Logarithmic Depth for Exact In-Context Gradient Descent with Linear Attention

Yingze Li[1] **Dong Wang**[1] **Xianglong Liu**[1] **Zou Qingyun**[2]
Chunnan Wang[1] **Zhiyu Liang**[1] **Hongzhi Wang**[1*]
[1]Harbin Institute of Technology [2]National University of Singapore

Abstract

How much Transformer depth is needed to compute the result of k gradient-descent updates? Constructions assigning one update to each layer give a linear upper bound. We establish a tight $\Theta(\log k)$ characterization for exact endpoint computation by attention-only, unnormalized linear self-attention with a fixed coordinate workspace. Each layer raises polynomial degree by at most a factor of three, giving a depth lower bound independent of width, heads and context length. A matching construction stores matrix columns in tokens and uses attention sums to implement doubling and increment operations on matrix powers and geometric sums. The construction covers arbitrary positive-definite quadratic objectives and extends to prediction from regression demonstrations at an unseen query. We also prove a barrier to uniform approximation before convergence and a lower bound for prediction risk in a scalar model with a supplied spectrum. Controlled experiments study polynomial approximation and spectral conditioning, while broader regression experiments examine endpoint accuracy across architectures and task families.

1 Introduction

Transformers can adapt to new tasks at inference time, without updating any parameters, simply by conditioning on examples provided in their input. This capability, known as in-context learning (ICL), has been observed across a range of tasks, from linear regression and decision trees to natural language classification [Garg et al., 2022, Akyürek et al., 2023, Brown et al., 2020]. Given a handful of demonstrations of an unseen task, the model infers the task's structure and generalizes to new inputs without any explicit learning algorithm being specified in its architecture [Xie et al., 2022, Akyürek et al., 2023, Li et al., 2023]. This raises a question about computation in ICL. How many Transformer layers are needed to compute a predictor defined by a sequence of optimization updates?

von Oswald et al. [2023] showed how a Transformer can perform gradient descent (GD) on the least-squares objective defined by the in-context examples, assigning one update to each layer. This result became foundational for the theoretical study of ICL [Akyürek et al., 2023, Bai et al., 2023, Li et al., 2023], with extensions to preconditioned gradient descent [Ahn et al., 2023], Newton's method [Fu et al., 2024], and momentum updates [Vladymyrov et al., 2024]. These constructions give an upper bound on the depth of the computation. The final iterate also admits a closed-form expression, raising the possibility of a shallower implementation. *How many layers are necessary and sufficient to compute the same endpoint as a prescribed sequence of GD updates?*

Our result. We show that the minimum depth grows logarithmically with the number of prescribed updates. This characterization holds for exact endpoint computation by attention-only, unnormalized

*Corresponding author: wangzh@hit.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

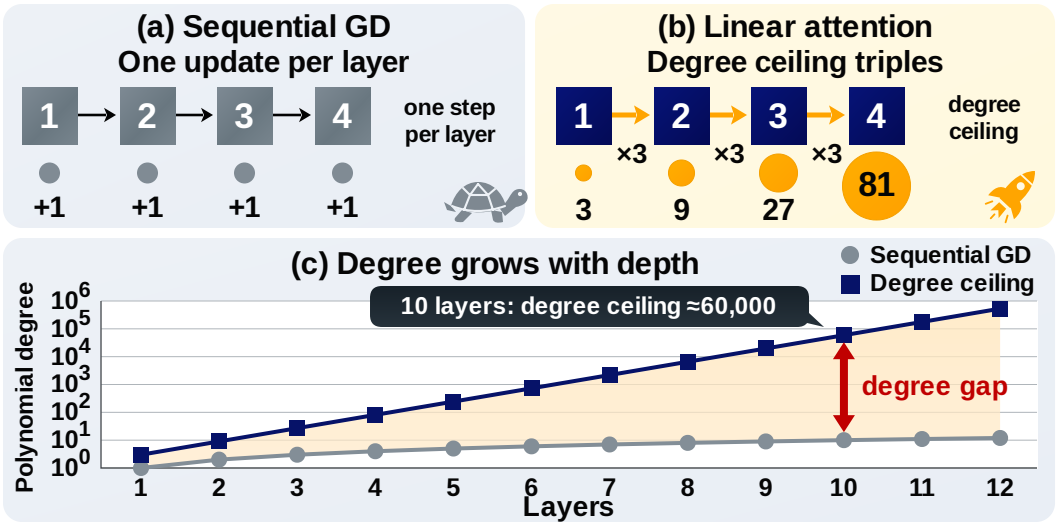

Figure 1: **Depth compounds algebraic capacity.** On the scalar family used for the degree lower bound, sequential GD increases polynomial degree by one per layer. The linear-attention degree ceiling grows exponentially, reaching about 60,000 at ten layers. Figure 2 illustrates how composition gives an exact endpoint construction with logarithmic depth.

linear self-attention with affine input encoding and a fixed coordinate workspace. It covers quadratic objectives with arbitrary positive-definite covariance. The construction also forms the empirical objective from regression demonstrations and produces the corresponding GD prediction at an unseen query. Its workspace size is independent of the update count.

The idea is to compose calculations across layers (Figures 1 and 2). Each layer can reuse and multiply quantities computed by earlier layers, combining progressively larger groups of updates into the final result. Polynomial degree makes the depth requirement precise. The GD endpoint demands increasing degree as the update count grows, while each attention layer can increase degree by only a fixed factor. This gives the necessary depth. An explicit construction stores matrix columns in workspace tokens and uses attention to multiply them, establishing that logarithmic depth is also sufficient.

Contributions. We make three contributions.

❶ **Depth lower bounds from polynomial degree.** We bound the growth of polynomial degree with depth independently of width, head count and context length. This yields a lower bound for exact computation and a bound for uniform approximation in a stated regime before convergence. A separate scalar model gives a depth lower bound for low prediction risk.

❷ **Exact constructions for quadratics and regression prompts.** We give explicit attention projections that compute the GD endpoint for general covariance in logarithmic depth. Forming empirical moments from demonstrations and reading out a query prediction extends the characterization to the stated augmented regression prompt.

❸ **Empirical study of endpoint prediction.** We compare endpoint error across model depths, attention kernels and spectral conditioning, using a polynomial oracle for the distribution of the controlled scalar task. The experiments distinguish approximation accuracy from exact computation and examine how the depth dependence changes across regression settings.

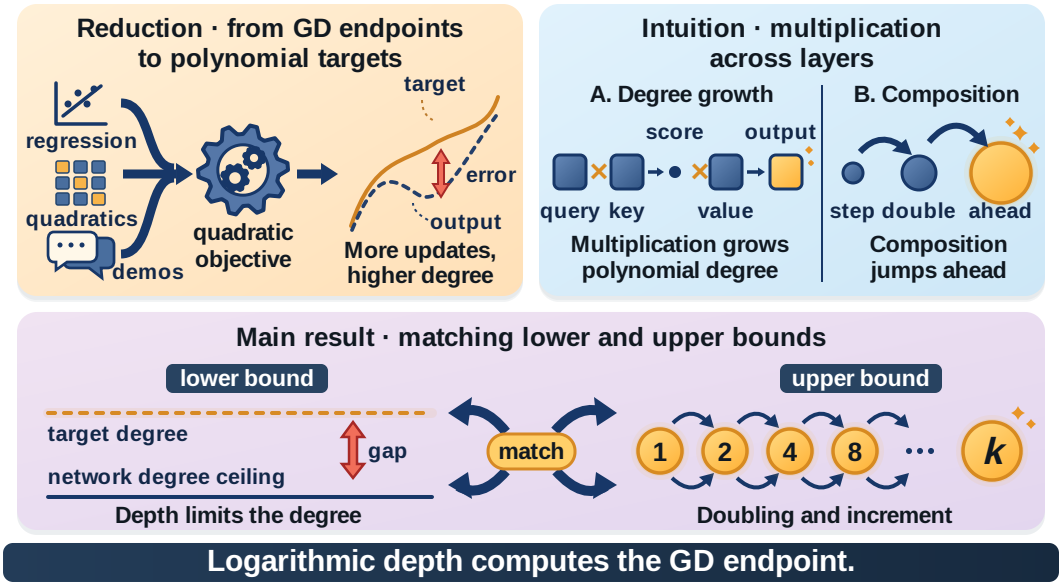

Figure 2: **From regression prompts to logarithmic depth.** The GD endpoint is a polynomial in the inputs. Comparing its degree with the network's degree ceiling gives the lower bound. Matrix composition implements doubling and increment for the upper bound. The two bounds characterize exact computation in the stated linear-attention model with fixed coordinate workspace.

2 Background and Motivation

2.1 Background

ICL as algorithm learning during inference. Transformers can adapt to a new task at inference time by conditioning on examples in the prompt, a behavior observed in large language models and in controlled studies of function classes [Brown et al., 2020, Garg et al., 2022, Akyürek et al., 2023]. An algorithmic account studies how the model constructs a predictor for the task inside its forward pass, using linear regression as a canonical testbed [Akyürek et al., 2023, Li et al., 2023, Bai et al., 2023, von Oswald et al., 2023]. Other accounts explain ICL through implicit Bayesian inference [Xie et al., 2022], the distribution of tasks [Chan et al., 2022, Raventós et al., 2023], training dynamics [Zhang et al., 2024, Song et al., 2024, Makkuva et al., 2024, Ren and Liu, 2024, Kim and Suzuki, 2024], softmax and statistical optimality effects [Li et al., 2024c, Collins et al., 2024, Kim et al., 2024, Oko et al., 2024, Yang et al., 2024, Bu et al., 2024], task vectors [Li et al., 2024a], and connections to memory, RL, diffusion, and Bayesian optimization [Li et al., 2024b, Shi et al., 2024, Wang et al., 2023b, Müller et al., 2023, Ye et al., 2023]. Our question concerns algorithm learning, where the computation performed by depth can be stated exactly [Akyürek et al., 2023, von Oswald et al., 2023, Li et al., 2023].

The shared optimization framework. The optimization view studies a prompt as defining a regression objective, usually least squares or a close variant, and asks which optimizer the Transformer implements in its activations [von Oswald et al., 2023, Akyürek et al., 2023, Li et al., 2023]. von Oswald et al. [2023] showed that a single linear self-attention layer can match one gradient-descent update on a regression loss, while deeper constructions simulate multiple updates [von Oswald et al., 2023, Mahankali et al., 2024]. This same framework underlies independent results on algorithm identification [Akyürek et al., 2023], preconditioned descent [Ahn et al., 2023, Vladymyrov et al., 2024], optimality at depth one [Mahankali et al., 2024], second-order iteration [Fu et al., 2024], broader statistical estimators [Bai et al., 2023, Li et al., 2023], and looped or weight-tied Transformers [Gatmiry et al., 2024, Chen et al., 2025, Giannou et al., 2023]. The spectrum of the regression Hessian controls GD convergence. In its eigenbasis, each coordinate follows a scalar quadratic with difficulty parameter σ, condition number κ, and a k-step convergence polynomial $p_k(\sigma) = 1 - (1 - \eta\sigma)^k$ [von

Oswald et al., 2023, Ahn et al., 2023, Fu et al., 2024]. We study the depth needed to compute a prescribed GD endpoint, with the input encoding and workspace specified explicitly.

Depth as a computational resource. Transformer depth has also been studied through expressivity, circuit complexity, and lower bounds [Yun et al., 2020, Pérez et al., 2019, Sanford et al., 2023, Peng et al., 2024, Kozachinskiy et al., 2025, Sanford et al., 2024, Merrill and Sabharwal, 2023, Wei et al., 2022a, Friedman et al., 2023, Anil et al., 2022, Merrill and Sabharwal, 2024]. Optimization interpretations of ICL use linear or attention-only models related to kernelized attention, fast-weight programmers, modern Hopfield networks, and architectures for efficient attention [von Oswald et al., 2023, Katharopoulos et al., 2020, Choromanski et al., 2021, Schlag et al., 2021, Ramsauer et al., 2021, Kitaev et al., 2020, Zaheer et al., 2020, Dai et al., 2019, Tay et al., 2021, Jaegle et al., 2021, Lee et al., 2019, Dao et al., 2022], while practical ICL is shaped by prompting methods and benchmarks [Wei et al., 2022b, Kojima et al., 2022, Wang et al., 2023a, Zhou et al., 2023, Zhao et al., 2021, Wang et al., 2019b,a, Hendrycks et al., 2021, Biderman et al., 2023, Liu et al., 2022]. The degree argument in Lemma 2 draws on algebraic complexity. Polynomial degree gives arithmetic circuit lower bounds in Strassen [1973], Baur and Strassen [1983], Shpilka and Yehudayoff [2010], and Raz [2013]. We apply this technique to the trilinear attention update. These literatures motivate depth as the resource to measure; the optimization framework above gives a concrete continuous computation on which to measure it.

Parallel composition and logarithmic depth. Sanford et al. [2024] construct causally masked softmax Transformers for k-hop induction by doubling discrete pointers and accumulating the binary expansion of k. Their matching lower bound uses a reduction to Massively Parallel Computation, assumes the one-versus-two-cycles conjecture, and restricts precision and the product of width and head count. For attention-only unnormalized linear attention with inputs of degree at most one, our lower bound follows directly from the trilinear layer map. Each layer raises polynomial degree by at most a factor of three, independently of width, heads, and context length. Our upper construction stores matrix columns in coordinate slots and uses attention sums to compose matrix powers and geometric sums. Both settings connect logarithmic depth to repeated composition, through discrete pointer maps and continuous matrix polynomials.

2.2 The Gap and Our Motivation

The gap. Prior GD constructions implement k updates using k layers, or by repeatedly applying a looped block [von Oswald et al., 2023, Ahn et al., 2023, Fu et al., 2024, Gatmiry et al., 2024, Chen et al., 2025]. This gives an upper bound [von Oswald et al., 2023, Ahn et al., 2023, Fu et al., 2024, Gatmiry et al., 2024, Chen et al., 2025], but it does not establish that k layers are necessary. The open question is whether the final k-step iterate admits a shallower computation that agrees on every task instance.

Our motivation. The attention update is trilinear in the token coordinates [Vaswani et al., 2017, Katharopoulos et al., 2020, von Oswald et al., 2023]. Composing layers can multiply polynomial degree in the task spectrum. For the scalar difficulty parameter σ, L layers can represent degree at most 3^L. This suggests that depth may encode the final result of a long optimization trajectory through algebraic composition. Our bounds establish logarithmic depth for exact endpoint computation in the stated linear-attention model.

3 The Degree-k Target Induced by Gradient Descent

3.1 Diagonal Quadratics as the Demand Side

Task family. We study the task of minimizing diagonal quadratic objectives. Each task instance is specified by a diagonal matrix $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ with $0 < \lambda_{\min} \le \sigma_j \le \lambda_{\max}$ and a linear term $\mathbf{g} \in \mathbb{R}^d$. The objective is

$$f(\mathbf{w}) \;=\; \tfrac{1}{2}\,\mathbf{w}^\top \mathbf{\Sigma}\,\mathbf{w} \;-\; \mathbf{g}^\top \mathbf{w}, \tag{1}$$

which has the unique minimizer $\mathbf{w}^* = \mathbf{\Sigma}^{-1}\mathbf{g}$. Gradient descent from $\mathbf{w}_0 = \mathbf{0}$ with learning rate η produces the iterates

$$\mathbf{w}_k \;=\; \left[\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Sigma})^k\right] \mathbf{w}^*. \tag{2}$$

The diagonal coordinates satisfy $(\mathbf{w}_k)_j = \left[1 - (1 - \eta\sigma_j)^k\right] \cdot g_j/\sigma_j$.

From quadratics to regression. The diagonal family exposes the scalar dependence on the spectrum used in optimization interpretations of ICL [von Oswald et al., 2023, Ahn et al., 2023]. Given demonstrations $(x_i, y_i)_{i=1}^N$, least squares has empirical moments $\widehat{\Sigma} = X^\top X/N$ and $\widehat{g} = X^\top y/N$. Its GD prediction at a new query is $x_q^\top w_k$. We first study inputs that supply (Σ, g) directly, then construct these moments from demonstrations in Theorem 7. The matrix construction applies to arbitrary SPD Σ without computing an eigenbasis.

Network architecture. We consider attention-only linear self-attention Transformers, as used in optimization interpretations of ICL [von Oswald et al., 2023]. The input consists of n tokens in $\mathbb{R}^{d_{\text{model}}}$, including a designated output token q. An L-layer, H-head Transformer updates every token $i \in \{1, \ldots, n\}$ by

$$\mathbf{z}_i^{(\ell+1)} = \mathbf{z}_i^{(\ell)} + \sum_{h=1}^{H} \mathbf{W}_V^{(\ell,h)} \left(\sum_{j=1}^{n} \mathbf{z}_j^{(\ell)} \alpha_{j,i}^{(\ell,h)} \right), \quad \ell = 0, \ldots, L-1, \tag{3}$$

with bilinear scores

$$\alpha_{j,i}^{(\ell,h)} = {\mathbf{z}_j^{(\ell)}}^\top {\mathbf{W}_K^{(\ell,h)}}^\top \mathbf{W}_Q^{(\ell,h)} \mathbf{z}_i^{(\ell)}. \tag{4}$$

Here $\mathbf{W}_Q^{(\ell,h)}, \mathbf{W}_K^{(\ell,h)}, \mathbf{W}_V^{(\ell,h)} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are the query, key, and value weight matrices of head h in layer ℓ. All tokens, including the output token, contribute to each unmasked sum. The model uses residual connections and unnormalized linear attention, with no MLP or LayerNorm. Initial token coordinates are affine in the tracked input variables; a fixed linear readout extracts the output from $\mathbf{z}_q^{(L)}$. The constructions use fixed coordinate workspace tokens, whose contents and count are specified with each theorem. Network parameters may depend on k, the fixed input dimensions and the step size, but are shared across task instances.

3.2 A Convergence Polynomial of Degree k

For inputs supplying (Σ, g), exact computation means outputting w_k for every instance using the same network weights. For demonstration inputs, the required output is $x_q^\top w_k$ for the empirical objective. We fix a positive step size η; the quantitative bound for uniform approximation uses $\eta = 1/\lambda_{\max}$.

The diagonal case gives a scalar explanation of the degree requirement. The GD iterate depends on σ through the *convergence polynomial*

$$p_k(\sigma) \;=\; 1 - (1 - \eta\sigma)^k, \tag{5}$$

a polynomial of degree exactly k in σ. On the scalar subfamily $g = \sigma$, the minimizer is one and $w_k = p_k(\sigma)$, so the inputs remain affine in σ. This subfamily supplies the scalar exact lower bound. For independent variables (Σ, g), the endpoint $w_k = \eta \sum_{t=0}^{k-1} (I - \eta\Sigma)^t g$ has total degree k; with g fixed, its degree in Σ is $k - 1$. The degree count thus depends on the specified input variables.

Problem 1 (Depth complexity of in-context gradient descent)**.** For fixed input dimensions and the stated affine prompt encodings, determine the minimum depth $L^*(k)$ that computes the prescribed GD endpoint exactly on the task family. Separately, determine the depth required to approximate p_k uniformly to error ε.

Polynomial approximation complexity. Write $\Lambda := [\lambda_{\min}, \lambda_{\max}]$, $\kappa = \lambda_{\max}/\lambda_{\min}$, and let Π_D denote real polynomials of degree at most D. For fixed η and Λ, the best degree-D uniform approximation error of f and the degree needed to reach error ε are

$$E_D(f; \Lambda) := \inf_{q \in \Pi_D} \|f - q\|_{L^\infty(\Lambda)}, \qquad d_\varepsilon(k, \kappa) := \min\{\, D \in \mathbb{N}_0 : E_D(p_k; \Lambda) \le \varepsilon \,\}. \tag{6}$$

The notation suppresses the dependence on the fixed step size and spectral scale; our quantitative approximation bound sets $\eta = 1/\lambda_{\max}$. The depth lower bounds in Section 4.3.1 compare the degree supplied by L layers with the degree required to approximate p_k at accuracy ε.

4 The Exponential Depth Effect of Linear Attention

4.1 Degree Growth Mechanism

A trilinear update. Consider polynomial degree in the scalar spectrum parameter σ, or total degree in the prompt variables. The attention score (4) multiplies key by query, so if the previous layer carries degree D polynomials in σ, the score can carry degree $2D$. Multiplying that score by the value vector adds one more degree-D factor, while heads, residuals, and linear maps only add or mix coordinates. One layer therefore turns degree D into degree at most $3D$.

Lemma 2 (Degree Growth Lemma)**.** *With affine input encoding, every token coordinate and fixed linear readout of an L-layer linear self-attention Transformer has polynomial degree at most 3^L in the tracked variables.*

The proof takes the maximum degree across all tokens and applies $D_{\ell+1} \le 3D_\ell$ to the full attention sum (Appendix A.1). Width, heads, and context length enter through sums and fixed linear maps, so they do not change this ceiling.

4.2 Depth Required for Exact Computation

Comparing target and output degrees. The Degree Growth Lemma gives $P_L \in \Pi_{3^L}$. The target $p_k(\sigma) = 1 - (1 - \eta\sigma)^k$ has degree exactly k on Λ. If P_L matches p_k pointwise on Λ, the two polynomials must agree as elements of Π_k (since Λ contains infinitely many points). But $P_L \in \Pi_{3^L}$, forcing $3^L \ge k$.

Theorem 3 (Exact computation lower bound)**.** *Let $P_L(\sigma)$ be the output polynomial at the query of any L-layer linear self-attention Transformer defined by* (3) *and* (4)*. If Λ is a nondegenerate interval and $P_L(\sigma) = p_k(\sigma)$ for every $\sigma \in \Lambda$, then $3^L \ge k$, hence $L \ge \lceil \log_3 k \rceil$. The bound is independent of the number of heads, model dimension, and context length.*

Proof. Lemma 2 gives $P_L \in \Pi_{3^L}$. The polynomial $p_k - P_L$ vanishes on the infinite set Λ, hence is identically zero, so $k = \deg p_k \le \deg P_L \le 3^L$. □

4.3 Approximation and Exact Constructions

4.3.1 Uniform approximation

Theorem 4 (Approximation lower bound)**.** *If a linear-attention output approximates p_k on Λ to absolute uniform error ε, then $3^L \ge d_\varepsilon(k, \kappa)$. For $\eta = 1/\lambda_{\max}$, $\kappa \ge 2$, $k \le \kappa$, and $0 < \varepsilon \le c_0/(e\sqrt{k})$, where $c_0 = (2e^{3/2}\sqrt{\pi})^{-1}$, this gives $d_\varepsilon \ge \sqrt{k}/2$ and $L \ge \frac{1}{2}\log_3 k - \log_3 2$.*

The general implication follows from $P_L \in \Pi_{3^L}$; the quantitative estimate uses the Chebyshev coefficients of p_k (Appendix A.4). With $\eta = 1/\lambda_{\max}$ and fixed κ and ε, the constant polynomial one suffices once $k \ge \kappa \log(1/\varepsilon)$.

4.3.2 Constructing the endpoint in logarithmic depth

The upper bound uses matrix multiplication inside the attention sum. Let $A = I - \eta\Sigma$, $P_m = A^m$, and $S_m = \sum_{t=0}^{m-1} A^t$, so that $w_m = \eta S_m g$. Matrix columns stored in d workspace tokens give $XY = \sum_j (Xe_j)(e_j^\top Y)$. Two heads can therefore implement each of the updates

$$(P_m, S_m) \longmapsto (P_m^2, S_m + P_m S_m), \qquad (P_m, S_m) \longmapsto (AP_m, S_m + P_m), \tag{7}$$

which advance m to $2m$ and $m+1$. Reading the binary digits of k gives $\ell_{\mathrm{bin}}(k) = \lfloor \log_2 k \rfloor + \nu(k) - 1$ such layers, where $\nu(k)$ counts the one bits.

Theorem 5 (Exact construction for general covariance)**.** *For fixed $d, k \ge 1$ and $\eta > 0$, an affine encoding of (Σ, g) into d workspace tokens and one output token admits exact computation of w_k for every SPD Σ and every g with*

$$L = \ell_{\mathrm{bin}}(k) + 1 \le 2\lceil \log_2 k \rceil + 1, \qquad H = 2, \qquad d_{\mathrm{model}} = 6d.$$

Every W_Q, W_K, W_V has entrywise maximum and induced infinity norm at most $\max(1, \eta)$. For $k = 2^p$, the depth is $p + 1$.

Table 1: Scope of the theoretical results. All exact statements use affine input encodings and a fixed linear readout.

Input or architecture	Result	Conditions or reference
Given (Σ, g)	Exact GD endpoint, $\Theta(\log k)$ depth	Full linear attention; Corollary 6
Demonstrations and query	Exact empirical GD prediction, $\Theta(\log k)$	Fixed coordinate workspace; Theorem 7
Scalar spectral factor	Uniform error requires $3^L \geq d_\varepsilon$	Quantitative regime in Theorem 4
Supplied scalar (σ, g)	Lower bound for prediction risk	Distribution and accuracy conditions in Theorem 19
Linear attention with MLP	Degree-p activation ceiling $(3p)^L$; ReLU exact lower bound	Corollary 22
Softmax or smooth nonpolynomial blocks	Empirical comparisons	Outside the polynomial ceiling

Proof sketch. Workspace j stores $(e_j, P_m e_j, Ae_j, S_m e_j)$ in separate blocks; the output token stores g. For doubling, one head sums $P_m e_j (e_j^\top (P_m - I) e_i)$, so the residual leaves $P_m^2 e_i$. A second head adds $P_m S_m e_i$ to $S_m e_i$. Increment similarly adds $AP_m e_i$ while a second head moves $P_m e_i$ from the power block into the sum block. The final layer reads out $\eta \sum_j S_k e_j g_j$. Invariants for every token and all projection matrices are given in Appendix B. □

Corollary 6 (Exact depth complexity)**.** *For fixed input dimension, exact computation of the prescribed GD endpoint on the full SPD quadratic family has depth complexity $\Theta(\log k)$ under the stated affine encoding and linear-attention architecture.*

The lower bound follows from the degree-k restriction in Section 3.2; the construction supplies the matching upper bound.

4.3.3 From demonstrations to a query prediction

The empirical moments needed by the construction can themselves be formed by attention. Let X have rows $x_i^\top$, set $\widehat{\Sigma} = X^\top X/N$ and $\widehat{g} = X^\top y/N$, and let w_k be GD from zero on $\|Xw - y\|^2/(2N)$.

Theorem 7 (Exact prediction from demonstrations)**.** *Fix $N, d, k \geq 1$ and $\eta > 0$. A prompt with N demonstrations (x_i, y_i), d fixed coordinate workspace tokens, and an unseen query x_q admits exact output $x_q^\top w_k$ with $L = \ell_{\mathrm{bin}}(k) + 3$, $H = 3$, and $d_{\mathrm{model}} = 8d + 3$. Initial coordinates are affine in (X, y, x_q). Conversely, any such linear-attention model that is exact on a set of prompts with nonempty interior satisfies $3^L \geq 2k + 1$.*

Proof sketch. One layer forms the columns of $\widehat{\Sigma}$ in the workspace and $\widehat{g}$ in the query. A separate input block for x_q keeps the query out of these empirical moments. The binary chain and vector readout compute w_k, and a final head forms $x_q^\top w_k$ through the query's attention to itself. The target has degree exactly $2k + 1$ in the raw prompt variables, whereas the network has degree at most 3^L. Appendices B.4 and B.5 give the complete construction and the nonzero leading coefficient. □

Prediction risk and architectural scope. A separate result concerns the optimal scalar predictor when the spectrum and linear term are supplied. For independent Gaussian g, the component of an arbitrary polynomial estimator that is linear in g must approximate $1/\sigma$. Under the distribution and accuracy conditions of Theorem 19, excess prediction risk at most ε requires $L = \Omega(\log \kappa + \log\log(1/\varepsilon))$. Table 1 collects the input models and architectural scope; Appendices D.2 and D.3 give consequences for MLP blocks and weight sharing.

Table 2: **Reported critical depths across attention kernels.** The threshold is median finite evaluation NMSE $< 10^{-5}$.

Kernel	$k = 8$	16	32	64	128
Linear	3	3	3	6	8
Softmax	2	2	3	4	6
Unnormalized exponential	2	2	2	2	2
ReLU	2	2	2	3	3

5 Experiments

5.1 Experimental Setup and Metrics

Task and metric. The polynomial task provides three demonstrations $(\sigma_j, ap_k(\sigma_j))$ at fixed Chebyshev roots on $[1/\kappa, 1]$ and a query $(\sigma_q, 0)$. The target is $ap_k(\sigma_q)$, where $p_k(\sigma) = 1 - (1 - 0.5\sigma)^k$. In the original sweep, $a \sim \mathcal{N}(0, 1)$ and σ_q is uniform on this interval, with $\kappa = 1000$. We measure finite evaluation error by

$$\widehat{\mathrm{NMSE}} = \frac{\sum_i (\widehat{y}_i - y_i)^2}{\sum_i y_i^2}.$$

The empirical critical depth $L_{\mathrm{crit}}^{\mathrm{reg}}$ is the smallest tested depth whose median NMSE across seeds is below the stated threshold, usually 10^{-5}. This statistic describes prediction at finite error; the theorems for exact computation and uniform approximation use their respective equality and worst-case criteria.

Model and sweep. The original sweep uses attention-only linear self-attention with $H = 4$, $d_{\mathrm{model}} = 64$, and no softmax, MLP, or LayerNorm. Its 620 runs comprise 11 values $k \in \{4, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128\}$, 11 depths $L \in \{1, 2, 3, 4, 5, 6, 8, 10, 12, 16, 20\}$, and five seeds, plus 15 runs at $L = 7$ for $k \in \{64, 96, 128\}$. Adam uses 8,000 scheduled steps for $L \leq 6$, 12,000 for $7 \leq L \leq 12$, and 15,000 for larger depths. The historical sweep uses float32 model computation and newly sampled evaluation queries; the separate control below uses float64 throughout and fixed independent evaluation sets. Appendices E and G give both protocols.

5.2 Depth, Kernels, and Approximation Degree

Trained endpoint performance. The reported crossings for linear attention in Table 2 range from three layers at $k = 8$ to eight at $k = 128$. Changing the attention kernel changes this empirical profile. Softmax has crossings from two to six layers, while unnormalized exponential and ReLU kernels remain at two or three on this grid. These observations compare the endpoint predictors learned under each architecture.

A control with a matched population metric. We trained 1,200 attention-only models without LayerNorm in float64, using $k \in \{4, 8, 16, 32, 64, 128\}$, $\kappa \in \{100, 1000\}$, depths 1 through 10, and ten seeds. Here $a \sim \mathrm{Unif}[-1, 1]$, and each (k, κ) has a shared, independent evaluation set of 65,536 queries. For the uniform query distribution, we also compute the least polynomial degree $d_\varepsilon^{L^2}$ achieving population NMSE at most ε. The oracle degrees agree at 4,096 and 8,192 quadrature nodes in all 48 tested (k, κ, ε) cells. At $\varepsilon = 10^{-12}$ and $k = 128$, they are 26 and 27 for the two condition numbers, illustrating how a finite accuracy target can require much less degree than exact equality. Of the 1,200 training runs, 1,131 completed with finite error and 69 stopped at nonfinite loss; the frozen rule requiring complete depth grids leaves the control's critical depths undefined. Appendix G reports the oracle degrees and training outcomes.

5.3 Regression and Probes of Pretrained Models

Broader in-context regression. The regression prompts contain 64 demonstrations (x_i, y_i) and a query $(x_q, 0)$, with $x \sim \mathcal{N}(0, \Sigma)$ and $y = x^\top w^\star$ in dimension 16. The model predicts the query label; training and evaluation use independently sampled tasks. At an NMSE threshold of 10^{-2}, the

Table 3: **Exploratory fits against class count across all four pretrained models.** The fit is $L_{\text{crit}}^{\text{probe}} = a + b \log_2 C$ on the selected tasks.

Model	Layers	Tasks	b	a	R^2
Qwen2.5 3B	36	6/8	3.42	20.59	0.91
Qwen2.5 1.5B	28	5/8	2.50	14.43	0.75
Pythia 2.8B	32	5/8	3.24	16.76	0.20
Pythia 6.9B	32	6/8	-0.52	32.70	0.54

reported crossings for diagonal and rotated covariances are $3, 4, 5$ for $\kappa = 10, 100, 1000$. Appendix I specifies the five covariance families, task prior, training schedule, and both thresholds. Endpoint error measures the final predictor and leaves the internal optimization trajectory unidentified.

Condition number at fixed depth. A float64 control uses $d \in \{4, 16\}$, $N = 2d$ demonstrations, and diagonal covariances or a fixed rotation of them with $\kappa \in \{10, 30, 100, 300, 1000, 3000\}$. Its task prior, protocol with ten seeds, and training outcomes appear in Appendix G.2. At $L = 6$ and $\kappa = 10$, all four median NMSE values lie between 1.883×10^{-4} and 3.031×10^{-4}. At $\kappa = 3000$, three complete cells have NMSE approximately one, and the fourth has an incomplete seed set. The control shows a conditioning effect at this depth; numerical instability at larger depths leaves the preregistered fits of critical depth unestimable.

Exploratory layer probes. We evaluate Qwen2.5 and Pythia models [Qwen Team, 2024, Biderman et al., 2023] using one demonstration per class and 200 query examples per classification task. A logit lens applies the final normalization and language model head to each layer. We define $L_{\text{crit}}^{\text{probe}}$ as the earliest layer reaching 95% of accuracy at the final layer and use C for class count. Table 3 reports all four fits from the original evaluation together. The Qwen2.5 3B fit uses six tasks with five distinct C values and has an intercept of 20.59 in a 36-layer model. The weaker Pythia 2.8B fit and negative Pythia 6.9B slope show that the association varies across models. These are exploratory correlations across tasks; C is distinct from the prescribed GD update count k. Appendix F gives the task inclusion and measurement protocol.

Limitations and Scope

The exact characterization applies to unnormalized linear attention with affine input encoding and a fixed coordinate workspace. The construction uses width proportional to the task dimension and weights that depend on the prescribed update count. Transferring these guarantees to causal prompts or standard softmax Transformers requires further analysis. The growing bound for uniform approximation depends on conditioning, accuracy and the horizon before convergence. The risk result uses a supplied scalar spectrum with stated distributional assumptions. Measured endpoint loss describes prediction accuracy, and the probes of pretrained models give exploratory associations across tasks. Identifying the computations learned by training remains a separate problem.

6 Conclusion

Exact GD endpoint computation has logarithmic depth complexity in the stated linear-attention model. The lower bound follows from polynomial degree, and the upper bound follows from matrix multiplication and binary composition inside attention. The same construction forms empirical moments from demonstrations and predicts at an unseen query. These results show how depth can compose the endpoint of many optimization updates, with explicit input, workspace and architectural requirements.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62232005 and 62202126. The authors declare no competing interests.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8ed3d610ea4b68e7afb30ea7d01422c6-Abstract-Conference.html.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay V. Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/fb7451e43f9c1c35b774bcfad7a5714b-Abstract-Conference.html.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/b2e63e36c57e153b9015fece2352a9f9-Abstract-Conference.html.

Walter Baur and Volker Strassen. The complexity of partial derivatives. *Theoretical Computer Science*, 22(3):317–330, 1983.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Taiji Suzuki, Qingfu Zhang, and Hau-San Wong. Provably transformers harness multi-concept word semantics for efficient in-context learning. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/73efab19ebde03ff0958f4f155483f57-Abstract-Conference.html.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html.

Bo Chen, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Bypassing the exponential dependency: Looped transformers efficiently learn in-context by multi-step gradient descent. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4447–4455, 2025. arXiv:2410.11268.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function lipschitzness. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a8633d27d782f66fe660c2fb4bae446e-Abstract-Conference.html.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1285. URL https://doi.org/10.18653/v1/p19-1285.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Dan Friedman, Alexander Wettig, and Danqi Chen. Learning transformer programs. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/995f693b73050f90977ed2828202645c-Abstract-Conference.html.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b2d4051f03a7038a2771dfbbe5c7b54e-Abstract-Conference.html.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 15130–15152. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/gatmiry24b.html.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 11398–11442. PMLR, 2023. URL https://proceedings.mlr.press/v202/giannou23a.html.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 4651–4664. PMLR, 2021. URL http://proceedings.mlr.press/v139/jaegle21a.html.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pages 5156–5165. PMLR, 2020. URL http://proceedings.mlr.press/v119/katharopoulos20a.html.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 24527–24561. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/kim24af.html.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c11daad0a48ea5f3c5c6390c7b060720-Abstract-Conference.html.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Alexander Kozachinskiy, Felipe Urrutia, Hector Ignacio Jimenez Orellana, Tomasz Steifer, Germán Pizarro, Matías Daniel Fuentes Álvarez, Francisco Meza Vásquez, Cristian Buc Calderon, and Cristobal Rojas. Strassen attention, split VC dimension and compositionality in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2501.19215.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pages 3744–3753. PMLR, 2019. URL http://proceedings.mlr.press/v97/lee19d.html.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6 (2):167–195, 2015.

Dongfang Li, Zhenyu Liu, Xinshuo Hu, Zetian Sun, Baotian Hu, and Min Zhang. In-context learning state vector with inner and momentum optimization. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/0ed52d7f6f641f228405d48a611e0684-Abstract-Conference.html.

Ji-An Li, Corey Yishan Zhou, Marcus K. Benna, and Marcelo G. Mattar. Linking in-context learning in transformers to human episodic memory. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/0ba385c3ea3bb417ac6d6a33e24411bc-Abstract-Conference.html.

Shuai Li, Zhao Song, Yu Xia, Tong Yu, and Tianyi Zhou. The closeness of in-context learning and weight shifting for softmax regression. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024c. URL http://papers.nips.cc/paper_files/paper/2024/hash/726ab29b61a749b36d2593648716ae3c-Abstract-Conference.html.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 19565–19594. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23l.html.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/0cde695b83bd186c1fd456302888454c-Abstract-Conference.html.

Arvind Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations (ICLR)*, 2024.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Hyeji Kim, Michael Gastpar, and Chanakya Ekbote. Local to global: Learning dynamics and effect of initialization for transformers. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9cdb4f8c4dfa13284d2d5a6e7853e5a2-Abstract-Conference.html.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics (TACL)*, 2023.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NjNGlPh8Wh.

Samuel Müller, Matthias Feurer, Noah Hollmann, and Frank Hutter. Pfns4bo: In-context learning for bayesian optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 25444–25470. PMLR, 2023. URL https://proceedings.mlr.press/v202/muller23a.html.

National Institute of Standards and Technology. NIST Digital Library of Mathematical Functions. https://dlmf.nist.gov/. Chapter 18, Table 18.3.1 and Eq. 18.14.1. Accessed 18 October 2026.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8d6c356c82252032a20c749dda36f2d8-Abstract-Conference.html.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 115–124, 2005.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture. In *Conference on Language Modeling (COLM)*, 2024.

Jorge Pérez, Javier Marinkovic, and Pablo Barceló. On the turing completeness of modern neural network architectures. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HyGBdo0qFm.

Qwen Team. Qwen2.5 technical report, 2024. arXiv:2412.15115.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David P. Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors,

Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/2e10b2c2e1aa4f8083c37dfe269873f8-Abstract-Conference.html.

Ran Raz. Tensor-rank and lower bounds for arithmetic formulas. *Journal of the ACM*, 60(6): 40:1–40:15, 2013.

Ruifeng Ren and Yong Liu. Towards understanding how transformers learn in-context through a representation learning lens. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/01a8d63f9cb6dcbaa3092ccddd2075ac-Abstract-Conference.html.

Clayton Sanford, Daniel J. Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/73bf692447f174984f30499ec9b20e04-Abstract-Conference.html.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 43276–43327. PMLR, 2024. arXiv:2402.09268.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 9355–9366. PMLR, 2021. URL http://proceedings.mlr.press/v139/schlag21a.html.

Chengshuai Shi, Kun Yang, Jing Yang, and Cong Shen. Transformers as game players: Provable in-context game-playing capabilities of pre-trained models. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ee5884bc4908108d672e1c8e940c6c7b-Abstract-Conference.html.

Amir Shpilka and Amir Yehudayoff. Arithmetic circuits: A survey of recent results and open questions. *Foundations and Trends in Theoretical Computer Science*, 5(3–4):207–388, 2010.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

Bingqing Song, Boran Han, Shuai Zhang, Jie Ding, and Mingyi Hong. Unraveling the gradient descent dynamics of transformers. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a7d36e5cb41a1f21c46db25cb1aafab9-Abstract-Conference.html.

Volker Strassen. Vermeidung von divisionen. *Journal für die reine und angewandte Mathematik*, 264:184–202, 1973.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 10183–10192. PMLR, 2021. URL http://proceedings.mlr.press/v139/tay21a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Max Vladymyrov, Johannes von Oswald, Mark Sandler, and Rong Ge. Linear transformers are versatile in-context learners. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/57a3c602f0a1c8980cc5ed07e49d9490-Abstract-Conference.html.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 35151–35174. PMLR, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang (Atlas) Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b3750390ca8b931fb9ca988647940cb-Abstract-Conference.html.

Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a. URL http://papers.nips.cc/paper_files/paper/2022/hash/4ebf1d74f53ece08512a23309d58df89-Abstract-Conference.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022b. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1112–1122, 2018.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9bfa0c155653e24120760a5ead819376-Abstract-Conference.html.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 39818–39833. PMLR, 2023. URL https://proceedings.mlr.press/v202/ye23c.html.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

Ruiqi Zhang, Jingfeng Wu, and Peter L. Bartlett. In-context learning of a linear transformer block: Benefits of the MLP component and one-step GD initialization. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/20b6b87ca17792337f414d948af7b0e8-Abstract-Conference.html.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, pages 649–657, 2015.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 12697–12706. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhao21c.html.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=92gvk82DE-.

A Degree Growth and Lower Bounds

A.1 Degree growth for all tokens

Lemma 8 (Multivariate degree growth)**.** *Suppose every initial token coordinate is a polynomial of total degree at most one in a parameter vector ξ. Under the full attention recurrence* (3)*, every token coordinate after L layers, and every fixed linear readout, has total degree at most 3^L in ξ.*

Proof. Let D_ℓ be the maximum total degree of any coordinate of any token at layer ℓ. Each score $z_j^\top W_K^\top W_Q z_i$ has degree at most $2D_\ell$, and its product with $W_V z_j$ has degree at most $3D_\ell$. The sums over all keys $j \in [n]$ and all heads, followed by residual addition, have degree at most $3D_\ell$. This includes the term $j = i$ and every term involving the query as a key, giving

$$D_{\ell+1} \le 3D_\ell, \qquad D_0 \le 1,$$

which proves $D_L \le 3^L$. Fixed linear readout cannot increase degree. The number of summands and coordinates does not enter the recurrence. □

The univariate statement in Lemma 2 follows by taking $\xi = \sigma$. Fixed positional or workspace coordinates have degree zero. The same argument permits any fixed affine input embedding.

A.2 Exact computation of the spectral factor

Proof of Theorem 3. The preceding lemma gives $P_L \in \Pi_{3^L}$. For fixed $\eta \neq 0$,

$$p_k(\sigma) = \sum_{j=1}^{k} (-1)^{j+1} \binom{k}{j} \eta^j \sigma^j$$

has degree exactly k. If $P_L = p_k$ on a nondegenerate interval, their difference has infinitely many roots and is identically zero. Comparing degrees gives $k \le 3^L$. □

A.3 Uniform approximation

Proof of the degree implication in Theorem 4. If $\|P_L - p_k\|_{L^\infty(\Lambda)} \le \varepsilon$, then

$$E_{3^L}(p_k; \Lambda) \le \|P_L - p_k\|_{L^\infty(\Lambda)} \le \varepsilon.$$

By the definition of d_ε over nonnegative integer degrees, $3^L \ge d_\varepsilon(k, \kappa)$. When $d_\varepsilon \ge 1$, taking logarithms yields $L \ge \lceil \log_3 d_\varepsilon \rceil$. □

A.4 The pre-convergence estimate

Theorem 9 (Uniform approximation of the convergence polynomial)**.** *Fix $\eta = 1/\lambda_{\max}$ and $\kappa \ge 2$. For integers $k \ge 1$ and $0 \le D < k$,*

$$E_D(p_k; \Lambda) \ge (1 - 1/\kappa)^k 4^{-k} \binom{2k}{k - D - 1}. \tag{8}$$

If also $k \le \kappa$ and $D + 1 \le \sqrt{k}$, then

$$E_D(p_k; \Lambda) > \frac{c_0}{\sqrt{k}} \exp\left(-\frac{(D+1)^2}{k}\right), \qquad c_0 = \frac{1}{2e^{3/2}\sqrt{\pi}}. \tag{9}$$

Consequently $d_\varepsilon(k, \kappa) \ge \sqrt{k}/2$ whenever $0 < \varepsilon \le c_0/(e\sqrt{k})$ and $k \le \kappa$.

Proof. Set $y = (\lambda_{\max} - \sigma)/(\lambda_{\max} - \lambda_{\min})$ and $\beta = (1 - 1/\kappa)^k$. The affine change of variable gives

$$p_k(\sigma) = 1 - \beta y^k, \qquad E_D(p_k; \Lambda) = \beta E_D(y^k; [0, 1]). \tag{10}$$

Under $x = 2y - 1$, the Chebyshev expansion is

$$y^k = \sum_{j=0}^{k} (2 - \delta_{j0}) 4^{-k} \binom{2k}{k - j} T_j(x). \tag{11}$$

For completeness, set $x = \cos\theta$, write $y^k = \cos^{2k}(\theta/2)$, and expand $(e^{i\theta/2} + e^{-i\theta/2})^{2k}/4^k$. Pairing conjugate terms gives the displayed coefficients.

For $j \geq 1$, the jth Chebyshev coefficient of a continuous function f satisfies

$$|c_j(f)| = \left| \frac{2}{\pi} \int_0^\pi f(\cos\theta)\cos(j\theta)\, d\theta \right| \leq 2\|f\|_\infty. \tag{12}$$

Every degree-D polynomial has zero coefficient at $j > D$. Therefore

$$E_D(f; [-1,1]) \geq \tfrac{1}{2}|c_j(f)| \qquad (j > D). \tag{13}$$

Using $j = D + 1$ in (11) gives

$$E_D(y^k; [0,1]) \geq 4^{-k}\binom{2k}{k-D-1}, \tag{14}$$

and (8) follows from (10).

To obtain the quantitative estimate, let $j = D + 1 \leq \sqrt{k}$, $K = k + \frac{1}{2}$, and $x_i = i - \frac{1}{2}$. The binomial ratio satisfies

$$R_{k,j} := \frac{\binom{2k}{k-j}}{\binom{2k}{k}} = \prod_{i=1}^{j} \frac{K - x_i}{K + x_i}.$$

With $r = j^2/K^2 < 1$, use $\sum_i x_i = j^2/2$ and $\sum_i x_i^{2m+1} \leq j^{2m+2}/2$ in the convergent $\operatorname{arctanh}$ series to obtain

$$\begin{aligned} -\log R_{k,j} &= 2\sum_{i=1}^{j} \operatorname{arctanh}(x_i/K) \\ &\leq \frac{j^2}{K}\left(1 + \frac{r}{3(1-r)}\right) \leq \frac{j^2}{K}\left(1 + \frac{1}{2k}\right) = \frac{j^2}{k}. \end{aligned}$$

The penultimate step follows because

$$\frac{r}{3(1-r)} \leq \frac{k}{3(K^2 - k)} = \frac{k}{3(k^2 + 1/4)} \leq \frac{1}{2k}.$$

Wallis' inequality gives

$$4^{-k}\binom{2k}{k} \geq \frac{1}{\sqrt{\pi(k + 1/2)}} \geq \frac{1}{e^{1/2}\sqrt{\pi k}}.$$

Combining these inequalities,

$$4^{-k}\binom{2k}{k-j} \geq \frac{1}{e^{1/2}\sqrt{\pi k}} e^{-j^2/k}. \tag{15}$$

For $k \leq \kappa$ and $\kappa \geq 2$, $\beta \geq (1 - 1/\kappa)^\kappa \geq 1/4 > 1/(2e)$. Here $(1 - 1/\kappa)^\kappa$ is increasing for $\kappa > 1$, since the derivative of its logarithm is $\log(1 - 1/\kappa) + 1/(\kappa - 1) > 0$. This proves the strict inequality (9).

For every integer $D < \sqrt{k}/2$, we have $D + 1 \leq \sqrt{k}$. For $k \geq 4$ this follows from $D + 1 < \sqrt{k}/2 + 1 \leq \sqrt{k}$; for $k \in \{1, 2, 3\}$ the only possible degree is $D = 0$. Equation (9) consequently gives $E_D(p_k; \Lambda) > c_0/(e\sqrt{k})$ for each such D. All degrees below $\sqrt{k}/2$ are excluded even when $\varepsilon = c_0/(e\sqrt{k})$, proving the last assertion. □

Corollary 10 (Uniform error at a prescribed depth)**.** *Every model satisfying Lemma 2 obeys* $\|P_L - p_k\|_{L^\infty(\Lambda)} \geq E_{3^L}(p_k; \Lambda)$. *Under the preceding theorem's hypotheses, if* $3^L + 1 \leq \sqrt{k}$, *this error exceeds* $c_0/(e\sqrt{k})$.

Proof. The first assertion follows from $P_L \in \Pi_{3^L}$; the second uses (9) with $D = 3^L$. □

Accuracy after convergence. For fixed κ and $\varepsilon \in (0,1)$,

$$\|p_k - 1\|_{L^\infty(\Lambda)} = (1 - 1/\kappa)^k \le e^{-k/\kappa}.$$

Thus degree zero suffices once $k \ge \kappa \log(1/\varepsilon)$. An accuracy requirement that shrinks with k can retain a growing bound. For $\kappa = 2$, $\lambda_{\max} = \eta = 1$, and $\Lambda = [1/2, 1]$, the leading coefficient of p_k has magnitude one. The monic Chebyshev extremal property on an interval of length $1/2$ gives $E_{k-1}(p_k; \Lambda) = 2 \cdot 8^{-k}$. Hence $\varepsilon = 8^{-k}$ requires degree k for every k.

A.5 Exact computation for general covariance

Theorem 11 (General covariance lower bound)**.** *Fix $\eta \neq 0$ and $w^\star \neq 0$. Let $\mathcal{U}$ be a nonempty open subset of the vector space of symmetric matrices, consisting of positive definite matrices. Suppose initial token coordinates are affine in (Σ, g), and a model outputs the exact GD iterate for every $\Sigma \in \mathcal{U}$ under $g = \Sigma w^\star$. Then $3^L \ge k$.*

Proof. On this restriction the input remains affine in the independent entries of Σ, and the target is $[I - (I - \eta\Sigma)^k]w^\star$. Its highest homogeneous part is $-(-\eta)^k \Sigma^k w^\star$. Choose a coordinate r with $w^\star_r \neq 0$ and restrict the polynomial to $\Sigma = tI$ to see the nonzero term $-(-\eta)^k t^k w^\star_r$. This restriction establishes that the degree-k component is nonzero; it need not lie inside $\mathcal{U}$. Exact equality on $\mathcal{U}$ is a polynomial identity in the independent matrix entries. Lemma 8 therefore gives $k \le 3^L$. □

For the full family with independently varying (Σ, g), the same conclusion follows directly. Its vector target $\eta \sum_{t<k} (I - \eta\Sigma)^t g$ has highest joint homogeneous part $\eta(-\eta)^{k-1}\Sigma^{k-1} g$ and total degree k. With g held fixed, its degree in Σ is $k - 1$ when $g \neq 0$. The affine restriction $g = \Sigma w^\star$ explains how the spectral factor of degree k appears in the exact lower bound.

B Exact Construction by Matrix Columns

B.1 State and input encoding

Fix $d, k \ge 1$ and $\eta > 0$. Write

$$A = I - \eta\Sigma, \qquad P_m = A^m, \qquad S_m = \sum_{t=0}^{m-1} A^t, \qquad w_m = \eta S_m g. \tag{16}$$

The polynomial expression for w_m is valid for every matrix Σ. For invertible Σ, it also equals $[I - (I - \eta\Sigma)^m]\Sigma^{-1} g$.

Use six coordinate blocks of dimension d, indexed by (u, a, b, s, g, o). There are d workspace tokens and one output query. At stage m, their states are

$$z_j = (e_j, P_m e_j, A e_j, S_m e_j, 0, 0) \quad (j \in [d]), \qquad z_q = (0, 0, 0, 0, g, 0). \tag{17}$$

Initialize at $m = 1$, so each workspace token contains $(e_j, (I - \eta\Sigma)e_j, (I - \eta\Sigma)e_j, e_j, 0, 0)$. This is a fixed affine encoding of (Σ, g). The blocks u and b remain unchanged throughout the computation.

Let E_r inject block r into the full token space. Thus $E_r^\top z = z_r$ and $E_r^\top E_c = \delta_{rc} I$ for these blocks. The matrix product identity

$$XY = \sum_{j=1}^{d} (X e_j)(e_j^\top Y) \tag{18}$$

turns a sum of value vectors times bilinear scores into matrix multiplication. We use the notation $M[r, c] = E_r E_c^\top$ and $V[r, c] = E_r E_c^\top$ for a score map and a value copy, respectively. Their separate names indicate their roles in an attention head.

B.2 Every projection matrix

The following table specifies all nonzero heads for one doubling layer, one increment layer, and the final vector readout. Every layer applies (3) to all tokens and sums over all $d + 1$ keys. Unused heads have zero value matrices.

Layer and head	W_K	W_Q	W_V
Doubling 1	$(E_u + E_a)E_u^\top$	$E_a E_a^\top - E_u E_u^\top$	$E_a E_a^\top$
Doubling 2	$E_u E_u^\top$	$E_u E_s^\top$	$E_s E_a^\top$
Increment 1	$E_u E_u^\top$	$E_u E_a^\top$	$E_a E_b^\top$
Increment 2	$E_u E_u^\top$	$E_u E_u^\top$	$-E_a E_a^\top + E_s E_a^\top$
Readout	$E_u E_u^\top$	$E_u E_g^\top$	$\eta E_o E_s^\top$

For a primitive score $M[r, c]$, the factorization is $W_K = E_r E_r^\top$, $W_Q = E_r E_c^\top$. The first doubling head uses the signed factorization

$$W_K^\top W_Q = E_u(E_u + E_a)^\top (E_a E_a^\top - E_u E_u^\top) = E_u(E_a - E_u)^\top.$$

Thus its score is $M[u, a] - M[u, u]$ exactly.

Lemma 12 (Doubling)**.** *The two doubling heads map the state* (17) *at m to the same state at $2m$.*

Proof. For workspace receiver i, head 1 has score $e_j^\top (P_m - I)e_i$ from workspace key j and value $P_m e_j$ in block a. Its update is $P_m(P_m - I)e_i$. Adding the residual gives $P_m^2 e_i = P_{2m} e_i$. Head 2 has score $e_j^\top S_m e_i$ and writes value $P_m e_j$ into block s. Its update is $P_m S_m e_i$, so the residual gives $(S_m + P_m S_m)e_i = S_{2m} e_i$. Both heads read the states at the start of the layer.

The query has zero u block, so its contribution as a key vanishes. Its u, a, s blocks are zero, so all scores for the query as a receiver also vanish. All blocks except a and s remain unchanged. □

Lemma 13 (Increment)**.** *The two increment heads map the state* (17) *at m to the same state at $m + 1$.*

Proof. For workspace receiver i, head 1 sums $\sum_j A e_j (e_j^\top P_m e_i) = A P_m e_i$ into block a. Head 2 has score $e_j^\top e_i = \delta_{ji}$. It subtracts $P_m e_i$ from block a and adds it to block s. After the residual, the new blocks are $A P_m e_i = P_{m+1} e_i$ and $(S_m + P_m)e_i = S_{m+1} e_i$. The same block zeros as in the doubling layer keep the query unchanged and remove its key contribution. □

Lemma 14 (Vector readout)**.** *After the chain reaches k, the readout head writes w_k into block o of the query, leaving every workspace token unchanged.*

Proof. For query receiver q, the score from workspace key j is $e_j^\top g = g_j$. Hence the new output block is $\eta \sum_j S_k e_j g_j = \eta S_k g$. Every workspace receiver has zero g block; the query has zero u block and contributes zero as a key. All output blocks were initially zero. □

B.3 Binary chain and resource count

Write the binary expansion of k from its leading bit to its last bit. Start at $m = 1$. For each remaining bit, apply one doubling layer; if that bit equals one, also apply one increment layer. There are $\lfloor \log_2 k \rfloor$ doublings and $\nu(k) - 1$ increments, where $\nu(k)$ counts the one bits. Therefore

$$\ell_{\text{bin}}(k) = \lfloor \log_2 k \rfloor + \nu(k) - 1 \le 2 \lfloor \log_2 k \rfloor. \tag{19}$$

Proof of Theorem 5. The preceding lemmas implement every operation of the binary chain in one layer with two heads. One readout layer gives $L = \ell_{\text{bin}}(k) + 1$, $H = 2$, $d_{\text{model}} = 6d$, and $n = d + 1$. Each row of every displayed W_Q, W_K, W_V has at most one nonzero entry, of magnitude at most $\max(1, \eta)$. This proves both

$$\max_{i,j} |W_{ij}| \le \max(1, \eta), \qquad \|W\|_{\infty \to \infty} \le \max(1, \eta).$$

These bounds apply to the layer matrices. The initialization is the affine map explicitly specified above. The readout from block o is linear. For $k = 1$, the chain has no layers and the single readout gives ηg. For $k = 2^p$, there are p doublings and depth $p + 1$. □

All matrix products in this proof are valid for an arbitrary square matrix A. Positive definiteness specifies the quadratic family. Stability of A controls convergence as k grows; the finite-k identities use polynomial operations throughout. The bounded projection matrices also do not imply bounded token states. For example, S_m is an unnormalized sum of m matrix powers.

B.4 Demonstrations and an unseen query

Add vector blocks x, x_q of dimension d and scalar blocks y, c, v. The width is now $8d + 3$. The N demonstration tokens carry x_i in block x and y_i in block y. The next d workspace tokens carry e_j in block u. The final query carries its input in block x_q and the constant one in block c. All unspecified blocks are zero. This fixed input format has $n = N + d + 1$ tokens and is affine in the raw prompt variables.

For scalar blocks, E_r is the corresponding coordinate injection. The notation $M[r, c]$ and $V[r, c]$ applies whenever the two blocks have equal dimension. Define

$$\widehat{\Sigma} = \frac{1}{N} X^\top X, \qquad \widehat{g} = \frac{1}{N} X^\top y,$$

where the rows of X are $x_i^\top$.

Lemma 15 (Moment formation)**.** *One layer with the following three heads initializes the workspace and query to* (17) *at $m = 1$, with $A = I - \eta\widehat{\Sigma}$ and $g = \widehat{g}$.*

$$\begin{aligned}
W_K^{(1)} &= E_x E_x^\top, & W_Q^{(1)} &= E_x E_u^\top, & W_V^{(1)} &= -\frac{\eta}{N}(E_a + E_b)E_x^\top, \\
W_K^{(2)} &= E_u E_u^\top, & W_Q^{(2)} &= E_u E_u^\top, & W_V^{(2)} &= (E_a + E_b + E_s)E_u^\top, \\
W_K^{(3)} &= E_y E_y^\top, & W_Q^{(3)} &= E_y E_c^\top, & W_V^{(3)} &= \frac{1}{N} E_g E_x^\top.
\end{aligned}$$

The coordinates carrying demonstrations and the query input remain unchanged.

Proof. Only workspace receivers have a nonzero u block. For workspace j, the first head sums $-\eta N^{-1} \sum_i x_i (x_i^\top e_j)$ into both a and b, and the second head adds e_j into a, b, s. The resulting blocks are Ae_j, Ae_j, e_j. The query's separate x_q block excludes its input from these moments. Only the query has nonzero c; the third head writes $N^{-1} \sum_i y_i x_i = \widehat{g}$ into its g block. Demonstration tokens have zero u, c and receive zero updates. The workspace and query have zero y and add no term to the third head. This exhausts the full attention sum for every receiver. □

Upper bound in Theorem 7. After moment formation, the demonstration tokens have zero u, a, b, s, g, o blocks. They neither contribute to nor receive any chain update. The workspace and query therefore obey the chain invariants already proved. The vector readout writes $w_k = \eta \sum_{t<k} (I - \eta\widehat{\Sigma})^t \widehat{g}$ into the query's o block. Apply one last head with

$$W_K = E_o E_o^\top, \qquad W_Q = E_o E_{x_q}^\top, \qquad W_V = E_v E_c^\top.$$

The only nonzero score is the query's attention to itself, $z_q^\top W_K^\top W_Q z_q = w_k^\top x_q$. Its value is the query's constant one, so the prediction coordinate becomes $x_q^\top w_k$. A fixed linear readout selects that coordinate.

There is one setup layer, $\ell_{\mathrm{bin}}(k)$ chain layers, and two readout layers. Thus $L = \ell_{\mathrm{bin}}(k) + 3$ and $H = 3$, padding chain and readout layers with zero heads. Every row of every projection again has at most one nonzero entry; its magnitude is at most $\max(1, \eta)$ because $N \geq 1$. The empirical covariance can be singular. The proof uses only polynomial matrix identities. When $k = 1$, the three layers output $\eta x_q^\top X^\top y / N$. □

B.5 Degree of the demonstration prediction

Theorem 16 (Demonstration prediction lower bound)**.** *Fix $N, d, k \geq 1$ and $\eta \neq 0$. For any affine encoding of the raw demonstrations and query, exact computation of $x_q^\top \eta \sum_{t<k} (I - \eta\widehat{\Sigma})^t \widehat{g}$ on a set of prompts with nonempty interior requires $3^L \geq 2k + 1$. The lower bound permits any fixed number of workspace tokens independent of the data.*

Proof. The target has total degree at most $2k + 1$ in (X, y, x_q), because $\widehat{\Sigma}$ and $\widehat{g}$ each have degree two. Its top homogeneous part is $\eta(-\eta)^{k-1} x_q^\top \widehat{\Sigma}^{k-1} \widehat{g}$. Within the same variable space, restrict to

$x_i = t_i e_1$ and $x_q = s e_1$. This part becomes

$$\eta(-\eta)^{k-1} s \left(\frac{1}{N} \sum_i t_i^2 \right)^{k-1} \left(\frac{1}{N} \sum_i t_i y_i \right).$$

The coefficient of $s\, t_1^{2k-1} y_1$ is $\eta(-\eta)^{k-1} N^{-k} \neq 0$, so the target degree is exactly $2k+1$. The output degree is at most 3^L by Lemma 8. Equality on a set with nonempty interior makes the two polynomials identical, giving $2k+1 \leq 3^L$. □

C A Lower Bound for Prediction Risk

We now study the optimal predictor in a scalar model that supplies the spectrum and linear term. Let $\sigma \sim \mu$ on $[1/\kappa, 1]$, and let $g \sim \mathcal{N}(0,1)$ be independent of σ. The optimal parameter is $w^\star = g/\sigma$. A network receives affine coordinates in (σ, g) and outputs an estimate $\widehat{w} = \Phi_L(\sigma, g)$ through a fixed linear readout. Given σ, the test input $x_q \sim \mathcal{N}(0, \sigma)$ is independent of g. Define the excess prediction risk by

$$\mathcal{R}(\Phi_L) = \mathbb{E}[(x_q \Phi_L - x_q w^\star)^2] = \mathbb{E}_\sigma \left[\sigma\, \mathbb{E}_g[(\Phi_L - g/\sigma)^2] \right]. \tag{20}$$

The estimator can be any output of the specified network; its dependence on g is unrestricted.

Lemma 17 (Reduction to one polynomial)**.** *If $D \geq 1$ and $\Phi(\sigma, g)$ has total degree at most D, then there is a polynomial $q \in \Pi_D$ with $q(0) = 1$ such that*

$$\mathcal{R}(\Phi) \geq \mathbb{E}_\mu[q(\sigma)^2/\sigma]. \tag{21}$$

Proof. Expand the polynomial in the probabilists' Hermite basis,

$$\Phi(\sigma, g) = \sum_{m=0}^{D} a_m(\sigma)\, \mathrm{He}_m(g), \qquad \deg a_m \leq D - m.$$

For a standard Gaussian, $\mathbb{E}[\mathrm{He}_m(g)\, \mathrm{He}_n(g)] = m!\, \delta_{mn}$ and $\mathrm{He}_1(g) = g$. These normalizations are listed in National Institute of Standards and Technology, Table 18.3.1. Independence gives, pointwise in σ,

$$\mathbb{E}_g[(\Phi - g/\sigma)^2] = \sum_{m=0}^{D} m! \left(a_m(\sigma) - \frac{\delta_{m1}}{\sigma} \right)^2 \geq \left(a_1(\sigma) - \frac{1}{\sigma} \right)^2.$$

Set $q = 1 - \sigma a_1$. Its degree is at most D, and $q(0) = 1$. Multiplication by σ and expectation over μ give the claim. □

Lemma 18 (Constrained polynomial approximation in L^2)**.** *Let $\Lambda = [a, b] \subset (0, \infty)$, and suppose μ has density at least $c > 0$ on Λ. Every $q \in \Pi_D$ with $q(0) = 1$ satisfies*

$$\|q\|_{L^2(\mu)} \geq \frac{\sqrt{c(b-a)}}{(D+1) T_D(\xi)}, \qquad \xi = \frac{a+b}{b-a} > 1. \tag{22}$$

Proof. Set $\widetilde{q}(x) = q((a+b)/2 + (b-a)x/2)$, so $\widetilde{q}(-\xi) = 1$. Exterior Chebyshev extremality gives

$$1 \leq T_D(\xi) \|\widetilde{q}\|_{L^\infty[-1,1]}.$$

Here is an elementary verification of this extremality step. For $D \geq 1$, take the nodes $x_j = \cos(j\pi/D)$, $0 \leq j \leq D$, and their Lagrange basis polynomials ℓ_j. For $t > 1$, the sign of $\ell_j(t)$ is $(-1)^j$. Interpolation of T_D, whose nodal values are $(-1)^j$, yields $\sum_j |\ell_j(t)| = T_D(t)$. Interpolation of any degree-D polynomial r therefore gives $|r(t)| \leq T_D(t) \|r\|_{L^\infty[-1,1]}$. Apply this to $r(x) = \widetilde{q}(-x)$ and $t = \xi$. For $D = 0$, the inequality is immediate.

Let P_j be the Legendre polynomial normalized by $P_j(1) = 1$. The standard identities $\int_{-1}^{1} P_j^2 = 2/(2j+1)$ and $|P_j(x)| \leq 1$ for $|x| \leq 1$ are given in National Institute of Standards and Technology,

Table 18.3.1 and Eq. 18.14.1. Expanding $\widetilde{q}$ in the orthonormal basis $\sqrt{(2j+1)/2}\,P_j$ and applying Cauchy's inequality gives

$$\|\widetilde{q}\|_{L^\infty[-1,1]} \le \frac{D+1}{\sqrt{2}}\|\widetilde{q}\|_{L^2[-1,1]},$$

because $\sum_{j=0}^{D}(2j+1)/2 = (D+1)^2/2$. Finally,

$$\|\widetilde{q}\|^2_{L^2[-1,1]} = \frac{2}{b-a}\int_a^b q(\sigma)^2\,d\sigma \le \frac{2}{c(b-a)}\|q\|^2_{L^2(\mu)}.$$

Combining the three estimates proves (22). □

Theorem 19 (Depth required for small prediction risk)**.** *In the scalar model* (20)*, assume $\kappa \ge 4$ and $d\mu/d\sigma \ge c_\mu > 0$ on $[1/\kappa, 1]$, with c_μ independent of κ. There is a constant $C = C(c_\mu)$ such that, if $E = \log(1/\varepsilon) \ge 1$ and $E \ge C(\log\kappa + \log E)$, every L-layer linear-attention estimator with $\mathcal{R}(\Phi_L) \le \varepsilon$ satisfies*

$$3^L \ge \frac{1}{8}\sqrt{\kappa}\log(1/\varepsilon), \qquad L = \Omega(\log\kappa + \log\log(1/\varepsilon))\,. \tag{23}$$

Proof. Put $D = 3^L$. The degree lemma and Lemmas 17 and 18 give, for $\xi = (\kappa+1)/(\kappa-1)$,

$$\mathcal{R}(\Phi_L) \ge \mathbb{E}_\mu[q^2/\sigma] \ge \|q\|^2_{L^2(\mu)} \ge \frac{c_\mu(1-1/\kappa)}{(D+1)^2 T_D(\xi)^2} \ge \frac{c_\mu/2}{(D+1)^2 T_D(\xi)^2}. \tag{24}$$

Write $\theta = \log((\sqrt{\kappa}+1)/(\sqrt{\kappa}-1))$, so $\xi = \cosh\theta$ and $T_D(\xi) = \cosh(D\theta) \le e^{D\theta}$. Risk at most ε therefore implies

$$D\theta \ge \tfrac{1}{2}E - \log(D+1) - c_1, \qquad c_1 = \max\{0, \tfrac{1}{2}\log(2/c_\mu)\}.$$

Moreover, $\theta = 2\,\mathrm{arctanh}(1/\sqrt{\kappa}) \le 2\sqrt{\kappa}/(\kappa-1)$.

Suppose $D < \sqrt{\kappa}E/8$. Since $\sqrt{\kappa}E \ge 2$, this gives $\log(D+1) \le \frac{1}{2}\log\kappa + \log E$. Set $\Lambda_E = \frac{1}{2}\log\kappa + \log E + c_1$. When $E \ge 6\Lambda_E$, the preceding inequalities yield

$$D \ge \frac{\kappa-1}{2\sqrt{\kappa}}(\tfrac{1}{2}E - \Lambda_E) \ge \frac{3}{8}\sqrt{\kappa}(\tfrac{1}{2}E - \Lambda_E) \ge \frac{1}{8}\sqrt{\kappa}E,$$

a contradiction. The condition $\kappa \ge 4$ gives the middle inequality. Taking $C = 6(1+c_1)$ ensures $C(\log\kappa + \log E) \ge 6\Lambda_E$, since $\log\kappa \ge \log 4 > 1$ and $E \ge 1$. Hence $D \ge \sqrt{\kappa}E/8$, proving (23). □

The normalization $\lambda_{\max} = 1$ fixes the scale of the absolute risk. For a different endpoint $B = \lambda_{\max}$, the substitutions $\widetilde{\sigma} = \sigma/B$ and $\widetilde{\Phi}(\widetilde{\sigma}, g) = B\Phi(B\widetilde{\sigma}, g)$ give $\widetilde{\mathcal{R}} = B\mathcal{R}$ and density $\widetilde{f}(u) = Bf(Bu)$. Both the risk tolerance and the density floor therefore change with scale. The theorem concerns the supplied scalar pair (σ, g); the empirical moments in a raw demonstration prompt have a different joint distribution. The exact demonstration theorem addresses that input format for the prescribed finite-k GD prediction.

D Consequences and Architectural Scope

D.1 Exact depth and a fixed eigenbasis

Corollary 20 (Exact depth complexity)**.** *For fixed d and $\eta > 0$, exact computation of w_k from affine encodings of (Σ, g) on the full SPD quadratic family has depth complexity $\Theta(\log k)$ in full linear self-attention. For fixed N, d, the same asymptotic depth characterizes exact demonstration-to-query prediction with the workspace of Theorem 7.*

Proof. For matrix inputs, combine the lower bound $3^L \ge k$ with the construction of depth $\ell_{\mathrm{bin}}(k)+1$. For demonstration inputs, combine $3^L \ge 2k+1$ with depth $\ell_{\mathrm{bin}}(k)+3$. □

Corollary 21 (A shared eigenbasis)**.** *Fix an orthogonal matrix U, a nondegenerate interval $[\lambda_{\min}, \lambda_{\max}] \subset (0, \infty)$, and $\eta > 0$. Let $\mathcal{F}_U = \{U^\top \operatorname{diag}(\sigma_1, \dots, \sigma_d) U \mid \sigma_j \in [\lambda_{\min}, \lambda_{\max}]\}$. Exact computation of $w_k(\Sigma, g)$ for every $\Sigma \in \mathcal{F}_U$ and every g has depth complexity $\Theta(\log k)$ under affine input encoding. The upper bound has the resources of Theorem 5, independently of U.*

Proof. The column construction applies directly to every $\Sigma \in \mathcal{F}_U$. For the lower bound, set $v = U^\top e_1$, restrict $g = \Sigma v$, and let σ_1 vary while fixing the other eigenvalues. The scalar projection $v^\top w_k$ is $p_k(\sigma_1)$. The input remains affine in σ_1 and v is fixed, giving $3^L \geq k$. For a uniform approximation guarantee controlling this projection to error ε, the corresponding implication is $3^L \geq d_\varepsilon(k, \kappa)$. Its growing quantitative consequence uses the conditions in Theorem 4. □

D.2 MLP blocks

Corollary 22 (Polynomial and ReLU activations)**.** *Append a fixed finite position-wise MLP after every linear-attention layer, with one coordinatewise activation between affine maps. If the activation is a polynomial of degree $p \geq 1$, the output degree is at most $(3p)^L$ for affine inputs. Exact computation of p_k therefore requires $(3p)^L \geq k$, and uniform approximation requires $(3p)^L \geq d_\varepsilon(k, \kappa)$. For ReLU activation, the output is piecewise polynomial with degree at most 3^L on each piece. Exact computation of p_k on a nondegenerate interval still requires $3^L \geq k$.*

Proof. Polynomial activation multiplies the current degree by at most p. After attention the degree recurrence is $D_{\ell+1} \leq 3pD_\ell$, including any affine MLP residual. The exact and approximation conclusions follow as in the main theorems.

For ReLU, fix a complete activation pattern of the finite network. On the corresponding input region every MLP is affine, and the degree bound is 3^L. Restricted to a scalar interval, the finite collection of polynomial sign conditions partitions the domain into finitely many points and intervals. At least one piece has a nonempty interval. If the output equals p_k everywhere, its polynomial on that piece equals p_k identically, forcing $3^L \geq k$. □

The ReLU statement bounds exact computation for a piecewise polynomial model. Uniform approximation also depends on the number and location of the pieces. Smooth nonpolynomial activations and LayerNorm require a different analysis. Polynomial surrogates for such components need explicit bounds on intermediate ranges and propagated error before they yield a depth statement.

D.3 Looped linear attention

Corollary 23 (Total depth under weight sharing)**.** *A block of L linear-attention layers repeated T times has output degree at most 3^{LT} under affine input encoding. Exact computation of p_k requires*

$$LT \geq \lceil \log_3 k \rceil. \tag{25}$$

Uniform approximation requires $3^{LT} \geq d_\varepsilon(k, \kappa)$. For the exact demonstration prediction, $3^{LT} \geq 2k + 1$.

Proof. Unrolling the repeated block gives a depth-LT linear-attention network. Weight sharing restricts its parameters and leaves the degree recurrence unchanged. □

The unrestricted order of total depth is attained by the binary construction with $T = 1$. This gives $L = \ell_{\text{bin}}(k) + 1$ for matrix inputs. Achievability for a prescribed number of unique layers and repeats is a separate resource question.

D.4 Outputting the full trajectory

The semigroup

$$(P_a, S_a) \oplus (P_b, S_b) = (P_a P_b, S_a + P_a S_b) = (P_{a+b}, S_{a+b}) \tag{26}$$

also organizes the intermediate iterates. An explicit realization of all $w_1, \dots, w_k$ follows by placing k copies of the column construction in disjoint feature blocks. Each copy has its own d workspace

tokens and one output token. The common inputs (Σ, g) are copied by an affine embedding, and copy m follows the binary chain for m.

Embed every projection of copy m in its own block and set it to zero elsewhere. Tokens belonging to other copies have zero coordinates in that block, so attention scores between copies are zero. Once a copy finishes its chain, pad it with layers whose heads are zero. Perform all vector readouts in the final layer. The full attention recurrence then computes each copy exactly and gives

$$L = \max_{1 \le m \le k} \ell_{\mathrm{bin}}(m) + 1 \le 2\lfloor \log_2 k \rfloor + 1, \quad H = 2k, \quad d_{\mathrm{model}} = 6dk, \quad n = k(d+1).$$

This construction makes the resource cost of multiple output positions explicit. It establishes logarithmic depth for the full trajectory with these larger width, head, and token budgets.

D.5 Fixed preconditioning

Corollary 24 (Exact preconditioned GD)**.** *Fix $P \succ 0$ and $\eta > 0$. On the family $\mathcal{F}_U$ of Corollary 21 with a nondegenerate spectral interval, exact computation of the P-preconditioned GD iterate for all linear terms has depth complexity $\Theta(\log k)$ under affine input encoding.*

Proof. Set $B = P^{-1}\Sigma$ and $h = P^{-1}g$. Since P is fixed, these are linear transformations of the supplied inputs. The recurrence becomes $w_{t+1} = (I - \eta B)w_t + \eta h$. The column construction with $A = I - \eta B$ therefore gives the exact endpoint in $\ell_{\mathrm{bin}}(k) + 1$ layers. For the lower bound, restrict to $\Sigma = tI$ with t varying over the spectral interval. Take a unit eigenvector v of P with eigenvalue $p > 0$ and set $g = tv$. The projected endpoint is $1 - (1 - \eta t/p)^k$, of degree k in t. Affine encoding and the degree lemma give the lower bound. □

The exact statement applies for every positive step size and condition number. Approximation statements use the eigenvalues of $P^{-1/2}\Sigma P^{-1/2}$ and their specified interval. The effective condition number is $\lambda_{\max}(P^{-1/2}\Sigma P^{-1/2})/\lambda_{\min}(P^{-1/2}\Sigma P^{-1/2})$; the condition number of P alone does not describe that interval. When P and Σ share an eigenbasis, the effective eigenvalues reduce to σ_j/p_j.

D.6 Fixed momentum

For fixed $\beta \in \mathbb{R}$ and $\eta > 0$, consider

$$w_{t+1} = ((1+\beta)I - \eta\Sigma)w_t - \beta w_{t-1} + \eta g, \qquad w_{-1} = w_0 = 0.$$

With $v_t = (w_t, w_{t-1})$, this is

$$v_{t+1} = Bv_t + \eta h, \qquad B = \begin{pmatrix} (1+\beta)I - \eta\Sigma & -\beta I \\ I & 0 \end{pmatrix}, \qquad h = \begin{pmatrix} g \\ 0 \end{pmatrix}.$$

The columns of B and the vector h are affine in (Σ, g). The construction in Appendix B applies to any square matrix, so with matrix dimension $2d$ it computes $v_k = \eta \sum_{t<k} B^t h$. Its resources are $L = \ell_{\mathrm{bin}}(k) + 1$, $H = 2$, $d_{\mathrm{model}} = 12d$, and $n = 2d + 1$; the readout selects the first d coordinates. Induction on the recurrence shows that the highest joint homogeneous part of w_k is $\eta(-\eta)^{k-1}\Sigma^{k-1}g$. It is nonzero, so the degree-k exact lower bound also applies on an open family. No characteristic roots are included in the input.

The Newton endpoint. Newton's method reaches $\Sigma^{-1}g$ in one update on a quadratic. This optimizer identity has a different representational demand. With affine (Σ, g) input, every linear-attention output at finite depth is polynomial, while already in one dimension g/σ cannot equal a polynomial on an open set. Indeed, fixing $g = 1$ would give a polynomial r with $\sigma r(\sigma) = 1$ on an interval, hence identically, which is impossible at $\sigma = 0$. Approximation of the inverse requires its own accuracy and conditioning analysis, as in Appendix C.

D.7 Softmax attention

Softmax replaces the bilinear score by

$$\alpha_{j,i} = \frac{\exp(z_j^\top W_K^\top W_Q z_i)}{\sum_{r=1}^n \exp(z_r^\top W_K^\top W_Q z_i)}.$$

This map is generally nonpolynomial, so the finite polynomial degree ceiling used in the linear-attention theorems does not apply. The softmax experiments in Section 5 measure endpoint performance under that different architecture. Their observed depth profiles are empirical observations; the exact upper and lower depth results in this paper concern full unnormalized linear attention.

E Protocol for the Original Polynomial Sweep

The original polynomial sweep uses three fixed context positions

$$\sigma_j = \frac{1+\kappa^{-1}}{2} + \frac{1-\kappa^{-1}}{2}\cos\left(\frac{(2j-1)\pi}{6}\right), \qquad j = 1, 2, 3.$$

For each task, it draws $a \sim \mathcal{N}(0,1)$ and $\sigma_q \sim \mathrm{Unif}[1/\kappa, 1]$, independently, and predicts $a[1 - (1 - 0.5\sigma_q)^k]$ from the three demonstrations and query. The Gaussian amplitude is shared by the demonstrations and target. The nodes and target polynomial are initially evaluated in float64; the sampler returns float32 tensors to the float32 model. Evaluation accumulates squared errors and target squares across 40 batches of 256 newly sampled tasks and takes their ratio. These draws continue the run's random number stream after training.

The main grid has 11 values of k, 11 depths $\{1, 2, 3, 4, 5, 6, 8, 10, 12, 16, 20\}$, and five seeds $\{42, 123, 456, 789, 1024\}$, for 605 runs. The additional depth $L = 7$ is evaluated for $k \in \{64, 96, 128\}$ with the same seeds, giving 15 further runs. Adam has learning rate $5 \times 10^{-4} \min(1, 6/L)$, 500 warmup steps, cosine decay to one percent of the base rate, batch size 256, and gradient norm clipping at one. The scheduled budget is 8,000 steps for $L \le 6$, 12,000 for $7 \le L \le 12$, and 15,000 for deeper models. The historical implementation skips a nonfinite loss and stops after more than 200 such events; its records retain the event count and final evaluation. The separately specified float64 controls below retain every nonfinite stop and require complete seed sets for a valid median.

F Early Exit across Models

Prompt and measurement. The probe uses one demonstration per class on eight classification tasks. The task collection includes SST-2 [Socher et al., 2013], Rotten Tomatoes [Pang and Lee, 2005], MNLI [Williams et al., 2018], AG News and Yahoo Answers [Zhang et al., 2015], and DBpedia [Lehmann et al., 2015, Zhang et al., 2015]. For each task, the first encountered example of each class provides the demonstration; query indices are sampled with seed 42 from the evaluation split after excluding the first C indices. There are 200 evaluated queries per task. The prompt contains labeled demonstrations followed by the query, with a maximum token length of 2,048. At each layer, the final normalization and language model head map the last hidden state to logits; classification selects the largest logit among the first tokens of the candidate label names. The critical layer is the earliest one with accuracy at least 95% of the accuracy at the final layer. The index rule does not guarantee disjoint demonstrations and queries when a class first occurs after the first C examples. These measurements are exploratory properties of this prompt and readout.

Task inclusion. The reported fit includes a task when its final accuracy is at least $1/C + 0.2$ and its critical layer is at least three. This rule depends on the observed outcomes and yields different task subsets for different models. Table 4 records the subsets explicitly. The slopes for each model, intercepts, and R^2 values are reported together in Table 3; the weak and reversed Pythia associations are part of the same comparison as the positive Qwen associations. Class count varies together with dataset identity, prompt length, and label vocabulary, so the fit describes their joint association across tasks.

G Controlled Experiments in Float64

Compute and reproducibility. Each control ran in four shards on four MetaX C550 GPUs, with four training workers per shard. The polynomial control's recorded elapsed time was 77,526 seconds; the regression control's was 51,369 seconds. Each elapsed time covers all four shards and includes stopped runs. Seeds and sampling protocols are fixed; the MetaX training kernels are not bitwise deterministic.

Table 4: **Task inclusion in the reported probe fits.** A check denotes a task included by the stated rule based on observed outcomes. All eight tasks were evaluated on every model.

Task	C	Qwen 3B	Qwen 1.5B	Pythia 2.8B	Pythia 6.9B
SST-2	2	✓	✓	✓	✓
Rotten Tomatoes	2	✓	✓		✓
MNLI	3	✓			
AG News	4	✓	✓	✓	✓
Emotion	6			✓	✓
Yahoo Answers	10	✓	✓	✓	✓
DBpedia	14	✓	✓	✓	✓
20 Newsgroups	20				

G.1 Polynomial Prediction and the Population Oracle

Task and architecture. The control provides $(\sigma_j, ap_k(\sigma_j))$ at the same three Chebyshev roots as the original sweep, together with $(\sigma_q, 0)$. It draws $a \sim \mathrm{Unif}[-1, 1]$ independently of $\sigma_q \sim \mathrm{Unif}[1/\kappa, 1]$ and uses $\eta = 0.5$. The model has four heads and width 64, updates all four tokens at every layer, and uses unnormalized linear attention with a fixed factor $1/4$. It has residual connections, affine input and output maps, and learned context/query type embeddings. It has no softmax, MLP, LayerNorm, dropout, or attention mask. All floating inputs, parameters, activations, losses, and evaluation calculations use float64.

Training and evaluation. The grid is $k \in \{4, 8, 16, 32, 64, 128\}$, $\kappa \in \{100, 1000\}$, and $L = 1, \ldots, 10$, with ten seeds $2026072500, \ldots, 2026072509$ per cell. Adam uses batch size 256, learning rate $5 \times 10^{-4} \min(1, 6/L)$, 500 warmup steps, cosine decay to one percent of the base rate, and gradient norm clipping at one. Each run has budget $\max(15000, 2000L \log_2 k)$. A nonfinite loss ends the run and is retained as an invalid outcome. The final checkpoint is evaluated on 65,536 independent tasks with seed $20260725 + 10000k + \kappa$; all depths and training seeds share that evaluation set within a (k, κ) cell. NMSE is the ratio of summed squared errors to summed squared targets. The thresholds are $10^{-3}, 10^{-5}, 10^{-8}, 10^{-12}$. A median at a given depth requires all ten seeds to have finite evaluations after completing their budgets. The frozen critical depth statistic requires a complete depth grid and uses the strict crossing predicate $\widehat{\mathrm{NMSE}} < \varepsilon$.

Population comparison. For the same query distribution, define

$$d_\varepsilon^{L^2} = \min\left\{ D : \min_{q \in \Pi_D} \frac{\mathbb{E}_\sigma[(p_k(\sigma) - q(\sigma))^2]}{\mathbb{E}_\sigma[p_k(\sigma)^2]} \le \varepsilon \right\}.$$

If an output $F(a, \sigma)$ has total degree at most D, its projection onto functions $aq(\sigma)$ is given by $q(\sigma) = \mathbb{E}_a[aF(a, \sigma)]/\mathbb{E}[a^2]$. This projection has degree at most D in σ and cannot increase population squared error against $ap_k(\sigma)$. Independence of a and σ cancels the amplitude variance in the normalized population error. Thus population NMSE at most ε requires $3^L \ge d_\varepsilon^{L^2}$. Finite evaluation NMSE is a sample statistic; a deterministic bound for that statistic would require an additional argument relating it to the population error.

The oracle projects p_k onto shifted Legendre polynomials of degrees zero through k using weighted Gaussian quadrature with Legendre nodes. The primary calculation uses 4,096 nodes, and a separate 8,192-node calculation gives the same minimum degrees in all 48 threshold cells. Table 5 gives the values at 10^{-12}. Even this accuracy can require a degree substantially smaller than k.

Training outcomes. All 1,200 planned run identifiers are recorded. There are 1,131 finite evaluations after complete budgets and 69 stops at nonfinite loss, all at depths at least seven. Every (k, κ) grid contains an incomplete depth, so every formal critical depth result is undefined under the frozen completeness rule. None of the complete medians over ten seeds reaches 10^{-12}. The population oracle has a stable numerical degree at both quadrature resolutions. The missing complete training grids leave comparisons with empirical critical depths unevaluable. At 10^{-12}, the approximation degrees are below $0.9k$ for $k \ge 8$. Invalid grids are recorded separately from right censoring.

Table 5: **Population approximation degree at NMSE 10^{-12}.** Both quadrature resolutions give these same values.

	$k=4$	8	16	32	64	128
$\kappa = 100$	4	7	10	14	19	26
$\kappa = 1000$	4	7	10	14	19	27

G.2 Condition Number in Demonstration Regression

Task. We draw $g \sim \mathcal{N}(0, I_d)$ and use $w^\star = \Sigma^{-1} g / \sqrt{\mathrm{tr}(\Sigma^{-1})}$. Each prompt contains $N = 2d$ noiseless demonstrations with $x_i \sim \mathcal{N}(0, \Sigma)$, $y_i = x_i^\top w^\star$, and an independent query from the same task. The population eigenvalues are geometrically spaced between $1/\kappa$ and one. The two families are diagonal covariance and its conjugation by a fixed orthogonal DCT-II matrix. The fixed rotation and shared latent sampling pair the task distributions. The condition number refers to the population covariance.

Protocol. The grid is $\kappa \in \{10, 30, 100, 300, 1000, 3000\}$, $d \in \{4, 16\}$, both covariance families, and depths 1 through 12, with the same ten fixed seeds as the polynomial control. The model has four heads, width 64, full unnormalized linear attention, and no LayerNorm or MLP. All floating computations use float64, with native float64 parameter initialization. Adam uses the polynomial control's learning rate schedule, batch size 128, and budget $\max(15000, 2000L)$. Evaluation uses 65,536 independent tasks per $(\kappa, d, \text{family})$, shared across depths and training seeds, with paired latent draws across the two covariance families. The frozen thresholds are 10^{-3} and 10^{-5}. A strict median requires ten finite runs with complete budgets; any incomplete depth makes its axis's formal critical depth undefined.

Outcomes. Of 2,880 recorded runs, 1,869 complete with finite evaluations and 1,011 stop at nonfinite loss or gradient norm. All 48 formal critical depth cells are undefined, so the preregistered fits against $\log_2 \kappa$, κ, $\sqrt{\kappa}$, and the error proxy after one population GD step are unestimable. Table 6 reports the descriptive comparison at $L = 6$, including the two incomplete seed sets. These measurements characterize performance at that fixed depth.

Table 6: **Strict median NMSE over ten seeds at depth six.** NA denotes an incomplete seed set. Columns distinguish dimension and covariance family.

κ	$d=4$, diagonal	$d=4$, rotated	$d=16$, diagonal	$d=16$, rotated
10	$1.991 \cdot 10^{-4}$	$1.883 \cdot 10^{-4}$	$2.639 \cdot 10^{-4}$	$3.031 \cdot 10^{-4}$
30	$3.041 \cdot 10^{-4}$	$3.444 \cdot 10^{-4}$	$1.430 \cdot 10^{-3}$	$2.081 \cdot 10^{-3}$
100	$5.546 \cdot 10^{-4}$	$5.391 \cdot 10^{-4}$	$4.913 \cdot 10^{-2}$	$5.326 \cdot 10^{-2}$
300	$1.475 \cdot 10^{-2}$	NA	$9.999 \cdot 10^{-1}$	$9.552 \cdot 10^{-1}$
1000	1.000	1.000	1.000	1.000
3000	1.000	NA	1.000	1.000

H LayerNorm Sweep with 20 Seeds

The reported normalization sweep uses the grid $k \in \{8, 16, 32, 64, 128\}$, $L \in \{1, \ldots, 12\}$, three normalization variants, and 20 independent seeds per cell (∼3,600 trained models). The table reports median NMSE. Where supplied, bracketed subscripts give bootstrap 95% confidence intervals ($n_{\mathrm{boot}} = 10^4$), and superscripts give the number of finite runs out of 20. The remaining entries provide point estimates or qualitative summaries, without complete records of uncertainty and sample counts. NA denotes cells without a finite median under this optimizer setting. The table reports the behavior at finite error of the three normalization variants, including cells with incomplete or nonfinite runs. The polynomial degree theorem applies to the attention-only architecture without LayerNorm.

(k, L)	No LN	Post LN	Pre LN
$(8, 1)$	$2.2{\cdot}10^{-2}$ $^{20}_{[2.2,2.2]\cdot10^{-2}}$	$2.8{\cdot}10^{-5}$ $^{20}_{[2.3,4.0]\cdot10^{-5}}$	$3.8{\cdot}10^{-5}$ $^{20}_{[3.0,4.3]\cdot10^{-5}}$
$(8, 2)$	$1.2{\cdot}10^{-6}$ $^{20}_{[0.8,1.8]\cdot10^{-6}}$	$5.3{\cdot}10^{-6}$ $^{20}_{[2.9,9.0]\cdot10^{-6}}$	$6.3{\cdot}10^{-6}$ $^{20}_{[4.5,9.0]\cdot10^{-6}}$
$(8, 3)$	$1.3{\cdot}10^{-7}$ $^{20}_{[0.8,1.8]\cdot10^{-7}}$	$3.1{\cdot}10^{-6}$ $^{20}_{[2.1,6.4]\cdot10^{-6}}$	$3.8{\cdot}10^{-6}$ $^{20}_{[2.2,7.9]\cdot10^{-6}}$
$(8, 4)$	$7.8{\cdot}10^{-8}$ $^{20}_{[6.0,15.]\cdot10^{-8}}$	$2.0{\cdot}10^{-6}$ $^{20}_{[1.2,6.0]\cdot10^{-6}}$	$2.3{\cdot}10^{-6}$ $^{20}_{[1.5,6.9]\cdot10^{-6}}$
$(8, 5)$	$1.5{\cdot}10^{-7}$ $^{11}_{[1.3,2.6]\cdot10^{-7}}$	$9.4{\cdot}10^{-7}$ $^{20}_{[5.7,52.]\cdot10^{-7}}$	$1.9{\cdot}10^{-6}$ $^{20}_{[1.0,6.3]\cdot10^{-6}}$
$(8, \geq 6)$	NA-	stable, $\sim 5{\cdot}10^{-7}$	stable, $\sim 5{\cdot}10^{-7}$
$(16, 1)$	$2.6{\cdot}10^{-2}$ $^{20}_{[2.6,2.7]\cdot10^{-2}}$	$3.5{\cdot}10^{-5}$ $^{20}_{[2.9,3.9]\cdot10^{-5}}$	$3.3{\cdot}10^{-5}$ $^{20}_{[2.9,3.7]\cdot10^{-5}}$
$(16, 2)$	$6.4{\cdot}10^{-6}$ $^{20}_{[4.1,9.4]\cdot10^{-6}}$	$4.9{\cdot}10^{-6}$ $^{20}_{[4.7,11.]\cdot10^{-6}}$	$6.4{\cdot}10^{-6}$ $^{20}_{[4.5,12.]\cdot10^{-6}}$
$(16, 3)$	$2.1{\cdot}10^{-7}$ $^{20}_{[1.8,3.2]\cdot10^{-7}}$	$2.7{\cdot}10^{-6}$ $^{20}_{[1.9,6.5]\cdot10^{-6}}$	$4.5{\cdot}10^{-6}$ $^{20}_{[3.3,7.4]\cdot10^{-6}}$
$(16, 4)$	$8.2{\cdot}10^{-8}$ $^{20}_{[7.5,13.]\cdot10^{-8}}$	$2.0{\cdot}10^{-6}$ $^{20}_{[1.5,4.6]\cdot10^{-6}}$	$2.3{\cdot}10^{-6}$ $^{20}_{[1.7,6.8]\cdot10^{-6}}$
$(16, \geq 6)$	NA-	stable, $\sim 7{\cdot}10^{-7}$	stable, $\sim 1{\cdot}10^{-6}$
$(32, 1)$	$2.0{\cdot}10^{-2}$	$3.6{\cdot}10^{-5}$	$3.5{\cdot}10^{-5}$
$(32, 3)$	$9.2{\cdot}10^{-7}$	$4.8{\cdot}10^{-6}$	$5.5{\cdot}10^{-6}$
$(32, 4)$	$1.5{\cdot}10^{-7}$	$3.5{\cdot}10^{-6}$	$3.3{\cdot}10^{-6}$
$(32, \geq 6)$	NA-	stable, $\sim 5{\cdot}10^{-7}$	stable, $\sim 8{\cdot}10^{-7}$
$(64, 1)$	$1.2{\cdot}10^{-2}$	$2.8{\cdot}10^{-5}$	$2.8{\cdot}10^{-5}$
$(64, 3)$	$6.1{\cdot}10^{-6}$	$5.7{\cdot}10^{-6}$	$6.3{\cdot}10^{-6}$
$(64, 4)$	$7.6{\cdot}10^{-7}$	$4.8{\cdot}10^{-6}$	$6.4{\cdot}10^{-6}$
$(64, \geq 6)$	NA-	stable, $\sim 9{\cdot}10^{-7}$	stable, $\sim 1{\cdot}10^{-6}$
$(128, 1)$	$6.2{\cdot}10^{-3}$	$2.4{\cdot}10^{-5}$	$2.4{\cdot}10^{-5}$
$(128, 3)$	$5.4{\cdot}10^{-5}$	$5.6{\cdot}10^{-6}$	$1.0{\cdot}10^{-5}$
$(128, 4)$	$7.4{\cdot}10^{-6}$	$6.6{\cdot}10^{-6}$	$9.2{\cdot}10^{-6}$
$(128, \geq 8)$	NA-	stable, $\sim 2{\cdot}10^{-6}$	stable, $\sim 2{\cdot}10^{-6}$

I ICL Linear Regression across Covariance Families

The reported regression experiment covers two thresholds and five covariance families (Figure 3). Each prompt contains $N = 64$ demonstrations and a query, with $x_i \sim \mathcal{N}(0, \Sigma)$, $y_i = x_i^\top w^\star$, and $w^\star \sim \mathcal{N}(0, I_d/d)$ in dimension $d = 16$. Covariances are normalized to trace d and drawn from the stated family for each training batch; the rotated family uses a newly sampled orthogonal basis. Evaluation fixes the covariance within each family and condition number and uses 2,048 tasks with a random number stream independent of training. The model has $d_{\text{model}} = 64$, $H = 4$, no MLP or LayerNorm, and batch size 128. The scheduled budgets are 8,000 steps for $L \leq 6$ and 12,000 for $L \in \{8, 10, 12\}$, with five seeds per cell. The tested depth grid is $\{1, 2, 3, 4, 5, 6, 8, 10, 12\}$. $L_{\text{crit}}^{\text{reg}}$ is the smallest tested depth whose median NMSE crosses the stated threshold; NA marks an unobserved crossing in the reported results.

	L_{crit} at $\varepsilon = 10^{-2}$			L_{crit} at $\varepsilon = 10^{-3}$		
Family	κ=10	κ=100	κ=1000	κ=10	κ=100	κ=1000
diagonal	3	4	5	5	8	NA
rotated diagonal	3	4	5	6	8	NA
low-rank	4	8	NA	8	NA	NA
Wishart dense	3	4	4	5	8	8
power law	3	4	8	6	8	NA

Slope fit per family. OLS regression of L_{crit} on $\log_2 \kappa$ at $\varepsilon = 10^{-2}$.

Family	Slope	Intercept	R^2
diagonal	0.30	2.00	1.00
rotated diagonal	**0.30**	**2.00**	**1.00**
Wishart dense	0.15	2.67	0.75
power law	0.75	0.00	0.89
low-rank	1.20	0.00	1.00 (n=2)

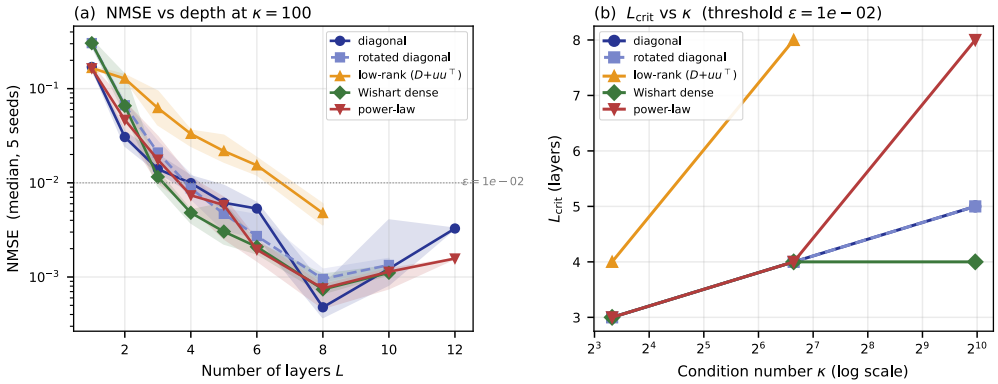

Figure 3: **Prediction error across covariance families. (a)** NMSE versus depth at $\kappa = 100$ for five covariance families (median over 5 seeds, range across seeds). **(b)** L_{crit} at $\varepsilon = 10^{-2}$ versus $\log_2 \kappa$. The diagonal and rotated covariance curves coincide at the displayed 10^{-2} threshold; the full table also reports the 10^{-3} threshold.

These fits summarize the reported threshold crossings on three condition numbers. The low-rank fit uses only two observed crossings. The comparison concerns trained endpoint performance; an exact representation result does not require different training runs to have identical critical depths.

J Additional Experimental Details

This section retains the reported tables of endpoint error, discrete degree estimates, architecture ablations, and regression results from the original experiments. The new float64 control protocols and training outcomes appear in Appendix G. The method for estimating effective degree used to populate Table 8 is given in Appendix J.1 below.

J.1 Discrete Estimate of Effective Degree

The architectural ceiling $D_{\text{arch}}(L) = 3^L$ bounds the exact polynomial degree of an output of linear attention under the assumptions of Lemma 2. To describe a trained predictor, the reported experiments use a separate estimate based on 64 sampled query values.

Definition 25 (Effective approximation degree)**.** For a scalar function f on Λ and a tolerance $\tau > 0$, define

$$D_{\text{eff}}(f;\tau) = \min\{D \in \mathbb{N}_0 : \inf_{q \in \Pi_D} \|f - q\|_{L^\infty(\Lambda)} \le \tau\}. \tag{27}$$

Reported sampling procedure. For each trained predictor $\widehat{p}$, evaluate at $N = 64$ Chebyshev roots after mapping Λ to $[-1, 1]$. Writing $t_n = \cos((2n-1)\pi/(2N))$, the coefficients of the resulting interpolant are

$$c_0 = \frac{1}{N}\sum_{n=1}^{N} \widehat{p}(\sigma_n), \qquad c_j = \frac{2}{N}\sum_{n=1}^{N} \widehat{p}(\sigma_n) T_j(t_n), \quad 1 \le j < N.$$

The discrete truncation estimate is

$$\widehat{D}(\widehat{p};\tau) = \min\left\{ D : \sum_{j=D+1}^{N-1} |c_j| \le \tau \right\}, \qquad \tau = 10^{-6}.$$

Table 8 reports the original estimates over three seeds. For the interpolant, the coefficient tail bounds the uniform error of truncation because $|T_j(t)| \le 1$ on $[-1, 1]$. For the underlying predictor, contributions above the sampling resolution can alias into the retained coefficients. Thus this estimate from a finite grid alone supplies no certificate for the continuous approximation degree in Definition 25.

Relation to the exact degree. When $\widehat{p}$ is a polynomial of degree at most 3^L, $D_{\mathrm{eff}}(\widehat{p};\tau) \le 3^L$ follows by choosing $q = \widehat{p}$. This statement about the exact function is distinct from the numerical truncation estimate. Approximating p_k to a fixed positive tolerance can require much less than degree k, as the population oracle in Table 5 illustrates for its stated metric.

Table 7: **Repeated training at depth two.** Reported terminal NMSE ranges for $L = 2$ with 50 seeds and 50,000 scheduled training steps. All reported errors exceed 10^{-5}.

k	NMSE range (50 seeds)
32	$[1.07, 1.37] \times 10^{-5}$
64	$[2.44, 3.16] \times 10^{-4}$
128	$[7.61, 10.8] \times 10^{-4}$

Table 8: **Reported discrete truncation estimates for trained models.** The estimates use 64 sampled nodes and three seeds. The exact polynomial degree ceiling 3^L appears in the header; estimates from the finite grid require the qualifications in Appendix J.1.

k	$L=1$ ($3^1=3$)	$L=2$ ($3^2=9$)	$L=3$ ($3^3=27$)	$L=4$ ($3^4=81$)
4	1	3	3	3
8	1	4	4	4
16	1	5	6	6
32	1	8	8	8
64	1	8	12	12

Table 9: L_{crit} **at** $\varepsilon = 10^{-5}$ **across LayerNorm variants (20 seeds).** Entries marked † use fewer than 10 finite runs under the same optimizer setting. Empty entries have no finite median in the reported grid.

k	No LN	Pre LN	Post LN	Stable depth (No LN)
8	2	2	2	up to $L = 4$
16	2	2	2	up to $L = 4$
32	3	2	2	up to $L = 5^{\dagger}$
64	3	2	2	up to $L = 5^{\dagger}$
128	4	4	2	up to $L = 6^{\dagger}$

Table 10: **Reported errors at two condition numbers.** Median NMSE by depth and condition number (10 seeds, $d_{\text{model}} = 256$, $H = 8$). The smallest reported error at $\kappa = 100$ occurs at depth four.

	$L=4$	$L=6$	$L=8$	$L=12$
$\kappa = 100$	**3.2e-3**	6.7e-2	6.7e-2	5.7e-2
$\kappa = 1000$	4.0e-3	4.0e-3	**3.2e-3**	6.1e-3

Table 11: **Depth benefit on real regression data.** Median NMSE on real datasets with in-context learning (3 seeds, $n_{\text{ctx}} = 40$). The datasets are Diabetes [Efron et al., 2004] and Superconductor [Hamidieh, 2018]. Best depth per dataset in bold.

Dataset	d	$L=1$	$L=2$	$L=3$	$L=4$	$L=8$	$L=12$
Diabetes	10	1.1e-4	2.1e-5	**1.3e-5**	2.2e-5	3.0e-5	2.3e-5
Superconductor	81	2.6e-2	1.6e-2	9.4e-3	9.8e-3	6.5e-3	**5.3e-3**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the exact depth characterization for the specified linear-attention model. Table 1 distinguishes exact computation, uniform approximation, prediction risk, and empirical comparisons.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The Limitations and Scope subsection before the conclusion collects the architectural, workspace, accuracy, and empirical limitations. The quantitative approximation conditions also appear with Theorem 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The main theorems state their assumptions. Appendices A.1, B and C provide the complete degree arguments, explicit attention projections, token invariants, resource counts, and risk proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The main text and appendix disclose the models or datasets, protocols, metrics, baselines, and experimental settings needed to reproduce or verify the main claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The paper provides detailed experimental protocols and complete mathematical constructions. A single publicly accessible code and data package covering all reported experiments is not provided with this revision.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The experiments specify the relevant task construction, splits or sampling protocol, model settings, decoding/training parameters, baselines, and evaluation metrics in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Synthetic results use medians across stated seed counts; Appendix I shows ranges across seeds. Appendix G reports finite-run counts and incomplete cells. The pretrained-model fits are descriptive exploratory summaries without significance tests.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: Appendix G reports the MetaX C550 hardware, worker configuration, and recorded elapsed times for the 1,200-run polynomial and 2,880-run regression controls, including nonfinite stops. Complete per-model memory and timing records and aggregate compute for the original sweeps are unavailable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work studies mathematical models and existing research datasets and models. The authors have checked conformance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A].

Justification: The work is primarily a foundational or diagnostic study of implicit optimization and Transformer depth and does not introduce a direct user-facing deployment artifact or application-specific system with clear societal-impact claims.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release a new high-risk model, scraped dataset, or deployment artifact requiring controlled-access safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No].

Justification: The pretrained models and benchmark datasets are identified in Section 5.3 and Appendices E and F. Relevant publications are cited, but the paper does not document a complete license and terms-of-use inventory for every dataset redistribution used in the historical experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A].

Justification: The paper does not release a new dataset, benchmark, trained model, or user-facing software asset as part of the submission.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The work does not involve crowdsourcing or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The work does not involve human-subject experiments, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes].

Justification: Qwen2.5 and Pythia are objects of the exploratory layer-probe experiment. Section 5.3 and Appendix F identify the model families and describe their prompts, readouts, and task-selection protocol.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.