inverse-depth scaling across all tested data spectra. This overview compares depth-dependent performance across models and tasks. Colors distinguish data spectra, and all panels are plotted on a log-log scale.

1 INTRODUCTION

Neural scaling laws describe how model loss decreases with the size of the dataset and the model, providing a practical basis for predicting performance and allocating computational resources (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022). Despite their predictive success, these laws remain largely empirical, and the mechanisms underlying their behavior are not fully understood.

[1]Here, linear attention means that attention scores are computed directly as query–key inner products and that the model contains no nonlinear components.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Clarifying their origins could help establish the scope of these laws and reveal how scaling can be made more efficient.

Recent theoretical work has begun to examine model-size scaling more closely by separating the roles of width and depth (Bahri et al., 2024; Liu et al., 2025; Bordelon et al., 2026). Increasing width provides more representational capacity within each layer, whereas increasing depth adds successive processing steps; the two can therefore improve performance through different mechanisms. Here, we focus on how models use additional layers and how this gives rise to scaling with depth. Previous work on linear-attention models ties power-law depth scaling to a power-law data spectrum (Bordelon et al., 2026). Yet it does not account for nonlinear attention in practical large language models. In contrast, empirical work on these models suggests that inverse-depth scaling can arise through an ensemble of layers with similar behavior (Liu et al., 2026a), without explaining the underlying reason for this layerwise similarity. We therefore ask:

Our research question:

How does nonlinearity in attention affect depth scaling and its underlying mechanisms?

We study this question in the common setting of in-context learning (ICL), where a model uses examples in its context to make predictions for a new query. Our tasks model the second half of an induction head's computation: retrieving the token that matches the query and copying its associated value (Elhage et al., 2021; Olsson et al., 2022; Lv et al., 2025). To isolate the role of nonlinearity, we compare controlled Transformers with linear or nonlinear attention across independently varied data spectra.

Figure 1 summarizes the resulting picture. In an easy task, the linear-attention model follows a spectrum-dependent power law, while a single-layer nonlinear-attention model is enough to solve the task. In a difficult task, the linear-attention model quickly plateaus, whereas the nonlinear-attention model exhibits inverse-depth loss scaling. Unlike the linear-attention model, which cannot selectively focus on relevant tokens, the nonlinear-attention model can attend to them directly, allowing strong and weak spectral directions to be learned in parallel and enabling better utilization of depth. Further analysis decomposes the loss into cross-layer and within-layer terms and connects this decomposition to the central limit theorem. The asymptotic behavior of these terms explains the origins of the loss plateau and the reducible loss.

Finally, we relax several simplifying assumptions of the controlled setup to move closer to standard Transformers and observe similar signs of inverse-depth scaling. These results suggest that nonlinearity in attention can be a plausible source of depth scaling in real models and may make scaling less dependent on the global covariance structure of the data.

Our contributions are the following:

- We discover that nonlinearity in attention can qualitatively affect depth scaling, enabling emergent inverse-depth loss scaling across all tested spectra.
- We explain how nonlinear attention enables parallel learning of strong and weak spectral directions, in contrast to the sequential learning behavior of linear-attention models.
- We connect inverse-depth scaling to the central limit theorem through a loss decomposition that identifies the origins of the loss plateau and the gains from depth.

2 Problem Setup

2.1 Task Setup

In-context learning is a common setting in which predictions rely on interactions among context tokens. These interactions expose the structure of the data, providing a useful setting for studying the origins of depth scaling. We consider a controlled ICL problem in which each context contains P inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ with labels generated by a teacher vector $\boldsymbol{\beta} \in \mathbb{R}^d$:

$$y_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i, \qquad i = 1, \dots, P. \tag{1}$$

Let λ_k and ω_k denote the eigenvalues of the input and teacher covariances, respectively. To control input and task structure independently, we vary these spectra as

$$\lambda_k \propto k^{-a}, \qquad \omega_k \propto k^{-b}, \qquad k = 1, \dots, d. \tag{2}$$

Here, $a, b \geq 0$ are the input-spectrum and teacher-spectrum exponents, respectively. They independently control how rapidly the corresponding variances decay across feature directions. Setting either exponent to zero gives an isotropic covariance for the corresponding distribution.

In natural language, different contexts may share similar features but express them along different directions. Following Bordelon et al. (2026), we model this variation by independently sampling an orthogonal rotation matrix $O \in \mathbb{R}^{d\times d}$ for each context. Conditioned on O, the input and teacher covariances are

$$\begin{aligned} \Sigma &= \mathrm{Cov}(\boldsymbol{x}_i \mid O) = O\,\mathrm{diag}(\lambda_1, \dots, \lambda_d)O^\top, \\ \Omega &= \mathrm{Cov}(\boldsymbol{\beta} \mid O) = O\,\mathrm{diag}(\omega_1, \dots, \omega_d)O^\top. \end{aligned} \tag{3}$$

Conditioned on O, we sample the teacher vector independently of the context and test inputs, which are drawn i.i.d.:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \Omega)\,, \\ \boldsymbol{x}_1, \dots, \boldsymbol{x}_P, \boldsymbol{x}_{\mathrm{test}} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)\,. \end{aligned} \tag{4}$$

The model must therefore infer the relevant directions from each context rather than rely on fixed directions across contexts.

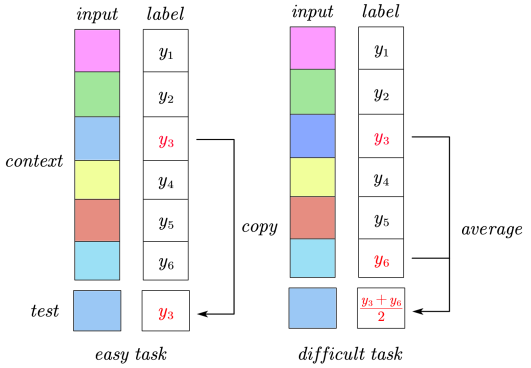


Figure 2: Both tasks model the second half of an induction head: the easy task copies the most relevant token's label, while the difficult task averages the labels of the top-k relevant tokens.

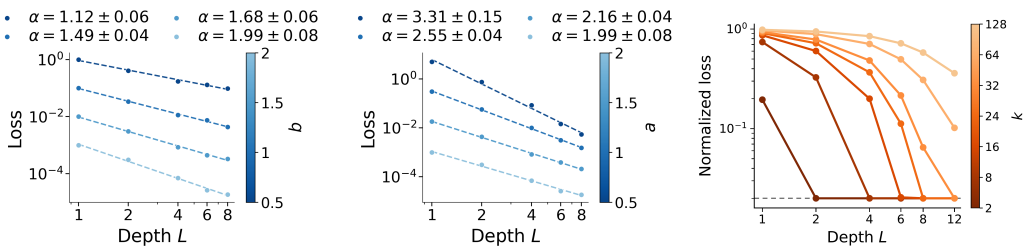
We first construct an *easy task* using a widely studied ICL setting (Garg et al., 2022; von Oswald et al., 2023; Ahn et al., 2023; Bordelon et al., 2026). Given context inputs and their teacher-generated labels, the model must predict the label assigned to a test input by the same teacher. Following Bordelon et al. (2026), we consider the infinite-context limit $P \to \infty$, in which a context token can be found arbitrarily close to the test token. The task can therefore be approximated by retrieving the closest context token and copying its label. This retrieval-and-copying operation corresponds to the second half of an induction-head circuit identified in language models (Elhage et al., 2021; Olsson et al., 2022; Lv et al., 2025). The task can thus be viewed, in an idealized sense, as an exact-match problem:

$$\boldsymbol{x}_{\mathrm{test}} = \boldsymbol{x}_j, \qquad y^*_{\mathrm{test}} = y_j, \qquad j \in \{1, \dots, P\}, \quad P \to \infty. \tag{5}$$

However, in practical ICL and language settings, prediction often requires combining information from multiple context tokens rather than retrieving a single one. Motivated by this need, we construct a *difficult task* that extends this induction-head-inspired operation from top-1 retrieval to top-k aggregation. We measure relevance by cosine similarity and average the selected labels. The selected index set and target are defined by

$$\begin{aligned} S_k(\boldsymbol{x}_{\mathrm{test}}) &= \mathrm{TopK}_{i\in\{1,\dots,P\}} \frac{\boldsymbol{x}_i^\top \boldsymbol{x}_{\mathrm{test}}}{\|\boldsymbol{x}_i\|\,\|\boldsymbol{x}_{\mathrm{test}}\|}, \\ y^*_{\mathrm{test}} &= \frac{1}{k} \sum_{i\in S_k(\boldsymbol{x}_{\mathrm{test}})} y_i. \end{aligned} \tag{6}$$

Here, TopK returns the indices of the k largest scores. Figure 2 summarizes these induction-head-inspired top-1 copying and top-k aggregation tasks.



(a) Fixed $a = 2$ (b) Fixed $b = 2$ (c) Normalized mode-wise loss

Figure 3: On the easy task, the linear-attention model exhibits spectrum-dependent depth scaling and learns spectral modes sequentially. (a) Varying b at fixed $a = 2$. (b) Varying a at fixed $b = 2$. Dashed curves show power-law fits, with the corresponding exponents indicated. (c) Normalized loss of representative spectral modes. Values below 2% are displayed at the threshold marked by the dashed line.

2.2 Model Setup

Large language models combine positional encodings, multiple attention heads, MLP blocks, normalization, and task-specific readouts. While important in practice, these components make it difficult to isolate the mechanisms underlying depth scaling. Building on Bordelon et al. (2026), we therefore use a simplified architecture that retains the components needed for context-dependent retrieval and label propagation, allowing a controlled comparison between linear and nonlinear attention.

Embeddings. Each context provides P input–label pairs $(\boldsymbol{x}_i, y_i)$, where the teacher vector $\boldsymbol{\beta}$ generates y_i from $\boldsymbol{x}_i$ according to equation 1. The model has access to all context input–label pairs and the test input, but cannot access the test label. The corresponding entry at the test position is initialized to zero. We collect the token inputs and the observed labels into

$$X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_P, \boldsymbol{x}_{\text{test}}] \in \mathbb{R}^{d\times(P+1)}, \qquad \boldsymbol{y}^{\text{in}} = (y_1, \dots, y_P, 0) \in \mathbb{R}^{1\times(P+1)}. \tag{7}$$

Let $N \geq d$ denote the hidden dimension, avoiding a representational bottleneck. Each token state lies in $\mathbb{R}^N$. With encoder matrices $W_x \in \mathbb{R}^{N\times d}$ and $W_y \in \mathbb{R}^{N\times 1}$, we encode input and label information into separate streams:

$$H_x = W_x X \in \mathbb{R}^{N\times(P+1)}, \qquad H_y^0 = W_y \boldsymbol{y}^{\text{in}} \in \mathbb{R}^{N\times(P+1)}. \tag{8}$$

This encoding separates the input features used for token matching from the labels to be propagated. The input features determine which context tokens are relevant to the query, and attention passes their label information to the test token.

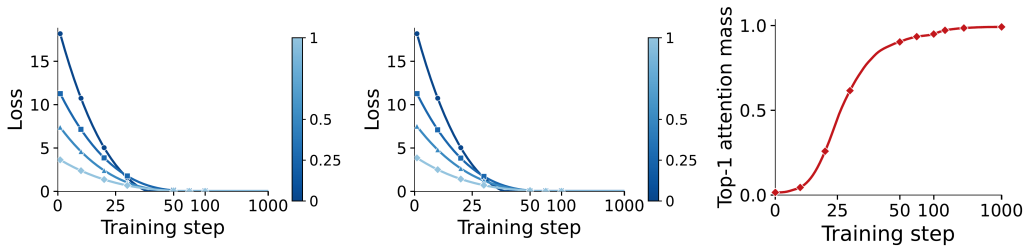
The model has L residual layers, indexed by $\ell \in \{1, \dots, L\}$. At each layer, attention uses the input features to determine how label information is propagated. For simplicity, we first keep H_x fixed across layers, while H_y^ℓ is updated as label information is passed among tokens. We use layer-specific query and key matrices $W_q^\ell, W_k^\ell \in \mathbb{R}^{N\times N}$ and a shared value matrix $W_v \in \mathbb{R}^{N\times N}$. Thus, each layer can select relevant tokens differently while using a common transformation to propagate label information.

Nonlinear attention. We first define the model with nonlinear attention. We apply pre-norm to the input-stream hidden state for better training stability and alignment with practical models, and denote the normalized state by $\widehat{H}_x$. At layer ℓ, the query, key, and value representations are

$$Q^\ell = W_q^\ell \widehat{H}_x, \qquad K^\ell = W_k^\ell \widehat{H}_x, \qquad V^\ell = W_v H_y^{\ell-1}, \qquad Q^\ell, K^\ell, V^\ell \in \mathbb{R}^{N\times(P+1)}. \tag{9}$$

Let M denote the attention-mask matrix. The attention matrix and residual update are

$$\begin{aligned} A^\ell &= \mathsf{M} \odot \operatorname{softmax}\big((Q^\ell)^\top K^\ell\big) \in \mathbb{R}^{(P+1)\times(P+1)}, \\ H_y^\ell &= H_y^{\ell-1} + \frac{1}{L} V^\ell (A^\ell)^\top \in \mathbb{R}^{N\times(P+1)}. \end{aligned} \tag{10}$$



(a) Fixed $b = 0$ (b) Fixed $a = 0$ (c) Top-1 attention mass, $(1, 0)$

Figure 4: A single nonlinear-attention layer solves the easy task by focusing on the matching token. (a) Varying a at fixed $b = 0$. (b) Varying b at fixed $a = 0$. (c) Top-1 attention mass approaches one during training.

Linear attention. The linear-attention model removes nonlinearities and retains the raw bilinear attention scores. Its attention matrix and residual update are

$$A^\ell = \frac{1}{PN}\mathsf{M} \odot \left((Q^\ell)^\top K^\ell\right) \in \mathbb{R}^{(P+1)\times(P+1)},$$
$$H_y^\ell = H_y^{\ell-1} + \frac{1}{L}V^\ell(A^\ell)^\top \in \mathbb{R}^{N\times(P+1)}. \tag{11}$$

Attention masks. As in large language models, we use attention masks to control information flow between tokens. For the linear-attention model, we adopt the masking scheme of Bordelon et al. (2026), allowing successive layers to extract and propagate information at different levels of granularity. For the nonlinear-attention model, we initially disable context updates to focus on how different layers retrieve information from a fixed context. In both models, the test token can attend to all context tokens to access their label information. Note that in the nonlinear-attention model, softmax is computed only over tokens that can be attended to, so disabled positions do not affect the distribution. Further details on the masks and context updates are given in Appendix C.3.

Output and loss. We obtain the prediction by applying a linear readout $W_o \in \mathbb{R}^{1\times N}$ to the final test-token state $h^L_{y,\text{test}}$. The model is trained to minimize the mean squared error over independently sampled contexts:

$$\widehat{y}_L = W_o h^L_{y,\text{test}}, \qquad \mathcal{R}_L = \mathbb{E}_{\text{context}}\left[(\widehat{y}_L - y^*)^2\right]. \tag{12}$$

Throughout, $\mathcal{R}_L$ denotes the evaluation mean squared error of a depth-L model; we write $\mathcal{R}_L(a, b)$ when making its dependence on the data-spectrum exponents explicit.

All models are trained until the loss has converged. In Appendix C, we relax several simplifying assumptions to better reflect the architectures of standard Transformers and large language models, and extend our experiments and analyses to these settings.

3 Easy Task Analysis

3.1 Linear-Attention Model

Figure 3(a,b) shows that the linear-attention model's loss on the easy task decreases with depth according to a power law across all tested settings, with an exponent that depends on the data spectra.

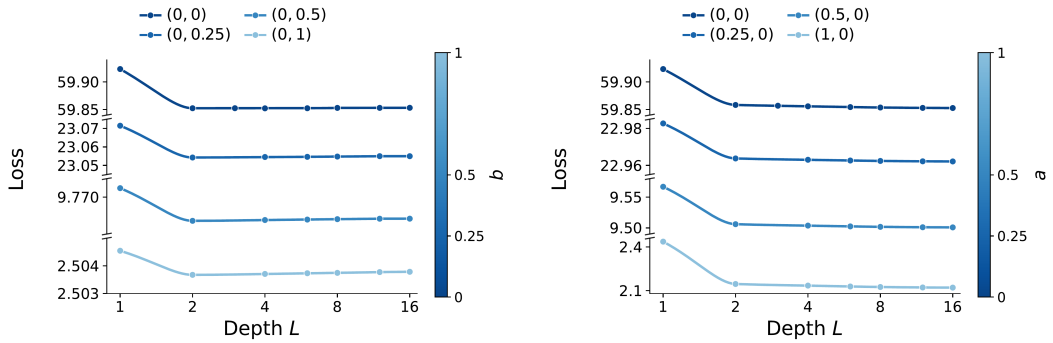
Heuristically, we can interpret the depth-L linear-attention model's output as a polynomial in the data covariance:

$$\widehat{y}_L = \boldsymbol{\beta}^\top p_L(\Sigma)\boldsymbol{x}_{\text{test}}, \qquad p_L(\Sigma) = \sum_{r=1}^{L} a_r \Sigma^r. \tag{13}$$

The target is

$$y^* = \boldsymbol{\beta}^\top \boldsymbol{x}_{\text{test}}. \tag{14}$$

Therefore, the model needs to approximate the identity map with $p_L(\Sigma)$; Appendix A.1 derives this covariance-polynomial representation. Greater depth allows the model to use higher powers of the



(a) Fixed $a = 0$ (b) Fixed $b = 0$

Figure 5: On the difficult task, linear attention rapidly reaches a spectrum-dependent loss plateau with little benefit from further depth. (a) Varying b at fixed $a = 0$. (b) Varying a at fixed $b = 0$.

covariance, incorporating information at different levels of granularity to improve the approximation. The asymptotic behavior of this approximation gives rise to power-law depth scaling. Different data spectra change the covariance Σ in the polynomial, leading to different convergence rates with depth. This produces spectrum-dependent depth scaling.

This mechanism also has a feature-learning interpretation. We decompose the power-law input covariance into individual spectral modes and normalize each mode's loss, which allows us to compare how fully each mode has been learned. Small k indexes stronger features with large eigenvalues, whereas large k indexes weaker features with small eigenvalues. Figure 3(c) reveals a clear separation in their learning dynamics: stronger modes are learned almost perfectly within a few layers, while weaker modes require substantially greater depth. Appendix A.4 further shows that individual modes can converge exponentially at different characteristic depths, with their superposition yielding spectrum-dependent power-law scaling.

3.2 Nonlinear-Attention Model

We next evaluate the nonlinear-attention model on the easy task. Figure 4(a,b) shows that a single-layer nonlinear-attention model achieves zero loss on the easy task across all tested data spectra. Unlike the spectrum-dependent improvement of the linear-attention model, its performance does not improve further with depth.

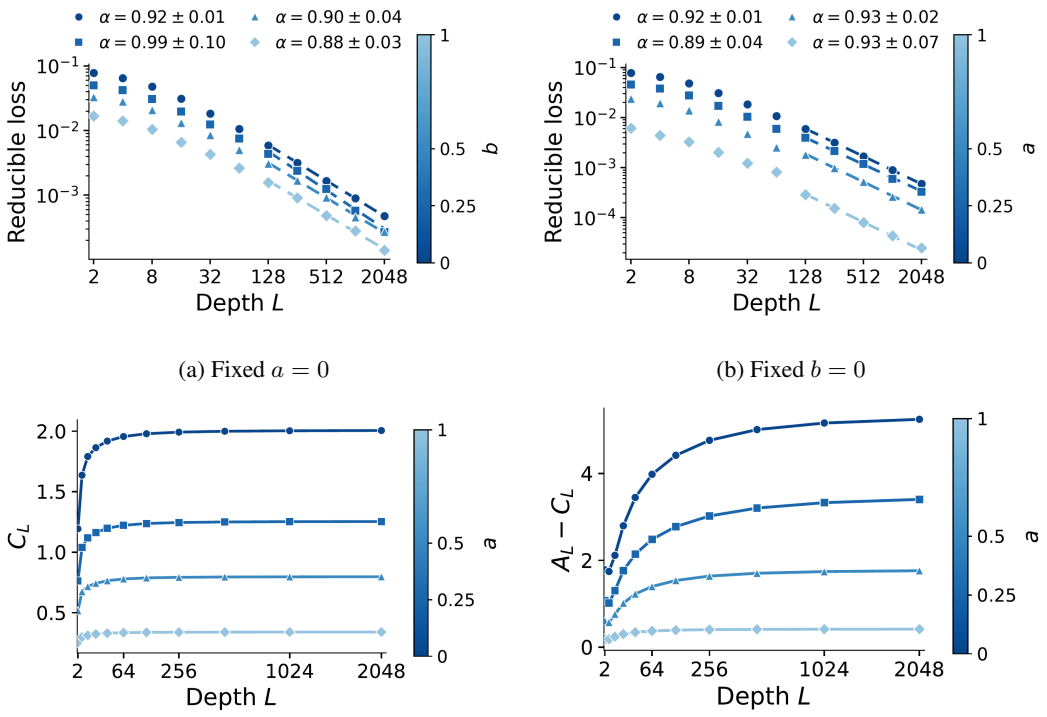
The key distinction is that nonlinear attention can selectively focus on the relevant token in each context, whereas linear attention cannot fully exclude contributions from irrelevant tokens. During training, the queries and keys attain large norms, amplifying differences between attention logits and placing softmax in an effectively low-temperature regime. Almost all attention is then assigned to the matching token, allowing its label to be copied in a single layer. Figure 4(c) supports this explanation: during training, attention shifts from a nearly uniform distribution toward the top-1 token, whose attention mass approaches one. The same qualitative behavior holds across the tested spectra.

4 Difficult Task Analysis

4.1 Linear-Attention Model

We next evaluate the linear-attention model on the difficult task. Here, we choose $k = 2$ as a representative case of top-k aggregation. Figure 5(a,b) shows that the loss reaches a spectrum-dependent plateau within one or two layers. Unlike the continued improvement on the easy task, additional depth provides little further benefit.

Intuitively, the selected tokens vary across contexts, so their average fluctuates around a shared mean. The target therefore contains a predictable component and a context-specific deviation. Linear attention can capture the shared component, but lacks the context-specific token selection needed to fully recover the deviation. This leaves an irreducible loss that sets the plateau.



(a) Fixed $a = 0$

(b) Fixed $b = 0$

(c) Shared-error term C_L at fixed $b = 0$

(d) Disagreement term $A_L - C_L$ at fixed $b = 0$

Figure 6: On the difficult task, nonlinear attention yields inverse-depth decay across all tested data spectra. (a) Varying b at fixed $a = 0$. (b) Varying a at fixed $b = 0$. (c,d) C_L and $A_L - C_L$ converge to spectrum-dependent constants.

We analyze the reducible part using the covariance-polynomial formulation in Section 3.1. The prediction retains the same form, while the predictable target has the leading-order form

$$\widehat{y}_L = \boldsymbol{\beta}^\top p_L(\Sigma)\boldsymbol{x}_{\text{test}}, \qquad \overline{y}^* \propto \boldsymbol{\beta}^\top \Sigma \boldsymbol{x}_{\text{test}}, \tag{15}$$

Unlike the identity map in the easy task, this leading target can already be represented by a degree-one polynomial. Additional depth therefore provides little further reduction in loss. Appendix A.3 provides the geometric illustration, detailed derivation, and comparison between predicted and measured loss floors.

4.2 Nonlinear-Attention Model

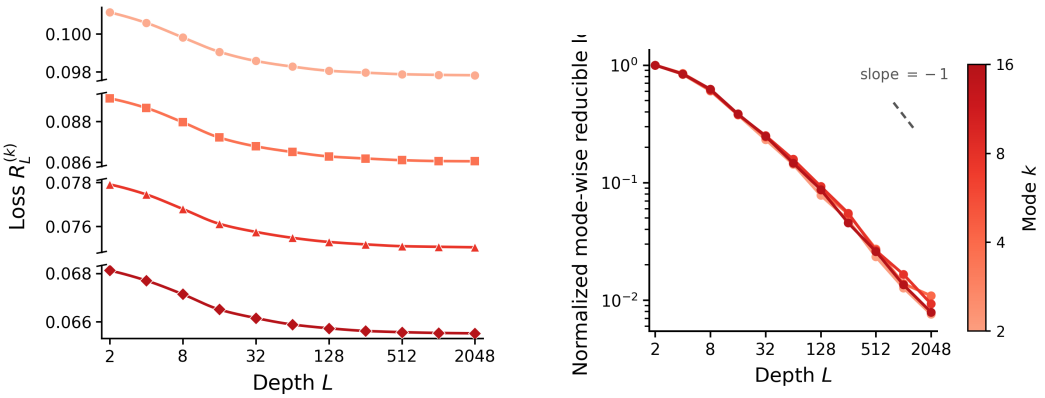
We next evaluate the nonlinear-attention model on the same difficult task with $k = 2$. Figure 6(a,b) shows that additional depth continues to reduce the loss, in contrast to the rapid saturation of the linear-attention model.

To quantify this depth dependence, we fit the evaluation loss to

$$\mathcal{R}_L(a, b) = L_0(a, b) + c(a, b)L^{-\alpha(a,b)}, \tag{16}$$

by minimizing squared residuals between the logarithms of the observed and modeled losses. The fitted exponents shown in Figure 6(a,b) are close to one across the tested data spectra, indicating emergent inverse-depth scaling in the deep regime.

With input and label information carried in separate streams and context states fixed across layers, our model admits an equivalent parallel representation: each layer forms a prediction from the same context, and the final output averages these predictions. Increasing depth therefore increases the number of predictors in the ensemble, suggesting an analogy with the central limit theorem, where averaging layer-specific fluctuations can yield a $1/L$ decay in squared error. To distinguish the error



(a) Mode-wise raw loss

(b) Normalized mode-wise reducible loss

Figure 7: With nonlinear attention, strong and weak spectral modes are learned in parallel. (a) Different modes are learned at similar rates. (b) Normalized mode-wise reducible losses nearly collapse onto a common curve, indicating parallel learning across modes.

reduced by ensembling from the error that remains, we derive the exact loss decomposition

$$\mathcal{R}_L(a,b) = C_L(a,b) + \frac{A_L(a,b) - C_L(a,b)}{L}. \tag{17}$$

Here, C_L measures error shared across predictors, whereas $A_L - C_L$ measures pairwise differences between their errors. A detailed derivation is provided in Appendix B. Figure 6(c,d) tracks these two terms across different data spectra. As depth increases, both converge to spectrum-dependent constants:

$$C_L(a,b) \longrightarrow C_\infty(a,b), \qquad A_L(a,b) - C_L(a,b) \longrightarrow \Delta_\infty(a,b), \qquad L \to \infty. \tag{18}$$

The loss plateau and the depth-dependent improvement have different origins. The error shared across predictors cannot be canceled by ensembling and sets the irreducible loss within this ensemble. For the top-2 task, this shared error reflects the mismatch between smooth softmax attention and the sharp target distribution $(1/2, 1/2, 0, \ldots, 0)$, which is difficult to match consistently across contexts.

In contrast, $A_L - C_L$ quantifies error variation across predictors. Ensembling allows these errors to partially cancel, reducing their contribution by a factor of $1/L$ and enabling continued improvement with depth. Detailed derivations and attention-mass diagnostics are provided in Appendices B and D.2, respectively.

We again examine individual spectral modes and normalize their excess losses to compare relative learning progress. Figure 7(a) shows similar improvement rates across modes, and the normalized curves in panel (b) nearly collapse. In contrast to the linear-attention model's sequential learning, nonlinear attention can focus directly on relevant tokens in each context and improve strong and weak features in parallel, even without context updates that expose progressively finer-grained information.

Building on our analysis of the differences between linear and nonlinear attention, we further compare their irreducible losses on the difficult task. Table 1 shows that the nonlinear-attention model achieves a substantially lower loss floor, clearly outperforming the linear-attention model.

We extend our setup toward standard Transformers by encoding inputs and labels in a shared space, untying the value matrix, and updating context states across layers. We observe similar signs of inverse-depth scaling (Appendix C). These results suggest that this mechanism may also contribute to depth scaling in realistic models with nonlinear attention, where data structure need not be the dominant factor.

5 Related Work

Empirical neural scaling laws relate loss to model size and dataset size, supporting performance prediction and resource allocation (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022).

Table 1: Nonlinear-attention models achieve substantially lower irreducible loss than linear-attention models on the difficult task.

(a, b)	$(0, 0)$	$(0, 0.25)$	$(0, 0.5)$	$(0, 1)$	$(0.25, 0)$	$(0.5, 0)$	$(1, 0)$
Nonlinear L_0	2.007	1.268	0.836	0.424	1.255	0.798	0.341
Linear L_0	59.853	23.055	9.766	2.504	22.962	9.501	2.120

Theoretical explanations have related them to data geometry and covariance spectra (Sharma & Kaplan, 2022; Bahri et al., 2024). More recent work identifies mechanisms arising from representation superposition and the dynamics of learning peaked distributions (Liu et al., 2025; 2026b; Kühn et al., 2026). Despite these advances, much of the literature treats model size as an aggregate quantity rather than resolving the different contributions of width and depth.

As a key component of model-size scaling, depth has been studied through several complementary accounts of how networks utilize layers. Compositional learning builds complex representations by successively combining simpler features, making hierarchical structure more efficiently learnable (Bengio et al., 2013; LeCun et al., 2015; Poggio et al., 2017; Cagnetta et al., 2024). The dynamical-systems perspective treats residual layers as discrete steps of an ordinary differential equation (E, 2017; Haber & Ruthotto, 2018; Chen et al., 2018; Sander et al., 2022; Chizat, 2025). The ensemble-averaging perspective views depth as the aggregation of related predictors (Veit et al., 2016; Lu et al., 2020; Liu et al., 2026a). These perspectives explain how networks use depth but do not fully explain the origins of depth scaling. We examine how nonlinearity in attention shapes its underlying mechanisms and resulting scaling laws.

More directly, Bordelon et al. (2026) build on the established ICL regression setting (Garg et al., 2022; von Oswald et al., 2023) to derive spectrum-dependent depth scaling for deep linear-attention models. A line of theoretical work has largely focused on analytically tractable linear models (Ahn et al., 2023; Zhang et al., 2024; Lu et al., 2025), leaving the mechanisms through which nonlinear attention affects depth scaling unresolved. Liu et al. (2026a) report inverse-depth scaling in LLMs and connect it to ensemble averaging, without explaining why this ensemble-like behavior emerges. Liu & Gore (2026) further argue that generic mechanisms may determine scaling exponents, while data and architecture affect the coefficients. Our work complements these studies by explaining how nonlinear attention enables ensemble-like behavior across layers and gives rise to inverse-depth scaling.

6 Discussion

Our results identify nonlinearity in attention as a source of depth scaling beyond power-law data spectra. Nonlinear attention can focus directly on relevant tokens in each context, allowing strong and weak spectral directions to be learned in parallel. Linear attention lacks this selective ability, tying its learning dynamics to global spectral strength. Similar focusing across layers supports an ensemble interpretation: aggregation reduces layer-specific errors while leaving a shared loss floor, connecting inverse-depth improvement to the central limit theorem. These findings suggest that practical depth scaling can originate from attention nonlinearity rather than data structure alone.

Our study has many limitations. Our simplified architecture omits components of full LLMs, so the identified mechanism is not guaranteed to operate in those models. Although architectural extensions show similar scaling signals, a clear theoretical account of their mechanisms is still lacking. Our experiments are restricted to ICL tasks, leaving generalization to real-world datasets and broader task classes untested. The identified mechanism need not be universal or dominant, and other processes may produce different scaling or interact with it. Our analysis also does not address how depth scaling couples to model width and dataset size.

Aggregation also raises a question about how efficiently depth is used. If a few layers perform the essential computation while others mainly average residual errors, separate parameters at every layer may be wasteful. Looped Transformers reuse parameters across depth (Dehghani et al., 2019; Geiping et al., 2025) and may retain aggregation benefits with fewer parameters. Testing this possibility could guide more parameter-efficient use of depth in future language models.

AI USE STATEMENT

We used generative AI tools to support brainstorming, articulate candidate mathematical claims, generate code under explicit human instructions, and assist with language editing. We did not use generative AI to generate synthetic datasets, conduct data analysis, or write mathematical proofs. All AI-assisted ideas and writing were critically reviewed by the authors, and all AI-generated code was independently inspected and verified. The authors performed the data analysis and mathematical proofs and take full responsibility for the final content of this work.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Blake Bordelon, Mary Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. In *International Conference on Learning Representations*, 2026.

Francesco Cagnetta, Leonardo Petrini, Umberto M. Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, 14(3):031001, 2024. doi: 10.1103/PhysRevX.14.031001.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1806.07366.

Lénaïc Chizat. The hidden width of deep ResNets: Tight error bounds and phase diagram. *arXiv preprint arXiv:2509.10167*, 2025. URL https://arxiv.org/abs/2509.10167.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1807.03819.

Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5:1–11, 2017. doi: 10.1007/s40304-017-0103-z.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1380.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34 (1):014004, 2018. doi: 10.1088/1361-6420/aa9a90.

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.

Jordan Hoffmann et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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.

Marcel Kühn, Yoon Thelge, and Bernd Rosenow. A boundary-layer mechanism for one-third scaling in online softmax classification. *arXiv preprint arXiv:2605.22341*, 2026.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539.

Yizhou Liu and Jeff Gore. Neural scaling universality: If exponents are fixed, time to understand coefficients. *arXiv preprint arXiv:2606.25008*, 2026. URL https://arxiv.org/abs/2606.25008.

Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. In *Advances in Neural Information Processing Systems*, 2025.

Yizhou Liu, Sara Kangaslahti, Ziming Liu, and Jeff Gore. Inverse depth scaling from most layers being similar. In *International Conference on Machine Learning*, 2026a.

Yizhou Liu, Ziming Liu, Cengiz Pehlevan, and Jeff Gore. Universal one-third time scaling in learning peaked distributions. *arXiv preprint arXiv:2602.03685*, 2026b.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean-field analysis of deep ResNet and beyond: Towards provable optimization via overparameterization from depth. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6426–6436, 2020. URL https://proceedings.mlr.press/v119/lu20b.html.

Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122.

Ang Lv, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Rui Yan. Language models "grok" to copy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 2025. URL https://aclanthology.org/2025.naacl-short.61/.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep—but not shallow—networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Grant M. Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Michael E. Sander, Pierre Ablin, and Gabriel Peyré. Do residual neural networks discretize neural ordinary differential equations? In *Advances in Neural Information Processing Systems*, volume 35, pp. 36520–36532, 2022. URL https://arxiv.org/abs/2205.14612.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

Andreas Veit, Michael J. Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1605.06431.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 35151–35174, 2023.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL https://www.jmlr.org/papers/v25/23-1042.html.

A Linear-Attention Model Theory

A.1 Covariance Polynomials in the Linear-Attention Model

We first derive the theory of the linear-attention model on the easy task. In general, its output is an ordered matrix polynomial of degree at most L in the data covariance. The isotropic setting analyzed by Bordelon et al. (2026) then follows as a special case of our theory.

Let U collect the context inputs, q denote the query, and S be the empirical input covariance:

$$\begin{aligned} U = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_P] \in \mathbb{R}^{d\times P}, \qquad q = \boldsymbol{x}_{\text{test}} \in \mathbb{R}^d, \\ y = U^\top \boldsymbol{\beta}, \qquad S = \frac{1}{P} UU^\top. \end{aligned} \tag{19}$$

The input and teacher covariances are

$$\begin{aligned} \Sigma &= O\,\mathrm{diag}(\lambda_1, \dots, \lambda_d) O^\top, \\ \Omega &= O\,\mathrm{diag}(\omega_1, \dots, \omega_d) O^\top, \end{aligned}$$

where the random orthogonal rotation O is fixed within a context and sampled independently across contexts. Contexts therefore share the same spectra but express their features in different directions. Within each context, conditioning on O, the context inputs and test token are sampled identically and independently, while the teacher vector is sampled independently from its corresponding distribution:

$$\boldsymbol{x}_1, \dots, \boldsymbol{x}_P, \boldsymbol{x}_{\text{test}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \qquad \boldsymbol{\beta} \sim \mathcal{N}(0, \Omega).$$

Define the effective query–key matrix

$$B_\ell = \frac{1}{N} W_x^\top (W_k^\ell)^\top W_q^\ell W_x, \qquad \ell = 1, \dots, L, \tag{20}$$

where N is the hidden-state dimension. Before masking, token i assigns context token j the weight

$$\frac{(W_q^\ell W_x \boldsymbol{x}_i)^\top (W_k^\ell W_x \boldsymbol{x}_j)}{NP} = \frac{\boldsymbol{x}_j^\top B_\ell \boldsymbol{x}_i}{P}.$$

Separate the context and test label states as

$$H_y^\ell = [Z_\ell, h_\ell], \qquad Z_\ell \in \mathbb{R}^{N\times P}, \quad h_\ell \in \mathbb{R}^N.$$

The linear-attention mask defined in Section 2.2 gives

$$\begin{aligned} Z_0 &= W_y \boldsymbol{\beta}^\top U, \qquad h_0 = 0, \\ Z_\ell &= Z_{\ell-1} - \frac{1}{LP} W_v Z_{\ell-1} U^\top B_\ell U, \\ h_\ell &= h_{\ell-1} + \frac{1}{LP} W_v Z_{\ell-1} U^\top B_\ell q. \end{aligned} \tag{21}$$

Write $Z_\ell = F_\ell U$, where F_ℓ maps inputs to context label states. Substituting $UU^\top/P = S$ yields

$$\begin{aligned} F_0 = W_y \boldsymbol{\beta}^\top, \qquad F_\ell \in \mathbb{R}^{N\times d}, \\ F_\ell = F_{\ell-1} - \frac{1}{L} W_v F_{\ell-1} S B_\ell, \\ h_\ell - h_{\ell-1} = (F_{\ell-1} - F_\ell) q. \end{aligned} \tag{22}$$

Therefore, the output is

$$\widehat{y}_L = W_o \sum_{\ell=1}^{L} (F_{\ell-1} - F_\ell) q = W_o (F_0 - F_L) q. \tag{23}$$

The first two steps show how the polynomial is built:

$$\begin{aligned} F_1 &= W_y \boldsymbol{\beta}^\top - \frac{1}{L} W_v W_y \boldsymbol{\beta}^\top S B_1, \\ F_2 &= W_y \boldsymbol{\beta}^\top - \frac{1}{L} W_v W_y \boldsymbol{\beta}^\top (S B_1 + S B_2) \\ &\quad + \frac{1}{L^2} W_v^2 W_y \boldsymbol{\beta}^\top S B_1 S B_2. \end{aligned}$$

With readout coefficients

$$m_r = W_o W_v^r W_y, \qquad r = 1, \dots, L,$$

substitution into the output gives

$$\mathcal{P}_L(S) = \sum_{r=1}^{L} \frac{(-1)^{r-1} m_r}{L^r} \sum_{1 \le \ell_1 < \cdots < \ell_r \le L} S B_{\ell_1} \cdots S B_{\ell_r}. \tag{24}$$

The prediction and easy-task target are

$$\widehat{y}_L = \boldsymbol{\beta}^\top \mathcal{P}_L(S) q, \qquad y^* = \boldsymbol{\beta}^\top q. \tag{25}$$

Thus learning the easy task amounts to approximating the identity on the relevant input directions. Define

$$\mathcal{D}_L(S) = I_d - \mathcal{P}_L(S).$$

Averaging over the teacher at fixed inputs gives

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\beta}}[(\widehat{y}_L - y^*)^2 \mid U, q, O] &= \mathbb{E}_{\boldsymbol{\beta}}[(\boldsymbol{\beta}^\top \mathcal{D}_L(S) q)^2 \mid U, q, O] \\ &= q^\top \mathcal{D}_L(S)^\top \mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\top \mid U, q, O] \mathcal{D}_L(S) q \\ &= q^\top \mathcal{D}_L(S)^\top \Omega \mathcal{D}_L(S) q. \end{aligned}$$

The evaluation loss $\mathcal{R}_L$ defined in equation 12 is therefore

$$\begin{aligned} \mathcal{R}_L &= \mathbb{E}[(\widehat{y}_L - y^*)^2] \\ &= \mathbb{E}_{U,q,O}[q^\top \mathcal{D}_L(S)^\top \Omega \mathcal{D}_L(S) q]. \end{aligned} \tag{26}$$

A.2 Constrained and Free Polynomial Depth Scaling

Repeated updates. Following the repeated-update construction of Bordelon et al. (2026), we impose isotropic matching and align the shared value map with the label embedding:

$$B_\ell = \kappa I_d \quad (\ell = 1, \dots, L), \qquad W_v W_y = v W_y. \tag{27}$$

Each degree-r product in equation 24 then equals $\kappa^r S^r$, and there are $\binom{L}{r}$ choices of r layers. Hence

$$\begin{aligned} \mathcal{P}_L(S) &= \sum_{r=1}^{L} \frac{(-1)^{r-1} m_r}{L^r} \binom{L}{r} \kappa^r S^r = p_L(S), \\ p_L(\lambda) &= \sum_{r=1}^{L} a_r \lambda^r, \qquad a_r = (-1)^{r-1} \binom{L}{r} (\kappa/L)^r m_r. \end{aligned} \tag{28}$$

In the infinite-context limit $P \to \infty$, we use $S \approx \Sigma$. Thus

$$I_d - p_L(\Sigma) = O \operatorname{diag}\big(1 - p_L(\lambda_1), \dots, 1 - p_L(\lambda_d)\big) O^\top.$$

Using $\mathbb{E}[qq^\top \mid O] = \Sigma$ in equation 26 gives the population-covariance approximation $\mathcal{R}_L^{\mathrm{pop}}$ to $\mathcal{R}_L$:

$$\begin{aligned} \mathcal{R}_L^{\mathrm{pop}} &= \mathbb{E}_O \operatorname{tr}\big[\mathcal{D}_L(\Sigma)^\top \Omega \mathcal{D}_L(\Sigma) \Sigma\big] \\ &= \mathbb{E}_O \operatorname{tr}\Big[O \operatorname{diag}\big(\lambda_k \omega_k [1 - p_L(\lambda_k)]^2\big)_{k=1}^{d} O^\top\Big] \\ &= \sum_{k=1}^{d} \lambda_k \omega_k [1 - p_L(\lambda_k)]^2. \end{aligned} \tag{29}$$

Define $r_L(\lambda) = 1 - p_L(\lambda)$. Since $p_L(0) = 0$, we have $r_L(0) = 1$, and therefore

$$\deg r_L \le L, \qquad r_L(0) = 1, \qquad \mathcal{R}_L^{\mathrm{pop}} = \sum_{k=1}^{d} \lambda_k \omega_k r_L(\lambda_k)^2. \tag{30}$$

For a specific data spectrum, write

$$\lambda_k = c_\lambda k^{-a}, \qquad \omega_k = c_\omega k^{-b},$$
$$a \geq 0, \qquad b \geq 0, \qquad \zeta = \frac{a+b-1}{a} \quad (a > 0).$$

where $c_\lambda, c_\omega > 0$ and $\lambda_1 = c_\lambda$. For $a > 0$, the change of variables from feature index to eigenvalue is

$$k = (c_\lambda/\lambda)^{1/a}, \qquad \left|\frac{dk}{d\lambda}\right| = \frac{c_\lambda^{1/a}}{a}\lambda^{-1-1/a}.$$

Substituting into equation 30 and replacing the sum by its continuum integral gives

$$\begin{aligned}\mathcal{R}_L^{\text{pop}} &= c_\lambda c_\omega \sum_{k=1}^{d} k^{-(a+b)} r_L(c_\lambda k^{-a})^2 \\ &\approx c_\lambda c_\omega \int_1^d k^{-(a+b)} r_L(c_\lambda k^{-a})^2 \, dk \\ &= \frac{c_\omega c_\lambda^{1-\zeta}}{a} \int_{\lambda_d}^{\lambda_1} \lambda^{\zeta-1} r_L(\lambda)^2 \, d\lambda.\end{aligned} \tag{31}$$

For $a + b > 1$, we have $\zeta > 0$, which makes the continuum weight integrable at zero. When d is sufficiently large, λ_d is close to zero and the finite-dimensional cutoff does not yet affect the loss. We may then approximate the dense spectrum by a measure on $[0, \lambda_1]$. This gives

$$\mathcal{R}_L^{\text{pop}} \propto \int_0^{\lambda_1} \lambda^{\zeta-1} r_L(\lambda)^2 \, d\lambda. \tag{32}$$

Because W_v is shared across layers,

$$m_r = W_o W_v^r W_y \propto v^r, \qquad \gamma = \kappa v.$$

Optimizing $W_o W_y$ to 1 allows the repeated-update construction to solve the easy task in the large-depth limit. Substituting into equation 28 gives

$$\begin{aligned}p_L(\lambda) = \sum_{r=1}^{L} (-1)^{r-1} \binom{L}{r} (\gamma\lambda/L)^r = 1 - (1 - \gamma\lambda/L)^L, \\ r_L(\lambda) = (1 - \gamma\lambda/L)^L.\end{aligned} \tag{33}$$

For the residual to vanish as $L \to \infty$, the scalar updates must contract rather than amplify positive-eigenvalue directions. At the largest eigenvalue this requires

$$|1 - \gamma\lambda_1/L| < 1 \quad \Longleftrightarrow \quad 0 < \gamma < \frac{2L}{\lambda_1}.$$

Using the optimal-scale estimate of Bordelon et al. (2026),

$$\gamma^* \approx \frac{L}{\lambda_1}, \qquad \frac{\gamma^*}{L} \approx \frac{1}{\lambda_1},$$

the continuum loss becomes, with $u = 2L\lambda/\lambda_1$,

$$\begin{aligned}\mathcal{R}_L^{\text{con}} &\approx \frac{c_\omega c_\lambda^{1-\zeta}}{a} \int_0^{\lambda_1} \lambda^{\zeta-1} (1 - \lambda/\lambda_1)^{2L} \, d\lambda \\ &\sim \frac{c_\omega c_\lambda^{1-\zeta}}{a} \left(\frac{\lambda_1}{2L}\right)^\zeta \int_0^\infty u^{\zeta-1} e^{-u} \, du \\ &\propto L^{-\zeta}.\end{aligned} \tag{34}$$

Here, $\mathcal{R}_L^{\text{con}}$ denotes the loss of the repeated-update reference within the population approximation. This exponent is consistent with the result of Bordelon et al. (2026).

Free polynomials. We then derive the depth-scaling exponent when the polynomial coefficients are free by removing the repeated-update conditions in equation 27. We assume only that the depth-L predictor is an arbitrary polynomial in Σ of degree at most L. Substitution into equation 26 gives the freely optimized population-polynomial reference $\mathcal{R}_L^{\text{free}}$:

$$\begin{gathered}\mathcal{P}_L(\Sigma) = p_L(\Sigma), \qquad \deg p_L \le L, \qquad p_L(0) = 0, \qquad S \approx \Sigma, \\ \mathcal{R}_L^{\text{pop}} = \sum_{k=1}^{d} \lambda_k \omega_k [1 - p_L(\lambda_k)]^2, \\ \mathcal{R}_L^{\text{free}} \propto \min_{\substack{\deg r \le L \\ r(0)=1}} \int_0^{\lambda_1} \lambda^{\zeta-1} r(\lambda)^2 \, d\lambda, \qquad r = 1 - p_L.\end{gathered} \tag{35}$$

Let ϕ_n be a degree-n polynomial orthonormal under this weight:

$$\int_0^{\lambda_1} \lambda^{\zeta-1} \phi_n(\lambda) \phi_m(\lambda) \, d\lambda = \delta_{nm},$$

where δ_{nm} equals one for $n = m$ and zero otherwise. Expanding the residual and using $r(0) = 1$ gives

$$\begin{gathered}r(\lambda) = \sum_{n=0}^{L} d_n \phi_n(\lambda), \qquad \sum_{n=0}^{L} d_n \phi_n(0) = 1, \\ \int_0^{\lambda_1} \lambda^{\zeta-1} r(\lambda)^2 \, d\lambda = \sum_{n=0}^{L} d_n^2.\end{gathered}$$

Cauchy–Schwarz now gives the minimum directly:

$$\begin{aligned}1 &= \left(\sum_{n=0}^{L} d_n \phi_n(0) \right)^2 \\ &\le \left(\sum_{n=0}^{L} d_n^2 \right) \left(\sum_{n=0}^{L} \phi_n(0)^2 \right).\end{aligned} \tag{36}$$

Equality is attained by coefficients proportional to the endpoint values:

$$d_n^* = \frac{\phi_n(0)}{\sum_{m=0}^{L} \phi_m(0)^2}, \qquad \mathcal{R}_L^{\text{free}} \propto \left[\sum_{n=0}^{L} \phi_n(0)^2 \right]^{-1}. \tag{37}$$

For this weight, the basis consists of shifted Jacobi polynomials:

$$\phi_n(\lambda) = \sqrt{\frac{2n+\zeta}{\lambda_1^\zeta}} \, P_n^{(\zeta-1,0)}(1 - 2\lambda/\lambda_1), \tag{38}$$

where $P_n^{(\zeta-1,0)}$ is the degree-n Jacobi polynomial. Using its value at 1 gives

$$\begin{aligned}\phi_n(0)^2 &= \frac{2n+\zeta}{\lambda_1^\zeta} \left[\frac{\Gamma(n+\zeta)}{\Gamma(\zeta)\Gamma(n+1)} \right]^2 \\ &\sim \frac{2}{\lambda_1^\zeta \Gamma(\zeta)^2} \, n^{2\zeta-1}.\end{aligned}$$

Consequently,

$$\begin{aligned}\sum_{n=0}^{L} \phi_n(0)^2 &\sim \frac{2}{\lambda_1^\zeta \Gamma(\zeta)^2} \int_0^L n^{2\zeta-1} \, dn \\ &= \frac{L^{2\zeta}}{\zeta \lambda_1^\zeta \Gamma(\zeta)^2}.\end{aligned} \tag{39}$$

Table 2: Measured depth-scaling exponents of the linear-attention model on the easy task lie between two reference values.

(a, b)	Constrained ζ	Measured α	Free 2ζ
$(2, 0.5)$	0.750	1.121	1.500
$(2, 1)$	1.000	1.485	2.000
$(2, 1.5)$	1.250	1.684	2.500
$(2, 2)$	1.500	1.987	3.000
$(0.5, 2)$	3.000	3.310	6.000
$(1, 2)$	2.000	2.552	4.000
$(1.5, 2)$	1.667	2.163	3.333

The two polynomial families therefore give

$$\mathcal{R}_L^{\text{con}} \propto L^{-\zeta}, \qquad \mathcal{R}_L^{\text{free}} \propto L^{-2\zeta}, \qquad \zeta = \frac{a+b-1}{a}. \tag{40}$$

Table 2 compares these reference exponents with the measured values.

The actual model is less constrained than the repeated-update construction of Bordelon et al. (2026), which relies on the two conditions in equation 27, but it is not as unconstrained as the idealized free polynomial, whose coefficients are optimized independently. It is therefore reasonable for the measured exponents to lie between these two reference values.

Note that this theoretical analysis has several requirements, including the population-covariance approximation $S \approx \Sigma$, corresponding to $P \to \infty$; a dense spectrum with $\zeta > 0$; and a sufficiently large spectral range λ_1/λ_d. At the evaluated depths, repeated updates require $1 \ll L \ll \lambda_1/\lambda_d$, whereas free polynomials require $1 \ll L \ll \sqrt{\lambda_1/\lambda_d}$.

A.3 Predicting the Linear-Model Loss Floor

In this subsection, we characterize the predictable component of the difficult-task target and the resulting irreducible loss floor for the linear-attention model. We further validate these theoretical predictions using Monte Carlo estimates.

For a test token q sampled independently of the context inputs, the difficult-task target is

$$\begin{aligned} S_k(q) &= \mathrm{TopK}_{1\le i\le P} \frac{q^\top \boldsymbol{x}_i}{\|q\|\,\|\boldsymbol{x}_i\|}, \\ t = \frac{1}{k}\sum_{i\in S_k(q)} \boldsymbol{x}_i \in \mathbb{R}^d, &\qquad y^* = \boldsymbol{\beta}^\top t. \end{aligned} \tag{41}$$

From Appendix A.1, assuming $S \approx \Sigma$,

$$\widehat{y}_L = \boldsymbol{\beta}^\top p_L(\Sigma) q, \qquad p_L(\Sigma) = \sum_{r=1}^{L} a_r \Sigma^r. \tag{42}$$

Write $u_L(q) = p_L(\Sigma)q$. Since $\widehat{y}_L - y^* = \boldsymbol{\beta}^\top(u_L - t)$, averaging first over the independent teacher gives

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\beta}}[(\widehat{y}_L - y^*)^2 \mid U, q, O] &= (t-u_L)^\top \mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\top \mid O](t-u_L) \\ &= (t-u_L)^\top \Omega (t-u_L). \end{aligned}$$

Thus $\mathcal{R}_L^{\text{pop}} = \mathbb{E}\|t-u_L\|_\Omega^2$, where $\|z\|_\Omega^2 = z^\top \Omega z$. At fixed q, O, the selected target t varies across contexts, whereas u_L does not. Separate its conditional mean from this variation:

$$\begin{aligned} \bar{t} = \mathbb{E}[t \mid q, O], &\qquad \delta = t - \bar{t}, \\ \mathbb{E}[\delta \mid q, O] &= 0. \end{aligned}$$

Substituting $t - u_L = \delta + (\bar{t} - u_L)$ and expanding gives

$$\|t-u_L\|_\Omega^2 = \|\delta\|_\Omega^2 + \|\bar{t}-u_L\|_\Omega^2 + 2\delta^\top \Omega(\bar{t}-u_L).$$

The cross term vanishes after conditioning because Ω and $\bar{t} - u_L$ are fixed given q, O:

$$\begin{aligned}\mathbb{E}\big[\delta^\top \Omega\big(\bar{t} - p_L(\Sigma)q\big)\big] &= \mathbb{E}\big[\mathbb{E}[\delta \mid q, O]^\top \Omega\big(\bar{t} - p_L(\Sigma)q\big)\big] \\ &= 0.\end{aligned}$$

Hence the population risk separates into

$$\begin{aligned}\mathcal{R}_L^{\mathrm{pop}} &= \mathbb{E}\|t - p_L(\Sigma)q\|_\Omega^2 \\ &= \underbrace{\mathbb{E}\|\delta\|_\Omega^2}_{\text{context-specific variation}} + \underbrace{\mathbb{E}\|\bar{t} - p_L(\Sigma)q\|_\Omega^2}_{\text{polynomial approximation error}}.\end{aligned} \tag{43}$$

We denote the first term by $\mathcal{R}_{\mathrm{irr}}$. It is independent of the polynomial degree, whereas increasing the degree can reduce only the second term.

Predictable component of the difficult-task target. Section 4.1 states that the predictable component is approximately proportional to Σq. We now derive this result by approximating cosine-similarity ranking with the inner-product scores $s_i = q^\top \boldsymbol{x}_i$. Conditional on q, O, the pair $(\boldsymbol{x}_i, s_i)$ is jointly Gaussian, with

$$\mathbb{E}[s_i \mid q, O] = 0, \qquad \mathrm{Var}(s_i \mid q, O) = q^\top \Sigma q =: V(q), \qquad \mathrm{Cov}(\boldsymbol{x}_i, s_i \mid q, O) = \Sigma q.$$

The Gaussian conditioning formula therefore gives

$$\mathbb{E}[\boldsymbol{x}_i \mid s_i, q, O] = \frac{\mathrm{Cov}(\boldsymbol{x}_i, s_i \mid q, O)}{\mathrm{Var}(s_i \mid q, O)} s_i = \frac{\Sigma q}{V(q)} s_i. \tag{44}$$

Let $s_{(1)} \geq \cdots \geq s_{(P)}$ be the ordered scores, and let $\boldsymbol{x}_{(j)}$ denote the input attaining $s_{(j)}$. Since $z_i = s_i / \sqrt{V(q)}$ are independent standard normal variables, define

$$\gamma_{P,k} = \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} z_{(j)}\right].$$

Applying equation 44 to the k selected inputs and averaging over their ordered scores gives

$$\begin{aligned}\bar{t}(q) &= \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} \boldsymbol{x}_{(j)} \;\middle|\; q, O\right] \\ &= \frac{\Sigma q}{V(q)} \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} s_{(j)} \;\middle|\; q, O\right] = \gamma_{P,k} \frac{\Sigma q}{\sqrt{q^\top \Sigma q}}.\end{aligned} \tag{45}$$

To remove the remaining query-dependent denominator, note that for $q \sim \mathcal{N}(0, \Sigma)$,

$$\mathbb{E}[V(q) \mid O] = \mathrm{tr}(\Sigma^2), \qquad \mathrm{Var}(V(q) \mid O) = 2\,\mathrm{tr}(\Sigma^4).$$

When $2\,\mathrm{tr}(\Sigma^4)/\,\mathrm{tr}(\Sigma^2)^2 \ll 1$, we can replace $V(q)$ by its mean. Combining this with the cosine-to-score approximation yields

$$\bar{t}(q) \approx c\Sigma q, \qquad c = \frac{\gamma_{P,k}}{\sqrt{\mathrm{tr}(\Sigma^2)}}.$$

Since the teacher is independent of the selected inputs given q, O,

$$\mathbb{E}[y^* \mid q, \boldsymbol{\beta}, O] = \boldsymbol{\beta}^\top \mathbb{E}[t \mid q, O] \approx c\, \boldsymbol{\beta}^\top \Sigma q. \tag{46}$$

Thus, the predictable component is proportional to Σq and can already be represented by the degree-one polynomial $p_1(\Sigma) = c\Sigma$. Higher polynomial degrees cannot remove the context-specific variation in equation 43, explaining both the rapid saturation and the nonzero loss floor of the linear-attention model.

Table 3: The theoretically predicted irreducible loss closely matches the actual loss plateau of the linear-attention model.

(a, b)	Theory-predicted loss	Measured loss	Difference
$k = 2$			
$(0, 0)$	59.3714 ± 0.0213	59.8502	0.80%
$(0.25, 0.25)$	9.7004 ± 0.0049	9.7387	0.39%
$(0.5, 0.5)$	2.2715 ± 0.0061	2.3274	2.40%
$(1, 0.25)$	1.2246 ± 0.0106	1.2259	0.11%
$(1, 1)$	0.4975 ± 0.0102	0.5036	1.22%
$k = 16$			
$(0.5, 0.5)$	0.3218 ± 0.0042	0.3083	4.38%
$(1, 0)$	0.4481 ± 0.0086	0.4452	0.67%
$(1, 0.25)$	0.3186 ± 0.0085	0.3222	1.11%
$(1.25, 0)$	0.3595 ± 0.0090	0.3659	1.75%

An explicit expression for the irreducible loss. Let $\mathbf{a} = (a_1, \ldots, a_L)^\top$ collect the polynomial coefficients and define

$$\Phi_L(q) = [\Sigma q, \ldots, \Sigma^L q], \qquad p_L(\Sigma)q = \Phi_L(q)\mathbf{a}. \tag{47}$$

The corresponding population loss expands as

$$\begin{aligned} \mathcal{R}_L^{\text{pop}}(\mathbf{a}) &= \mathbb{E}\|t - \Phi_L(q)\mathbf{a}\|_\Omega^2 \\ &= \mathbb{E}[t^\top \Omega t] - 2\mathbf{a}^\top \mathbb{E}[\Phi_L(q)^\top \Omega t] \\ &\quad + \mathbf{a}^\top \mathbb{E}[\Phi_L(q)^\top \Omega \Phi_L(q)]\mathbf{a} \\ &= T - 2\mathbf{a}^\top h + \mathbf{a}^\top G\mathbf{a}, \end{aligned} \tag{48}$$

where

$$T = \mathbb{E}[t^\top \Omega t], \qquad h = \mathbb{E}[\Phi_L(q)^\top \Omega t], \qquad G = \mathbb{E}[\Phi_L(q)^\top \Omega \Phi_L(q)]. \tag{49}$$

In particular, the entries of the Gram matrix are determined by the spectra:

$$G_{rs} = \mathbb{E}[(\Sigma^r q)^\top \Omega (\Sigma^s q)] = \operatorname{tr}(\Omega \Sigma^{r+s+1}), \qquad 1 \le r, s \le L. \tag{50}$$

Differentiating equation 48 with respect to $\mathbf{a}$ gives the normal equations $G\mathbf{a} = h$. Hence

$$\mathbf{a}_* = G^\dagger h, \qquad \mathcal{R}_{L,*}^{\text{pop}} = T - h^\top G^\dagger h. \tag{51}$$

The decomposition in equation 43 also gives

$$\min_{\mathbf{a}} \mathcal{R}_L^{\text{pop}}(\mathbf{a}) = \mathcal{R}_{\text{irr}} + \min_{\mathbf{a}} \mathbb{E}\|\bar{t} - \Phi_L(q)\mathbf{a}\|_\Omega^2. \tag{52}$$

Under the degree-one approximation $\bar{t} \approx c\Sigma q$ in equation 46, the second term is negligible. Combining equation 51 and equation 52 therefore gives the explicit prediction

$$\mathcal{R}_{\text{irr}} \approx T - h^\top G^\dagger h. \tag{53}$$

We evaluate this prediction by Monte Carlo sampling in the original cosine task. We independently sample n query–rotation–context triples. For sample i, t_i is the resulting cosine-selected target and $\Omega_i = O_i \operatorname{diag}(\omega_1, \ldots, \omega_d) O_i^\top$. The hats below denote empirical estimates computed from these samples:

$$\begin{aligned} \widehat{T} &= \frac{1}{n}\sum_{i=1}^n t_i^\top \Omega_i t_i, \qquad \widehat{h} = \frac{1}{n}\sum_{i=1}^n \Phi_L(q_i)^\top \Omega_i t_i, \\ \widehat{G} &= \frac{1}{n}\sum_{i=1}^n \Phi_L(q_i)^\top \Omega_i \Phi_L(q_i), \qquad \widehat{\mathcal{R}}_{\text{irr}} = \widehat{T} - \widehat{h}^\top \widehat{G}^\dagger \widehat{h}. \end{aligned} \tag{54}$$

Table 3 compares this Monte Carlo prediction of equation 53 with the actual loss plateau of the trained model. Their close agreement validates the predicted irreducible floor.

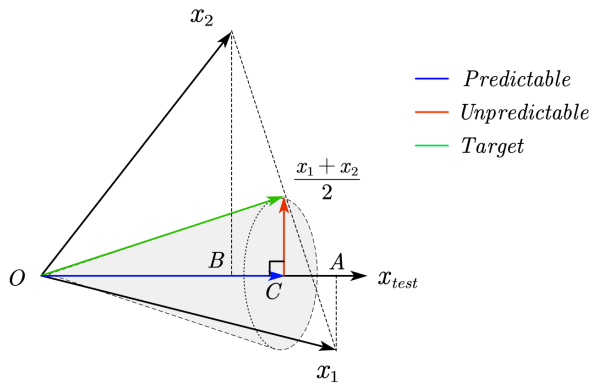


Figure 8: The linear-attention model captures only the average predictable component of the difficult-task target. The green target is the average of the two selected inputs. Blue shows the best predictable component $\overrightarrow{OC}$; red shows the context-specific residual.

Geometric interpretation. We now provide a heuristic geometric picture of the predictable component and irreducible loss, using the top-2 task as an example. In Figure 8, the horizontal ray represents the query x_{test}. The black vectors $\overrightarrow{Ox_1}$ and $\overrightarrow{Ox_2}$ are the two selected context inputs, and their average is the green target $\overrightarrow{Ot} = (x_1 + x_2)/2$. The blue vector $\overrightarrow{OC}$ represents the conditional mean $\bar{t}$, whereas the red vector $\overrightarrow{Ct} = t - \bar{t}$ is the context-specific residual. Linear attention can approximate the predictable blue component through the population covariance, but it lacks the context-specific selection required to follow the red residual for each realized context. This residual therefore produces the irreducible loss floor.

A.4 Power Laws from Mode-Wise Exponential Decay

As a heuristic model of Figure 3(c), we show how exponential decay at different rates can combine into a power law. Let feature k have initial loss weight $w_k \geq 0$ and characteristic depth $\tau_k > 0$:

$$\mathcal{R}_k(L) = w_k e^{-L/\tau_k}, \qquad \mathcal{R}_L = \sum_k w_k e^{-L/\tau_k}.$$

Assume the loss-weighted distribution of characteristic depths follows a power law, $\rho(\tau) = C\tau^{-\beta}$ for $\tau \geq \tau_0 > 0$, with $C > 0$ and $\beta > 1$. The aggregate loss is

$$\mathcal{R}_L \approx \int_{\tau_0}^{\infty} \rho(\tau) e^{-L/\tau}\, d\tau. \tag{55}$$

With $x = L/\tau$ and $d\tau = -Lx^{-2}\, dx$, this becomes

$$\begin{aligned} \mathcal{R}_L &\approx CL^{1-\beta} \int_0^{L/\tau_0} x^{\beta-2} e^{-x}\, dx \\ &\sim C\Gamma(\beta-1) L^{-(\beta-1)}, \qquad L \gg \tau_0. \end{aligned} \tag{56}$$

Thus exponential convergence of individual features produces power-law depth scaling when their characteristic depths have a broad power-law distribution. Stronger features converge first, while weaker features continue to contribute as depth increases.

B Nonlinear-Attention Model Theory

In this section, we derive the loss decomposition for the nonlinear-attention model on the difficult task studied in Section 4.2.

We first introduce the notation. Throughout this section, $\mathbb{E}$ denotes an expectation over contexts. Let L be the number of layers and P the number of context tokens. We use $\ell = 1, \ldots, L$ for layers and

$i = 1, \ldots, P$ for tokens. Write k for the number of labels selected by the top-k task. We measure the relevance of each context token to the test token using cosine similarity:

$$s_i = \frac{\boldsymbol{x}_{\text{test}}^\top \boldsymbol{x}_i}{\|\boldsymbol{x}_{\text{test}}\|_2 \|\boldsymbol{x}_i\|_2}, \qquad s_{i_{(1)}} \geq \cdots \geq s_{i_{(P)}}, \qquad S_k = \{i_{(1)}, \ldots, i_{(k)}\}. \tag{57}$$

Thus S_k contains the k context indices closest to the test token. For the label column $y = (y_1, \ldots, y_P)^\top$, define the target row $t_k = (t_{k,1}, \ldots, t_{k,P})$ by

$$t_{k,i} = \begin{cases} 1/k, & i \in S_k, \\ 0, & i \notin S_k, \end{cases} \qquad y^* = t_k y = \frac{1}{k} \sum_{i \in S_k} y_i. \tag{58}$$

Let $p_{\ell i}$ be the attention weight from the test token to context token i at layer ℓ, with softmax normalized over the P context sources since the test token does not attend to itself:

$$p_\ell = (p_{\ell 1}, \ldots, p_{\ell P}), \qquad \sum_{i=1}^{P} p_{\ell i} = 1. \tag{59}$$

With fixed context states, each layer reads the same label column y. The common scalar gain g_L includes the value map and readout:

$$\widehat{y}_L = \frac{1}{L} \sum_{\ell=1}^{L} f_\ell, \qquad f_\ell = g_L p_\ell y = g_L \sum_{i=1}^{P} p_{\ell i} y_i, \qquad g_L = W_o W_v W_y. \tag{60}$$

Define the layer error relative to the same target as $e_\ell = f_\ell - y^*$. For $L > 1$, assuming finite expected squared errors, let

$$A_L = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}[e_\ell^2], \qquad C_L = \frac{1}{L(L-1)} \sum_{\ell \neq m} \mathbb{E}[e_\ell e_m]. \tag{61}$$

The second sum is over ordered distinct pairs. Since $\widehat{y}_L - y^* = L^{-1} \sum_\ell e_\ell$, expanding the square gives

$$\begin{aligned} \mathcal{R}_L &= \mathbb{E}[(\widehat{y}_L - y^*)^2] \\ &= \frac{1}{L^2} \left[\sum_\ell \mathbb{E}[e_\ell^2] + \sum_{\ell \neq m} \mathbb{E}[e_\ell e_m] \right] \\ &= \frac{A_L}{L} + \frac{L-1}{L} C_L \\ &= C_L + \frac{A_L - C_L}{L}. \end{aligned} \tag{62}$$

Expanding pairwise differences and using equation 61,

$$\begin{aligned} \sum_{\ell \neq m} \mathbb{E}[(e_\ell - e_m)^2] &= 2(L-1) \sum_\ell \mathbb{E}[e_\ell^2] - 2 \sum_{\ell \neq m} \mathbb{E}[e_\ell e_m] \\ &= 2L(L-1)(A_L - C_L). \end{aligned}$$

Dividing by $2L(L-1)$ gives

$$A_L - C_L = \frac{1}{2L(L-1)} \sum_{\ell \neq m} \mathbb{E}[(e_\ell - e_m)^2] \geq 0. \tag{63}$$

C_L captures the error shared across layers, whereas $A_L - C_L$ measures their disagreement. Averaging reduces the latter contribution to $(A_L - C_L)/L$. The shared component cannot be removed through aggregation and therefore gives rise to the irreducible loss, while differences among layer errors can partially cancel and produce the reducible loss.

Importantly, this decomposition does not by itself imply inverse-depth scaling. We first establish the emergence of inverse-depth scaling empirically through systematic experiments and rigorous power-law fitting, and then use the decomposition to identify the components and mechanisms underlying this behavior.

This decomposition relies on the parallel interpretation of our controlled model: the layers act as independently parameterized predictors whose aggregate forms the final output. A layerwise decomposition is therefore natural in this setting. In the generalized settings considered below, however, it is neither the only possible decomposition nor necessarily the most natural one. For example, when inputs and labels share a single stream, the model becomes cascaded and each layer depends on the preceding layers. A mean-field treatment (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018) may then be more appropriate, separating each layer's output into a mean component coupled to the other layers and a layer-specific fluctuation. Developing a rigorous decomposition for more complex systems, including large language models, remains an important direction for understanding the role of depth in realistic models.

C Generalized Model Experiments and Analysis

We now relax some design choices from the controlled model in Section 2.2 to move the architecture closer to standard Transformers and LLMs. First, the original two-stream encoding separates token matching from label propagation: queries and keys interact only with the input stream, whereas values and the output map interact only with the label stream. Second, all layers share a value matrix, which improves training stability and keeps the model concise. Third, the nonlinear-attention model initially keeps the context states fixed, isolating the process by which different layers retrieve relevant information from the context. In the following subsections, we relax these assumptions one at a time and conduct analogous experiments and analyses for each extension.

C.1 Encoding Input and Label Information into One Space

Section 2.2 initializes separate input and label streams, $H_x = W_x X$ and $H_y^0 = W_y \boldsymbol{y}^{\text{in}}$. In standard LLMs, token and label information are commonly encoded in a shared representation space. We therefore embed both sources of information in a shared token state:

$$H^0 = W_x X + W_y \boldsymbol{y}^{\text{in}} \in \mathbb{R}^{N \times (P+1)}. \tag{64}$$

Queries, keys, and values are now all computed from the same evolving state:

$$Q^\ell = W_q^\ell H^{\ell-1}, \qquad K^\ell = W_k^\ell H^{\ell-1}, \qquad V^\ell = W_v H^{\ell-1}. \tag{65}$$

For nonlinear attention, the shared state is updated as

$$\begin{aligned} A^\ell &= \mathsf{M} \odot \operatorname{softmax}\big((Q^\ell)^\top K^\ell\big), \\ H^\ell &= H^{\ell-1} + \frac{1}{L} V^\ell (A^\ell)^\top. \end{aligned} \tag{66}$$

Thus token matching, information propagation, and residual updates all act within a common representation space, and each layer operates on the state produced by the preceding layers.

To define the two decomposition terms in this cascaded model, let $z_\ell = W_o h_{\text{test}}^\ell$ denote the test-token readout after layer ℓ. Define the base readout and layer increments as

$$b_L = z_0, \qquad d_\ell = z_\ell - z_{\ell-1}, \qquad \widehat{y}_L = b_L + \sum_{\ell=1}^{L} d_\ell. \tag{67}$$

We can express the final prediction as an average of effective layerwise predictors by setting

$$f_\ell = b_L + L d_\ell, \qquad e_\ell = f_\ell - y^*, \qquad \widehat{y}_L - y^* = \frac{1}{L} \sum_{\ell=1}^{L} e_\ell. \tag{68}$$

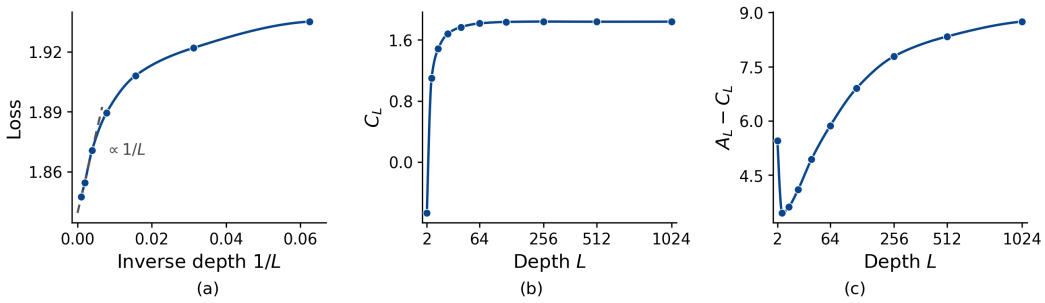


Figure 9: The single-stream nonlinear-attention model exhibits a similar inverse-depth scaling signal. (a) The model is evaluated under isotropic data spectra. (b,c) The shared-error term C_L and disagreement term $A_L - C_L$ both approach finite constants with depth.

For $L > 1$, the shared-error and disagreement terms are then defined as

$$\begin{aligned} C_L &= \frac{1}{L(L-1)} \sum_{\ell \neq m} \mathbb{E}[e_\ell e_m], \\ A_L - C_L &= \frac{1}{2L(L-1)} \sum_{\ell \neq m} \mathbb{E}[(e_\ell - e_m)^2] \\ &= \frac{1}{L-1} \sum_{\ell=1}^{L} \mathbb{E}[(f_\ell - \widehat{y}_L)^2]. \end{aligned} \tag{69}$$

With these definitions, the loss retains the exact identity

$$\mathcal{R}_L = C_L + \frac{A_L - C_L}{L}. \tag{70}$$

We evaluate this generalized model on the difficult task. Figure 9 shows that C_L and $A_L - C_L$ continue to approach finite constants, while the loss exhibits an approximate inverse-depth trend beyond moderate depth.

C.2 Layer-Decoupled Value Matrices

In standard LLMs, value matrices are typically parameterized independently across layers. We therefore replace the value matrix W_v shared in Section 2.2 with independently parameterized matrices W_v^ℓ, giving

$$Q^\ell = W_q^\ell H_x, \qquad K^\ell = W_k^\ell H_x, \qquad V^\ell = W_v^\ell H_y^{\ell-1}. \tag{71}$$

This modification allows each layer to apply a distinct transformation to the retrieved label information.

Let p_ℓ denote the attention weights from the test token to the context tokens, and let y collect their labels. The corresponding layerwise predictor and error are

$$g_\ell = W_o W_v^\ell W_y, \qquad f_\ell = g_\ell p_\ell y, \qquad e_\ell = f_\ell - y^*, \qquad \widehat{y}_L = \frac{1}{L} \sum_{\ell=1}^{L} f_\ell. \tag{72}$$

For $L > 1$, the shared-error and disagreement terms are

$$\begin{aligned} C_L &= \frac{1}{L(L-1)} \sum_{\ell \neq m} \mathbb{E}[e_\ell e_m], \\ A_L - C_L &= \frac{1}{2L(L-1)} \sum_{\ell \neq m} \mathbb{E}[(e_\ell - e_m)^2] \\ &= \frac{1}{2L(L-1)} \sum_{\ell \neq m} \mathbb{E}\Big[\big((g_\ell p_\ell - g_m p_m) y\big)^2\Big]. \end{aligned} \tag{73}$$

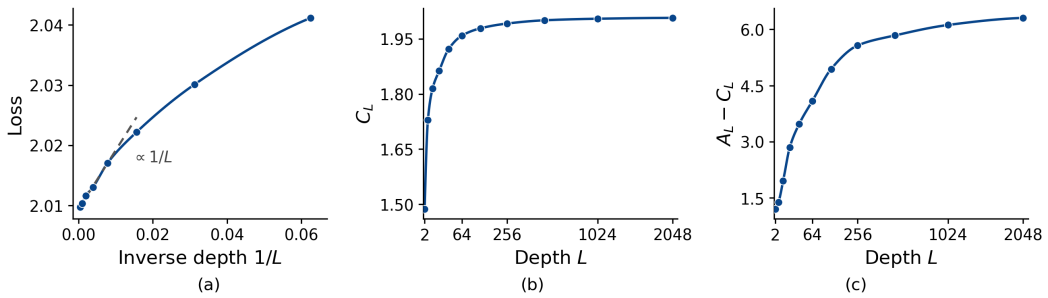


Figure 10: The nonlinear-attention model with layer-specific value matrices exhibits a similar inverse-depth scaling signal. (a) The model is evaluated under isotropic data spectra. (b,c) The shared-error term C_L and disagreement term $A_L - C_L$ both approach finite constants with depth.

The loss therefore retains the exact identity

$$\mathcal{R}_L = C_L + \frac{A_L - C_L}{L}. \tag{74}$$

We evaluate this generalized model on the difficult task. Figure 10 shows that C_L and $A_L - C_L$ continue to approach finite constants, while the loss displays an approximate inverse-depth trend beyond moderate depth. Thus, the same inverse-depth scaling signal remains visible after relaxing the shared-value assumption.

C.3 Allowing Context Updates in the Nonlinear-Attention Model

Our linear-attention model in Section 2.2 follows Bordelon et al. (2026), including their attention-masking scheme. Let $\mathbf{1}_P \in \mathbb{R}^P$ denote the all-ones vector. For a single test token, their mask can be written as

$$\mathsf{M} = \begin{pmatrix} -\mathbf{1}_P\mathbf{1}_P^\top & \mathbf{0}_{P\times 1} \\ \mathbf{1}_P^\top & 0 \end{pmatrix}, \tag{75}$$

where the upper-left block governs context-to-context updates, while the lower-left block allows the test token to attend to the context tokens. Intuitively, this sign structure lets linear attention extract data structure at progressively finer spectral resolutions. The negative context-to-context block preferentially subtracts components shared by highly similar context tokens from their label states. Dominant spectral features produce larger similarity-weighted updates and are therefore resolved earlier, whereas weaker features persist until greater depth and can be learned in deeper layers. This behavior is consistent with the sequential mode learning in Figure 3(c). By contrast, the positive test-to-context block allows retrieved information to accumulate in the test token's label state, which is initialized at zero and updated across layers.

To isolate how the test token retrieves and aggregates information from the context, the controlled nonlinear-attention model in Section 2.2 initially disables context updates by setting the upper-left block to zero:

$$\mathsf{M}(0) = \begin{pmatrix} \mathbf{0}_{P\times P} & \mathbf{0}_{P\times 1} \\ \mathbf{1}_P^\top & 0 \end{pmatrix}. \tag{76}$$

Standard Transformers, however, update context representations across layers. To examine how these updates affect the nonlinear-attention model, we parameterize the previously zero context-to-context block by a trainable scalar m. Specifically, we use

$$\mathsf{M}(m) = \begin{pmatrix} m\mathbf{1}_P\mathbf{1}_P^\top & \mathbf{0}_{P\times 1} \\ \mathbf{1}_P^\top & 0 \end{pmatrix}, \tag{77}$$

where the scalar m is shared across all layers and learned jointly with the other model parameters.

We evaluate this generalized model on the difficult task with isotropic data spectra, using top-2 and top-16 targets. Figure 11(a,b) shows qualitatively different depth dependence. For top-2, the loss continues to decrease substantially with depth and exhibits a signal consistent with inverse-depth scaling. For top-16, by contrast, increasing depth does not improve the loss.

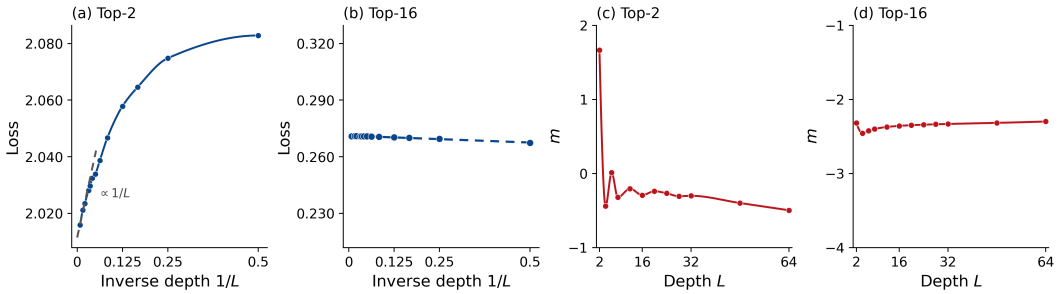


Figure 11: Enabling context updates preserves an inverse-depth scaling signal on the top-2 task but not on the top-16 task. Under isotropic data spectra, (a) the top-2 loss exhibits an inverse-depth scaling signal, whereas (b) the top-16 loss quickly saturates. (c,d) The learned context-update scalar m varies with depth for top-2 but stabilizes for top-16.

The learned context-update scalar m provides an additional diagnostic. Figure 11(c,d) shows that the learned scalar m varies substantially with depth for top-2, whereas for top-16 it remains near the stable negative value -2.3.

To interpret this task-dependent behavior, we first derive its exact effect in a two-layer model. Write the label state as $H_y^\ell = [Z_\ell, h_\ell]$, where $Z_\ell \in \mathbb{R}^{N\times P}$ contains the context states and $h_\ell \in \mathbb{R}^N$ is the test-token state. Let $y = (y_1, \ldots, y_P)^\top$, so that $Z_0 = W_y y^\top$ and $h_0 = 0$. Without changing the mechanism of interest we take $W_v = I_N$ for simplicity. Let $p_\ell \in \mathbb{R}^{1\times P}$ denote the test-to-context attention row at layer ℓ, and let $A_1 \in \mathbb{R}^{P\times P}$ denote the first layer's context-to-context attention matrix.

Applying equation 10 with the mask in equation 77 gives

$$\begin{aligned} Z_1 &= Z_0 + \frac{m}{2} Z_0 A_1^\top, \\ h_1 &= \frac{1}{2} Z_0 p_1^\top, \\ h_2 &= h_1 + \frac{1}{2} Z_1 p_2^\top. \end{aligned} \tag{78}$$

Substituting $Z_0 = W_y y^\top$ gives

$$h_2 = W_y \left[\frac{1}{2}(p_1 + p_2) y + \frac{m}{4} p_2 A_1 y \right]. \tag{79}$$

Defining the scalar readout gain $g = W_o W_y$, the prediction is

$$\widehat{y}_2 = \frac{g}{2}(p_1 + p_2) y + \frac{gm}{4} p_2 A_1 y. \tag{80}$$

The first term reads the original context labels directly. The second reads them after A_1 has mixed information across context tokens.

For arbitrary depth, the corresponding recursion is

$$\begin{aligned} Z_\ell &= Z_{\ell-1}\left(I_P + \frac{m}{L} A_\ell^\top\right), \\ h_\ell &= h_{\ell-1} + \frac{1}{L} Z_{\ell-1} p_\ell^\top. \end{aligned} \tag{81}$$

Consequently,

$$\widehat{y}_L = \frac{g}{L}\left[p_1 y + \sum_{\ell=2}^{L} p_\ell \left(I_P + \frac{m}{L} A_{\ell-1}\right) \cdots \left(I_P + \frac{m}{L} A_1\right) y \right]. \tag{82}$$

Thus, context updates affect prediction through operators $I_P + (m/L)A_\ell$. Whether and how they improve depth utilization depends on the spectral structure of A_ℓ.

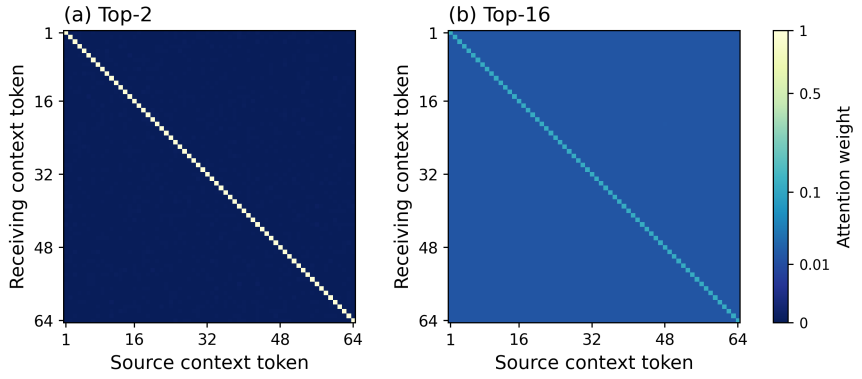


Figure 12: The learned context-attention structure depends strongly on the aggregation target. The context-attention matrix A_1 is (a) sharply concentrated on the diagonal for top-2 task, but (b) distributed broadly across context tokens for top-16 task.

Figure 12 visualizes the two context attention structures. For the top-2 task, the measured A_1 is nearly the identity. For the top-16 task, the measured A_1 is broader and flatter. This contrast is induced by the sharpness of the target distribution. The sharp top-2 target requires the model to preserve the identities of a few selected tokens, favoring well-separated embeddings and a nearly identity attention map. The flatter top-16 target instead favors stronger coupling across embeddings and therefore produces a broader attention map.

To quantify this contrast, we model each measured attention matrix by an identity component I_P and a uniform mixing component J_P. Specifically, for $k \in \{2, 16\}$, we fit ρ_k by least squares in

$$A_1^{(k)} \approx \rho_k I_P + (1 - \rho_k) J_P, \qquad J_P := \frac{\mathbf{1}_P \mathbf{1}_P^\top}{P}. \tag{83}$$

The identity term preserves each context embedding separately, whereas J_P replaces every embedding by their common average. The fitted results are

$$\begin{aligned} \rho_2 &= 0.9999, \quad & R^2_{\text{top-2}} &> 0.99999, \\ \rho_{16} &= 0.0989, \quad & R^2_{\text{top-16}} &= 0.99996. \end{aligned} \tag{84}$$

This model therefore confirms that the top-2 matrix is almost exactly the identity, while the top-16 matrix is dominated by uniform cross-token mixing.

For top-2, the near-identity structure gives $p_2 A_1 \approx p_2$, and

$$\widehat{y}_2 \approx \frac{g}{2} p_1 y + \frac{g}{2}(1 + m/2) p_2 y. \tag{85}$$

The context update therefore introduces no independent token-weighting direction: the prediction remains an aggregation of p_1 and p_2. Accordingly, its loss remains close to that of the fixed-context model and still exhibits approximate inverse-depth scaling over the observed depth range.

For top-16, let $u = \mathbf{1}_P^\top / P$. Since $p_2 \mathbf{1}_P = 1$, we have $p_2 J_P = u$ and hence

$$p_2 A_1 \approx \rho_{16} p_2 + (1 - \rho_{16}) u, \tag{86}$$

Equation 86 shows that the context update adds the uniform token-weighting direction u, beyond the direct rows p_1 and p_2. Defining $\bar{p} = (p_1 + p_2)/2$, the exact two-layer prediction is

$$\widehat{y}_2 = g \left(\bar{p} + \frac{m}{4} p_2 A_1 \right) y. \tag{87}$$

A negative m therefore produces a global shift that subtracts the common attention floor, rather than merely combining more direct attention rows. This additional degree of freedom allows the model to achieve low loss with few layers, explaining its early saturation. The same argument extends to models with more than two layers.

This distinction is also relevant to language modeling. Many language tasks require focusing on only a few relevant tokens and learning a sparse and peaked distribution. In such regimes, context

interactions may primarily reweight information that has already been selected. Depth utilization can therefore remain governed by selective retrieval and cross-layer aggregation, making the inverse-depth mechanism relevant even when context representations are not fixed.

Overall, these generalized models exhibit inverse-depth scaling signals that are qualitatively consistent with the controlled-model analysis over the accessible depth ranges. Training these models at much greater depths is substantially more computationally expensive, preventing experiments on the same scale as those conducted for the controlled model. We therefore interpret these results as evidence of similar scaling signals, rather than as a definitive guarantee of asymptotic inverse-depth scaling. Their appearance under more realistic architectural choices nevertheless suggests that the mechanism may remain relevant to practical models.

D More Experiments and Fitting Details

This section supplements the controlled-model results with additional experiments and provides the fitting procedures used to quantify the observed depth-scaling laws. We report fits for both the easy and difficult tasks, analyze context-dependent token selection, test broader targets and data spectra, and examine the convergence of the shared-error term.

D.1 Power-Law Fitting Procedures

We use a common log-space fitting procedure to quantify the depth-scaling laws in Sections 3.1 and 4.2. Each spectrum is fitted independently using

$$\mathcal{R}_L = L_0 + cL^{-\alpha}, \qquad c > 0, \quad \alpha > 0, \tag{88}$$

where L_0 is the asymptotic loss floor and α is the decay exponent.

For observations at depths $\{L_i\}$, we minimize the squared residuals between the logarithms of the observed and modeled losses:

$$(\widehat{L}_0, \widehat{c}, \widehat{\alpha}) = \arg\min_{L_0,c,\alpha} \sum_i \left[\log \mathcal{R}_{L_i} - \log\left(L_0 + cL_i^{-\alpha}\right)\right]^2. \tag{89}$$

For the easy task, we set $L_0 = 0$. As shown in Appendix A.1, given sufficient depth, width, dataset size, context length, and training time, the linear-attention model can solve this task exactly by using a covariance polynomial to approximate the identity map. For the difficult task, the decomposition in Appendix B contains an error component shared across layers, so we estimate L_0 jointly with c and α.

D.2 Complementary Token Selection Across Contexts

We further compare how nonlinear-attention models of different depths perform on the difficult task. We train an $L = 1$ model and an $L = 256$ model on the difficult task with $k = 2$ under identical conditions. We then sample 10,240 fresh contexts independently from the data distribution and evaluate both models on the same set.

For each context, we order the tokens by their similarity to the test token. Under the parallel interpretation developed in Appendix B, the final output aggregates the contributions of the individual layers. We therefore summarize their collective token selection by the aggregated attention distribution and compare it with the target distribution over the ordered tokens. Formally,

$$\begin{gathered}\bar{p}_L(c) = \frac{1}{L}\sum_{\ell=1}^{L} p_\ell(c), \qquad p^* = (1/2, 1/2, 0, \ldots, 0), \\ E_L(c) = \|\bar{p}_L(c) - p^*\|_2^2.\end{gathered} \tag{90}$$

Lower values indicate more accurate identification and weighting of the two target tokens. For each model separately, we sort the sampled contexts by $E_L(c)$ and denote the qth percentile by $Q_L(q)$. Thus, small q corresponds to contexts on which a model most closely matches the target distribution, whereas large q probes its high-error tail.

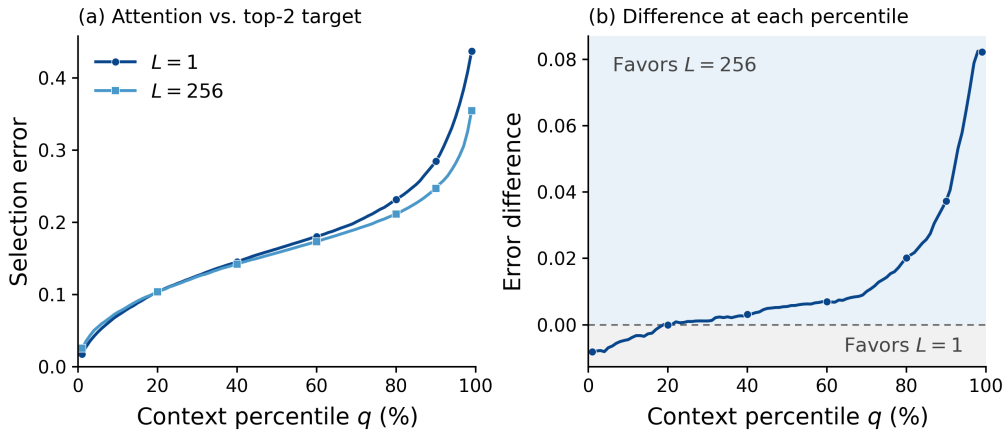


Figure 13: Greater depth makes token selection more reliable across contexts. (a) Selection-loss percentiles for $L = 1$ and $L = 256$. (b) Their difference $\Delta Q(q) = Q_1(q) - Q_{256}(q)$; positive values favor $L = 256$.

Figure 13(a) compares the resulting error distributions. Panel (b) reports

$$\Delta Q(q) = Q_1(q) - Q_{256}(q), \tag{91}$$

Positive $\Delta Q(q)$ therefore indicates that the deeper model has lower selection loss at percentile q. The one-layer model is slightly more accurate on the contexts in its lowest-error tail, but its error grows much more rapidly across harder contexts. The 256-layer model instead markedly reduces the upper tail, indicating more reliable token selection across the context distribution.

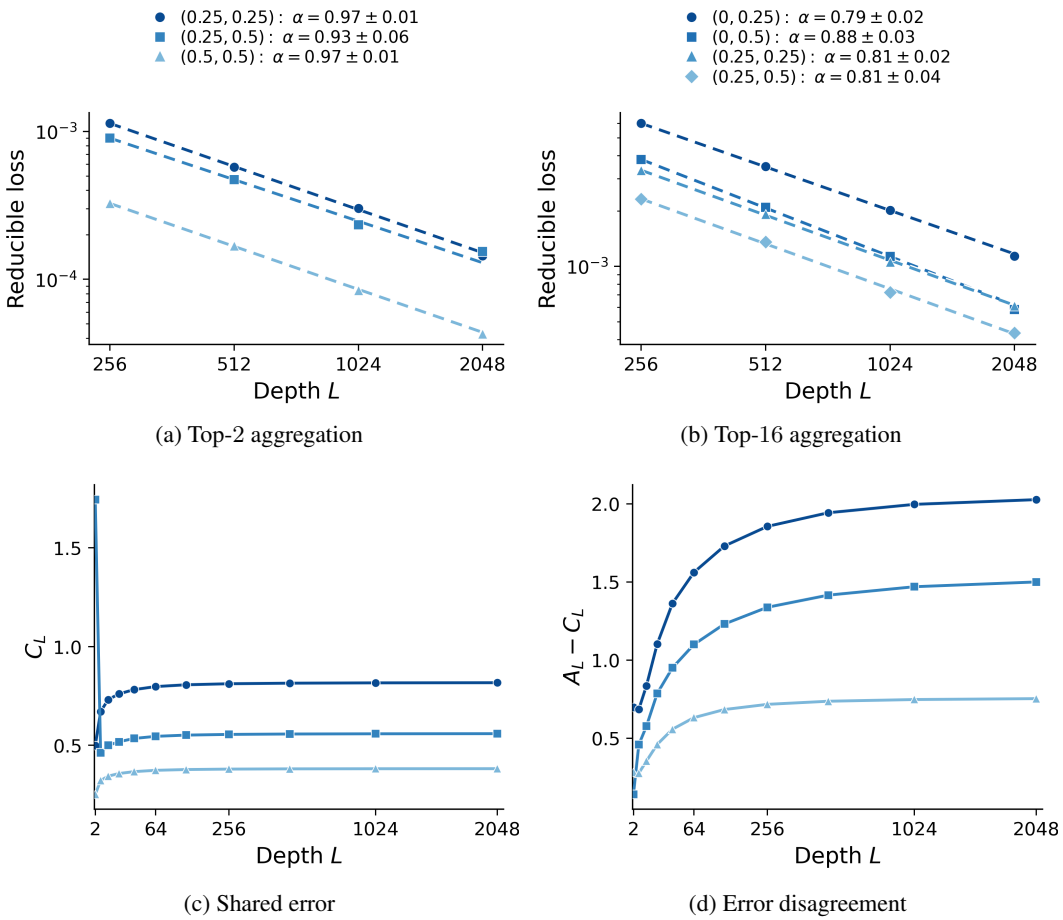
This pattern follows from the fact that each query–key parameterization must serve the entire context distribution. With only one layer, a single selection rule may fit some contexts particularly well but cannot identify the correct top-2 tokens uniformly. Greater depth supplies multiple independently parameterized rules whose errors can be complementary across contexts. Their aggregation improves coverage: for a context that is difficult for one rule, other layers may still identify the relevant tokens, thereby reducing the high-error tail.

The slight advantage of the one-layer model at the lowest percentiles suggests a possible tradeoff between specialization and coverage. A single selection rule may become highly accurate on a limited subset of contexts, whereas complementary rules across layers can cover a broader range. This observation suggests a testable hypothesis for settings with finitely many or clustered context types: different layers may specialize in different context classes, and their aggregation may combine these specialized behaviors. If this hypothesis holds, it would suggest that current architectures use depth inefficiently, as not every layer makes an indispensable contribution. Further exploring designs that exploit such layerwise specialization may help LLMs utilize depth better.

D.3 Broader Targets and Data Spectra

The main-text experiments with the nonlinear-attention model on the difficult task use top-2 aggregation and vary one spectral parameter while fixing the other at zero. We now test whether the observed depth scaling depends on either restriction. First, we retain the top-2 target but consider jointly structured input and task spectra, with both a and b nonzero. Second, we replace top-2 aggregation with top-16 aggregation, yielding a flatter target distribution, and evaluate several spectral settings. For each configuration, we fit the evaluation loss using the model and log-space objective defined in Appendix D.1.

Figure 14 shows approximately inverse-depth loss decay across both extensions, and Table 4 quantifies the corresponding exponents. For the jointly nonzero top-2 settings, panels (c,d) further show that both decomposition terms approach spectrum-dependent finite values, matching the convergence pattern observed in the main text. With jointly nonzero spectra, the top-2 exponents remain close to one. Note that the top-16 exponents deviate relatively more from one at the depths considered, suggesting that the top-16 task may enter the asymptotic inverse-depth regime only at greater depths. This contrast also suggests that inverse-depth scaling from cross-layer aggregation may be more pronounced for



(a) Top-2 aggregation

(b) Top-16 aggregation

(c) Shared error

(d) Error disagreement

Figure 14: Emergent inverse-depth scaling in the deep regime generalizes across broader targets and data spectra. (a) Performance on the top-2 task with both spectral parameters nonzero. (b) Performance on top-16 aggregation across representative spectra. (c) C_L and (d) $A_L - C_L$ converge to finite constants with depth for the three top-2 spectra.

Table 4: Fitted depth-scaling exponents remain close to one across broader targets and data spectra.

Target	(a, b)	L_0	$\alpha \pm \mathrm{SE}$	R^2
Top-2	$(0.25, 0.25)$	0.818068	0.97 ± 0.01	0.9998
	$(0.25, 0.5)$	0.559694	0.93 ± 0.06	0.9984
	$(0.5, 0.5)$	0.382188	0.97 ± 0.01	0.9999
Top-16	$(0, 0.25)$	0.173237	0.79 ± 0.02	0.9999
	$(0, 0.5)$	0.114436	0.88 ± 0.03	0.9998
	$(0.25, 0.25)$	0.115248	0.81 ± 0.02	0.9998
	$(0.25, 0.5)$	0.079785	0.81 ± 0.04	0.9994

sparse, sharply concentrated target distributions. These results show that the inverse-depth scaling behavior is not restricted to specific spectra or target distributions.

D.4 Convergence of the Shared-Error Term

The direct loss fits in Section 4.2 provide empirical evidence for approximate inverse-depth scaling over the tested depth range. As a consistency check on the decomposition in Appendix B, we examine how its shared-error term C_L approaches the fitted loss floor. For each spectrum, we fix L_0 to the corresponding estimate in Table 1 and model the gap as

$$L_0 - C_L = cL^{-\alpha}, \qquad c > 0, \quad \alpha > 0. \tag{92}$$

Following the log-space procedure in Appendix D.1, we estimate the two parameters by least squares:

$$(\widehat{c}, \widehat{\alpha}) = \underset{c>0,\,\alpha>0}{\arg\min} \sum_i \left[\log(L_0 - C_{L_i}) - \log\left(cL_i^{-\alpha}\right)\right]^2. \tag{93}$$

Table 5: The gap $L_0 - C_L$ decays approximately inversely with depth over the fitted range.

(a, b)	$\alpha \pm \mathrm{SE}$	R^2
$(0, 0)$	0.980 ± 0.007	0.99981
$(0, 0.25)$	0.983 ± 0.038	0.99396
$(0, 0.5)$	0.977 ± 0.014	0.99923
$(0, 1)$	0.982 ± 0.011	0.99946
$(0.25, 0)$	0.990 ± 0.015	0.99904
$(0.5, 0)$	0.986 ± 0.005	0.99988
$(1, 0)$	0.994 ± 0.005	0.99992

With L_0 fixed by the direct loss fits, the high R^2 values and small fitting uncertainties show that $L_0 - C_L$ is well described by a power law over the tested depths. Its fitted decay exponents are close to one, consistent with the inverse-depth trend observed in the loss.