A DEPTH–WIDTH SEPARATION FOR QUADRATIC IN-CONTEXT LEARNING WITH PROTECTED-LABEL BILINEAR TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quadratic in-context regression has $\Theta(d^2)$ task coefficients for inputs in $\mathbb{R}^d$, although each prediction is a scalar. Can depth reduce the token width needed to predict these tasks? We study quadratic regression with Gaussian inputs using a bilinear Transformer whose attention can access labels but whose feed-forward maps cannot read or modify them. For one or two blocks at width $w = o(d^2)$, the minimum achievable relative risk converges to one, uniformly over context length. Three blocks suffice at width $2d + 5$: an explicit network reaches relative risk at most ε with $O(d^2/\varepsilon)$ context examples. Thus the minimum width for fixed relative risk below one falls from $\Theta(d^2)$ to $\Theta(d)$ between two and three blocks. The construction evaluates the quadratic projection-kernel estimator without storing all quadratic features. Attention applies a context-dependent matrix to the query, and a later block completes the prediction using $O(d)$-dimensional states. This shows that sequential computation through context can replace explicit storage of a quadratic feature representation.

1 INTRODUCTION

Large language models can adapt their predictions to examples in a prompt without updating their parameters, a behavior known as in-context learning (ICL) (Brown et al., 2020). Regression provides a controlled setting for studying what a fixed Transformer computes from such examples (Garg et al., 2022). For linear regression, forward passes can implement estimation procedures such as gradient descent and ridge regression (Akyürek et al., 2023; von Oswald et al., 2023). These results illuminate ICL computation; nonlinear tasks raise the further question of what architectural resources that computation requires.

Quadratic regression exposes this resource question with minimal nonlinear structure. A quadratic task in d variables has $N = \binom{d+2}{2} = \Theta(d^2)$ monomial coefficients, yet predicts only a scalar at each query. An explicit feature representation is quadratic in size, but a predictor might compute the answer in stages without retaining every feature at once. Token width limits simultaneous intermediate state; depth allows successive rounds of local computation and context aggregation. Quadratic ICL thus separates task dimension from the state needed for one prediction, complementing resource-scaling analyses of linear ICL (Bordelon et al., 2026) and work on nonlinear representations (Cheng et al., 2024; Kim & Suzuki, 2024).

We study noiseless quadratic tasks with Gaussian inputs and independent Gaussian coefficients in the raw monomial basis. To track multiplication and context aggregation exactly, we use the bilinear Transformer of Sun et al. (2026): each block applies a local bilinear feed-forward map followed by linear attention. In its protected-label architecture, attention can read labels but feed-forward maps cannot read or modify the label coordinate; the query is excluded from attention sources. This access rule separates label transfer from local multiplication. Attention must first make labels available in feature coordinates before a feed-forward map can use them. It also determines the exact block count: if feed-forward maps may read labels directly, a two-block linear-width circuit computes the kernel estimator constructed below (Appendix B.1). Our two-to-three-block threshold concerns the protected-label model.

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

Prior work identifies a one-block feature-space bottleneck and gives linear-width constructions that cycle through feature groups as depth grows (Sun et al., 2026). This leaves a question: can a constant number of blocks avoid quadratic width while attaining a fixed prediction accuracy? We consider sufficiently long contexts so that the answer reflects an architectural limitation rather than a lack of examples.

We find a sharp two-to-three-block threshold. For any fixed target relative risk below one, one or two blocks require quadratic width even with arbitrarily many examples, whereas three blocks achieve the target with linear width and sufficiently long context.

The construction explains the width reduction. Attention applies a label-weighted operator defined by the context to the query, producing a linear-size vector; a later bilinear map combines that vector with the preserved query. The network never stores the full quadratic-size operator. For the lower bound, counting fixed features is insufficient because the query features may depend on the task. We instead bound the dimension of the query-function space available after two blocks and use the remaining task variance to bound prediction error.

Our contributions. The depth–width separation is the main result. We then identify the mechanisms behind both bounds and derive statistical consequences and a covariance extension.

- **A sharp two-to-three-block separation.** With sufficiently long context, one or two blocks require $\Theta(d^2)$ width for any fixed relative risk below one, whereas three blocks need only $\Theta(d)$. The shallow lower bound holds at every context length; the upper bound uses an explicit width-$2d + 5$ circuit (Theorem 3.1 and Corollary 3.2).
- **Structural mechanisms behind the separation.** The circuit applies a context-dependent matrix to the query using $O(d)$-dimensional states. The lower bound limits the adaptive query-function space of two-block networks and controls its remaining task variance by a spectral tail (Sections 3.1–5).
- **Statistical consequences and robustness.** At matched linear width, a third block gives a nonvanishing risk reduction; for $d \ll n \ll N$, this reduction matches the Bayes-optimal one to first order. The separation extends to uniformly well-conditioned Gaussian coefficient covariances in an orthonormal quadratic basis (Section 6 and Corollary F.1).

1.1 Related work

Nonlinear features and the closest depth–width comparison. We retain the Gaussian quadratic task distribution, bilinear-then-attention order, protected label, and query-exclusion source mask of Sun et al. (2026). Their Proposition 4 identifies a one-block feature-space bottleneck, and their Theorem 7 constructs block-coordinate descent at width $2d + 1$, visiting d feature groups per full cycle. These results leave open whether constant depth permits linear width at fixed relative risk. We resolve that question through a two-block lower bound and a three-block construction. The lower bound allows query features to depend on the task. Other nonlinear ICL analyses study functional-gradient computation (Cheng et al., 2024), feature learning in attention (Kim & Suzuki, 2024), low-dimensional single-index tasks (Oko et al., 2024), and inference-time feature adaptation (Nishikawa et al., 2025). Our result concerns the full class of quadratic functions and the computations representable at a given width and block count.

Learning algorithms and architectural resources. Function-class experiments (Garg et al., 2022) motivate analyses of gradient descent and regression estimators implemented in context (Akyürek et al., 2023; von Oswald et al., 2023). Linear-task theory also characterizes preconditioned updates (Ahn et al., 2023), optimal one-step linear attention (Mahankali et al., 2024), and predictors obtained through pretraining (Zhang et al., 2024). Joint scaling with depth, width, and context length (Bordelon et al., 2026) and depth separations for linear dynamical systems (Cole et al., 2025) study related resource questions. Here the nonlinear task has quadratically many coefficients, yet prediction can use linear token width at constant depth.

Kernel prediction and statistical information. Attention constructions realize polynomial and spline features (Hsu et al., 2026), Gaussian-kernel smoothing (Shen et al., 2026), and preconditioned

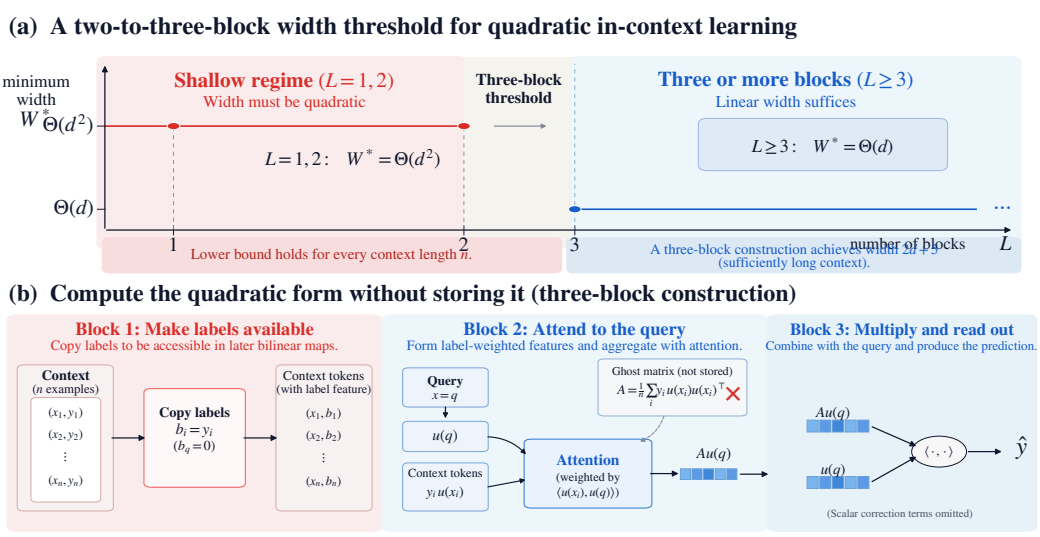

Figure 1: **The width threshold and its computational mechanism.** (a) For Gaussian quadratic tasks and the protected-label architecture, the minimum width at fixed relative risk below one changes from $\Theta(d^2)$ with one or two blocks to $\Theta(d)$ with three or more, given sufficiently long context (Corollary 3.2); the shallow lower bound holds for every context length. (b) The three-block circuit copies labels into feature coordinates, computes $Au(q)$ through attention without storing A, and combines it with the preserved $u(q) = (1, q)$. The final prediction includes scalar corrections omitted from the schematic (Section 5).

Richardson iteration for Gaussian kernel ridge regression (Yan et al., 2026). Our circuit evaluates the quadratic projection kernel through $O(d)$-dimensional intermediate vectors and admits an exact finite-context risk calculation. Bayesian accounts of ICL (Panwar et al., 2024) motivate comparing that risk with the information in the prompt. Whereas asymptotic analyses of kernel regression in the quadratic regime study empirical matrix inverses (Pandit et al., 2025), our Bayes comparison follows from noiseless task measurements and the population feature spectrum. It provides an architecture-independent lower bound on three-block risk.

2 Problem Setup

2.1 Quadratic Task Distribution

A prompt consists of context examples $(x_i, y_i)_{i=1}^n$ and an unlabeled query q. Let the raw quadratic feature map contain each monomial once:

$$\phi(x) = \big(1, x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, (x_j x_k)_{j<k}\big) \in \mathbb{R}^N, \qquad N = \binom{d+2}{2}.$$

For each prompt, independently sample

$$\begin{aligned} \theta \sim \mathcal{N}(0, I_N), \quad & x_1, \ldots, x_n, q \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d), \\ f_\theta(x) = \theta^\top \phi(x), \quad & y_i = f_\theta(x_i), \end{aligned} \tag{1}$$

with θ independent of the inputs. Labels are noiseless. The coefficient prior is defined in this raw-monomial basis and is therefore basis dependent. In an orthonormal Gaussian quadratic basis, the induced coefficients are correlated and have unequal variances.

2.2 PROTECTED-LABEL BILINEAR TRANSFORMER

We use the bilinear Transformer architecture of Sun et al. (2026) with the following label-access restriction. For integers $d, n \geq 1$ and $w \geq d+1$, each token has w feature coordinates and one label coordinate. The initial columns of $Z \in \mathbb{R}^{(w+1)\times(n+1)}$ are $z_i = (1, x_i, 0, \ldots, 0, y_i)$ for the context and $z_q = (1, q, 0, \ldots, 0, 0)$ for the query. Width w counts all non-label coordinates, including the constant, input, and workspace coordinates.

Each block applies a bilinear feed-forward sublayer followed by a linear-attention sublayer:

$$\mathsf{FF}_\ell(Z) = Z + (\widetilde{W}_{0,\ell} Z) \odot (\widetilde{W}_{1,\ell} Z), \tag{2}$$

$$\mathsf{Attn}_\ell(Z) = Z + \frac{1}{n} P_\ell Z M Z^\top Q_\ell Z, \qquad M = \mathrm{diag}(I_n, 0). \tag{3}$$

Here $\odot$ denotes coordinatewise multiplication, $W_{j,\ell} \in \mathbb{R}^{w\times w}$, and $\widetilde{W}_{j,\ell} = \left(\begin{smallmatrix} W_{j,\ell} & 0 \\ 0 & 0 \end{smallmatrix}\right)$. The matrices $P_\ell, Q_\ell \in \mathbb{R}^{(w+1)\times(w+1)}$ are arbitrary. We call the restriction that the feed-forward map neither reads nor modifies the label coordinate *protected-label access*; attention may access labels. The query is excluded from the attention source set, so every context-token update is independent of the query. An L-block network consists of L feed-forward–attention pairs, or $2L$ sublayers.

2.3 RELATIVE RISK AND WIDTH COMPLEXITY

Throughout, d is the input dimension, n the context length, w the feature width, and L the number of blocks. Parameters ω are task independent but may depend arbitrarily on (d, w, L, n); we impose no norm constraint. If $Z^{(L)}$ is the final token matrix, define the network prediction as $\widehat{f}_\omega(q) := -[Z^{(L)}]_{y,q}$. Its relative risk is

$$\mathcal{R}_n(\omega) := \frac{\mathbb{E}\big(\widehat{f}_\omega(q) - f_\theta(q)\big)^2}{V}, \qquad V := \mathbb{E} f_\theta(q)^2 = N + 2d. \tag{4}$$

The expectation is over the task and all inputs. The zero predictor has relative risk one. For target risk $\varepsilon \in (0,1)$, define the minimum required width by

$$W^\star(d, L, n, \varepsilon) := \min\{w \geq d+1 : \inf_\omega \mathcal{R}_n(\omega) \leq \varepsilon\},$$

where the infimum is over L-block networks of width w in this model, and $W^\star = \infty$ if no width reaches the target. We study how $W^\star$ scales with d as the block count L changes.

3 THE THREE-BLOCK WIDTH THRESHOLD

We state the separation before proving its lower- and upper-bound mechanisms in Sections 4 and 5.

For one or two blocks, the key risk bound is

$$\mathcal{R}_n \;\geq\; \frac{[N - 3w - 3]_+}{N + 2d}, \tag{5}$$

where $[t]_+ = \max\{t, 0\}$. When $w = o(d^2)$, the right-hand side tends to one. Because the zero predictor has risk one, the minimum achievable relative risk converges to one regardless of context length. Theorem 3.1 combines this lower bound with an explicit three-block construction.

Theorem 3.1 (Two-to-three-block separation)**.** *Under (1)–(4), the following hold.*

1. *Every one- or two-block network of width w satisfies (5), for every $n \geq 1$.*
2. *There is a three-block network of width $2d+5$, with weights independent of n, whose output is*

$$\widehat{f}_\omega(q) = \frac{1}{n} \sum_{i=1}^{n} y_i K(x_i, q), \tag{6}$$

$$K(x, q) = 1 + x^\top q + \frac{1}{2}\big((x^\top q)^2 - \|x\|^2 - \|q\|^2 + d\big). \tag{7}$$

This identity holds for arbitrary real inputs and labels. Under the task distribution its relative risk is

$$\mathcal{R}_n = \frac{\alpha_d}{n}, \qquad \alpha_d = N + 4d + 7 + \frac{2d-8}{V} = N + O(d). \tag{8}$$

3. *Multiplying the last attention matrix Q_3 by $n/(n+\alpha_d)$ gives, at the same width and depth,*

$$\mathcal{R}_n = \frac{\alpha_d}{n+\alpha_d}. \tag{9}$$

The second part has a direct statistical interpretation: the network averages label-weighted kernel evaluations, and its finite-context variance decreases as $1/n$. Because α_d is of order N, a context of order d^2/ε suffices for relative risk ε. The third part shrinks the same prediction toward zero to balance estimation variance against bias; it adds no block or feature coordinate.

Corollary 3.2 (Fixed-error block threshold)**.** *Fix $\varepsilon \in (0,1)$. There is a universal constant $C_0 > 0$ such that, uniformly for $n \geq C_0 N/\varepsilon$, as $d \to \infty$,*

$$W^\star(d, L, n, \varepsilon) = \begin{cases} \Theta(d^2), & L = 1, 2, \\ \Theta(d), & L \geq 3. \end{cases}$$

The comparison constants may depend on ε but not on n or L.

This identifies the minimum number of blocks required for linear width, with width measured up to constant factors. The context-length condition ensures that the finite-context variance is small enough for the architectural separation to be visible. Appendix D gives nonasymptotic bounds; identity blocks extend the construction to every $L \geq 3$.

3.1 How context aggregation saves feature storage

The predictor needs the context matrix only through its action on the query. Linear attention computes this product directly from the examples.

For the construction, let $u(x) = (1, x)$. The label-weighted empirical matrix A and the vector computed by attention in Figure 1 are

$$A = \frac{1}{n} \sum_i y_i u(x_i) u(x_i)^\top, \tag{10}$$

$$Au(q) = \frac{1}{n} \sum_i y_i u(x_i) \big(u(x_i)^\top u(q)\big).$$

The query stores only $Au(q)$ and the preserved $u(q)$, each with $d+1$ coordinates, rather than the $\Theta(d^2)$ entries of A. Their inner product gives $u(q)^\top Au(q)$. The full construction uses $2d+5$ feature coordinates per token.

The order of these operations explains why the construction uses three blocks. The vector $Au(q)$ becomes available only after block 2's feed-forward map, so its multiplication with the query occurs in block 3. Section 4 rules out alternative two-block computations at subquadratic width.

Remark 3.3 (Why label access changes the block count). The label-access restriction is essential to the exact block threshold. If feed-forward maps are allowed to read the label coordinate while leaving it unchanged, two blocks of width $3d+5$ implement the same kernel estimator. Appendix B.1 gives the full construction. In the protected-label setting, copying labels into feature coordinates takes a block before the matrix–vector product and inner product.

4 Why two blocks require quadratic width

After one block, a query can already contain label-dependent information. Its second feed-forward map can multiply that information by other query coordinates. Consequently, the relevant restriction is not the number of fixed input features: we must also account for how the task changes the features used for prediction.

The useful distinction is between choosing a space of query functions and choosing coefficients inside that space. A two-block network uses at most w linear measurements of the coefficient vector θ to choose a subspace spanned by at most $2w + 2$ functions. Further task dependence changes the coefficients, but does not enlarge the space. The following lemma makes this restriction precise.

Lemma 4.1 (Two-block query space)**.** *Fix a two-block network and context inputs X. There is a rank-$\leq w$ linear measurement $g = C_X\theta$ such that the network's output, as a function of q, belongs to a measurable subspace $\mathcal{V}_{X,g} \subset L^2(\mathcal{N}(0, I_d))$ of dimension at most $2w + 2$.*

Proof idea. Let $v(x)$ denote the w features after the first feed-forward map. The label-access restriction makes v independent of the task. For fixed context inputs, the first attention depends on the labels through the vector $g = n^{-1}\sum_i v(x_i)y_i$ and the scalar $h = n^{-1}\sum_i y_i^2$. The vector g contains at most w linear measurements of θ. Once it is fixed, the first query state has the form $z_1(q) = \nu(q) + hpa(q)$, where p is a fixed vector and the coordinates of ν and the function a lie in the span of $v_1, \ldots, v_w$.

Only the scalar h remains task dependent in this expression. Applying the second bilinear map produces $w + 1$ functions that do not involve h, together with terms in the span of $a, av_1, \ldots, av_w$. The quadratic terms in h add no new spanning function because a already lies in the span of the v_j. This gives a subspace of dimension at most $2w + 2$, regardless of the value of h.

Finally, the updated context is independent of the query. The second attention therefore takes a linear combination of these query functions, with coefficients that may depend on the entire context and task. The output stays in the same space. Appendix H.2 gives the complete expansion and measurability argument. □

The distinction matters even when the context is long enough to determine the entire task. Later attention coefficients may then depend on all labels, but the functions of the query that they multiply still come from the subspace selected through g. Knowing more coefficients cannot create a new query function outside that subspace. Thus the restriction in Lemma 4.1 is on the computation available at the query, not on how accurately the context identifies θ.

To turn this restriction into prediction error, let $G = \mathbb{E}\phi(q)\phi(q)^\top$ and let $\lambda_1 \geq \cdots \geq \lambda_N > 0$ be its eigenvalues. Their magnitudes measure how much task variance lies in the corresponding feature directions. Define the spectral tail $\tau_m(G) = \sum_{j>m} \lambda_j$, with $\tau_m = 0$ for $m \geq N$; it is the variance left after the largest m directions are removed.

For a Gaussian task, observing r linear directions and then predicting in a b-dimensional space determined by those observations leaves error at least $\tau_{r+b}(G)$ (Lemma H.1). The lower bound continues to hold under the stronger relaxation in which the prediction coefficients may depend on the full vector θ. Conditioning on the measurements leaves independent Gaussian variation in the unobserved directions. A b-dimensional subspace can account for at most b further directions; arbitrary coefficients do not change that dimension. The operator proof is in Appendix H.1.

Combining this fact with Lemma 4.1, conditional on the context inputs, gives

$$V\mathcal{R}_n(\omega) \geq \tau_{3w+2}(G). \tag{11}$$

For one block the final query label is in the fixed span of v, giving the stronger bound $V\mathcal{R}_n(\omega) \geq \tau_w(G)$. Under the raw-monomial coefficient prior, exactly one eigenvalue of G is below one, so $\tau_m(G) \geq [N - m - 1]_+$. Since $\operatorname{tr} G = N + 2d$, dividing by task variance proves (5). Appendix A gives the spectrum and explains why its nonuniform weights must be retained.

5 Computing quadratic predictions with linear width

We now specify how the network realizes the kernel estimator of Theorem 3.1, using the context matrix A from (10).

5.1 Three stages of computation

Use feature coordinates $(c, x_1, \ldots, x_d, t_0, \ldots, t_d, \rho, b, \zeta)$ and one protected label coordinate. Here $c = 1$ and the x coordinates preserve the input; the t coordinates hold a working vector; $\rho = \|x\|^2$

stores the squared input norm, b holds a label copy, and $\zeta = y\|x\|^2$ supplies a label-weighted norm. These roles describe the coordinates after their respective updates. This accounts for all $2d + 5$ feature coordinates. Appendix B specifies every matrix entry and simultaneous coordinate update.

The first block makes labels available to later feed-forward multiplication. Its attention copies each context label into b, while b remains zero at the query. Later feed-forward maps can then use this copy without reading the protected coordinate. The first feed-forward map also stores the coordinate squares needed for the scalar corrections.

In the second block, the feed-forward map forms the value vector $y_i u(x_i)$ at each context token. Attention weights it by $u(x_i)^\top u(q)$, sums over examples, and divides by n. Including the residual connection, the resulting working vectors are

$$t_i = y_i u(x_i) + Au(x_i) \quad (i \le n), \qquad t_q = Au(q), \tag{12}$$

where t_i denotes the entire t-coordinate group of token i. The context tokens are updated as well as the query. The construction preserves the scalar coordinates needed for the final readout, so these context updates do not interfere with the calculation.

The third feed-forward map multiplies each coordinate of $Au(q)$ by the corresponding preserved query coordinate. Their sum is $u(q)^\top Au(q)$. With the preserved context moments $m_0 = n^{-1}\sum_i y_i$ and $m_2 = n^{-1}\sum_i y_i\|x_i\|^2$, the final attention writes the negative of the predictor to the query-label coordinate. Thus

$$\widehat{f}_\omega(q) = \frac{1}{2}u(q)^\top Au(q) + \frac{d+1}{2}m_0 - \frac{1}{2}m_0\|q\|^2 - \frac{1}{2}m_2. \tag{13}$$

The first term supplies the squared inner products. The remaining terms center the square features, giving exactly the kernel in (7). This readout uses preserved context coordinates, not the updated vectors t_i. Substituting $u(q)^\top Au(q) = n^{-1}\sum_i y_i(1 + x_i^\top q)^2$ proves the estimator identity for arbitrary real inputs and labels.

5.2 Why the prediction error decreases with context length

The kernel is chosen to reproduce every quadratic function under Gaussian inputs. The functions 1, x_j, $(x_j^2 - 1)/\sqrt{2}$, and $x_j x_k$ for $j < k$ form an orthonormal basis of that space. Summing products of matching basis functions gives $K(x, q)$ and, for a fresh $X \sim \mathcal{N}(0, I_d)$ and any quadratic f,

$$\mathbb{E}_X f(X)K(X, q) = f(q), \qquad \mathbb{E}_q K(X, q)^2 = K(X, X). \tag{14}$$

The first identity says that one label-weighted kernel term is an unbiased estimate of the target at the query. Averaging n independent context examples therefore divides its variance by n. Averaging also over the Gaussian task coefficients gives

$$V\mathcal{R}_n = \frac{1}{n}\left(\mathbb{E}_X[\|\phi(X)\|^2 K(X, X)] - V\right). \tag{15}$$

Appendix C evaluates the Gaussian moments through order eight to obtain α_d in (8). The required context length scales with the quadratic feature dimension $N = \Theta(d^2)$, even though network width is $O(d)$.

For shorter contexts, the unbiased estimate has substantial finite-context variance. Multiplying it by a scalar a gives relative risk $(1 - a)^2 + a^2\alpha_d/n$, separating the introduced bias from the reduced variance. Its minimum occurs at the shrinkage factor $a = n/(n + \alpha_d)$ and gives (9). Only this factor depends on context length; the underlying three-block circuit is otherwise independent of n.

6 Statistical limits at linear width

The width threshold separates shallow and three-block networks when enough examples are available. To see the effect of depth at matched resources, fix the linear width of the construction and let context length scale with the quadratic feature dimension. A third block then gives a nonvanishing reduction in relative risk as dimension grows.

Corollary 6.1 (Matched-resource separation). *Fix $\gamma > 0$, let $n = \lfloor \gamma N \rfloor$, and set $w = 2d + 5$. Then*

$$\inf_{\omega:\, L \in \{1,2\}} \mathcal{R}_n(\omega) \longrightarrow 1, \qquad \limsup_{d \to \infty} \inf_{\omega:\, L=3} \mathcal{R}_n(\omega) \leq \frac{1}{1+\gamma}.$$

The shrinkage-scaled circuit in Theorem 3.1 has risk converging to $1/(1+\gamma)$.

The shallow limit follows from (5) and the zero predictor; the constructed limit follows from (9) and $\alpha_d/N \to 1$. Thus the risk reduction comes from an additional stage of computation, with both width and context length matched.

To assess how much error is due to limited data, let $\mathcal{R}_{\mathrm{Bayes}}(d, n)$ be the minimum relative risk over all measurable predictors based on the same context and query. This is the Bayes-optimal predictor with no architectural restriction. Each noiseless label is a linear measurement of the N task coefficients, and the n measurements have rank $\min(n, N)$ almost surely. The remaining task variance gives, uniformly in n,

$$\mathcal{R}_{\mathrm{Bayes}}(d, n) = (1 - n/N)_+ + O(d^{-1}), \qquad \mathcal{R}_{\mathrm{Bayes}}(d, n) = 0 \quad \text{if } n \geq N. \tag{16}$$

To see why this is the right sample benchmark, fix the context inputs X and form the matrix Φ_X with rows $\phi(x_i)^\top$. The labels reveal the component of θ in the row space of Φ_X and leave independent Gaussian uncertainty in its orthogonal complement. If Π_X projects onto that complement, then $\mathcal{R}_{\mathrm{Bayes}}(d, n) = V^{-1}\mathbb{E}_X \operatorname{tr}(G\Pi_X)$. The evaluation vectors span the full quadratic feature space: otherwise a nonzero quadratic polynomial would vanish at every input. Consequently, some evaluation minor is a nonzero polynomial in the input coordinates. Gaussian inputs have a density, and the zero set of that minor has measure zero, giving $\operatorname{rank} \Phi_X = \min(n, N)$ almost surely. Were G the identity, the risk would be exactly $(1 - n/N)_+$. For the raw-monomial prior, the spectrum of G differs from the identity by total mass only $O(d)$, compared with $V = \Theta(d^2)$; this gives the $O(d^{-1})$ correction in (16). Appendix E supplies the posterior calculation and an explicit finite-dimensional bound.

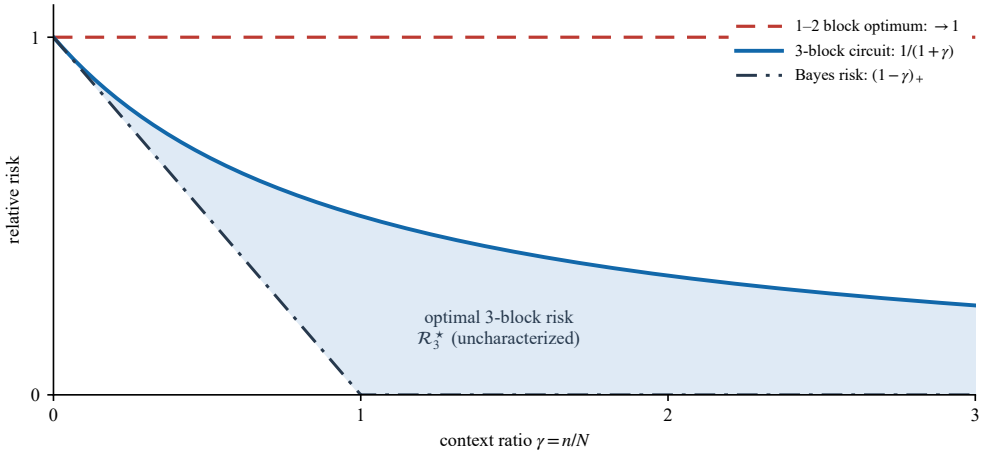

Figure 2: **Architectural and statistical limits at linear width.** As $d \to \infty$ with width $2d + 5$ and $n/N \to \gamma > 0$, the optimal one- and two-block relative risks tend to one (red dashed), the shrinkage-scaled three-block circuit attains $1/(1+\gamma)$ (blue solid), and Bayes risk tends to $(1-\gamma)_+$ (dark dash-dotted). The shaded region lies between the Bayes lower bound and the achievable three-block upper bound; $\mathcal{R}_3^\star$ denotes the optimal three-block risk. All curves are analytical.

Figure 2 separates three quantities: the shallow optimum, the risk achieved by the three-block construction, and Bayes risk. The two lower bounds limit different resources. Finite examples leave at least $\tau_n(G)/V$ risk for any predictor, while the two-block query-space restriction leaves at least $\tau_{3w+2}(G)/V$ risk. Since the spectral tail decreases with its index, a two-block network obeys $V\mathcal{R}_n(\omega) \geq \tau_{\min\{n, 3w+2\}}(G)$. Even when $n \geq N$ removes the sample limitation, subquadratic width keeps the shallow architectural bound near one. Conversely, increasing width cannot replace missing examples.

Define $\mathcal{R}_3^\star(d,n) = \inf_\omega \mathcal{R}_n(\omega)$ for exactly three blocks of width $2d+5$. In the regime $d \ll n \ll N$, the construction captures the available Bayes risk reduction to first order:

$$1 - \mathcal{R}_3^\star(d,n) \sim \frac{n}{N} \sim 1 - \mathcal{R}_{\mathrm{Bayes}}(d,n). \tag{17}$$

Here $n \gg d$ makes the $O(1/d)$ approximation error in (16) negligible relative to n/N. This gives a statistical interpretation of the construction beyond the width separation.

At proportional context length, $n/N \to \gamma \in (0,\infty)$, the established bounds instead leave a gap:

$$(1-\gamma)_+ \le \liminf_{d\to\infty} \mathcal{R}_3^\star(d,n) \le \limsup_{d\to\infty} \mathcal{R}_3^\star(d,n) \le \frac{1}{1+\gamma}. \tag{18}$$

For $n \ge N$, the Bayes-optimal predictor recovers the task by solving the interpolation system. The constructed network averages label-weighted kernel terms and therefore retains finite-context estimation variance. The proportional-context gap is therefore visible even after the task becomes identifiable: for $\gamma \ge 1$, Bayes risk tends to zero, but the risk of this kernel-averaging circuit tends to $1/(1+\gamma)$. This comparison identifies unused statistical information in the particular construction; it does not show that every three-block linear-width network has the same residual error. For $0 < \gamma < 1$, the gap similarly separates what is guaranteed by the sample-limited Bayes bound from what our explicit circuit attains. In both regimes the shallow optimum remains at one, so the achieved advantage of the third block does not depend on resolving the optimal three-block risk.

7 Conclusion

We established a sharp depth–width separation for quadratic in-context learning in the protected-label bilinear Transformer. For any fixed target relative risk below one and a sufficiently long context, one- and two-block networks require $\Theta(d^2)$ width, whereas three blocks require only $\Theta(d)$ width. Thus, within this architecture, one additional stage of computation reduces the required width from quadratic to linear in d.

Beyond the threshold itself, the key insight is the computation that enables this width reduction. A quadratic task has $N = \Theta(d^2)$ coefficients, yet prediction at a single query does not require the network to explicitly store a quadratic-dimensional representation. The context implicitly defines the label-weighted operator

$$A = \frac{1}{n}\sum_{i=1}^{n} y_i u(x_i) u(x_i)^\top,$$

but the three-block construction stores neither A nor the full quadratic feature vector. Instead, attention directly computes the query-dependent vector $Au(q)$, which has only $O(d)$ coordinates. A later bilinear operation combines this vector with the preserved query to produce the prediction. In this sense, depth replaces simultaneous feature storage with sequential, context-mediated computation.

The exact two-to-three-block threshold depends on the protected-label access rule. If feed-forward maps can read labels directly, the same estimator can be implemented with two blocks and linear width (Appendix B.1). Therefore, our result characterizes how label-dependent information moves through this architecture; it does not claim a universal two-to-three-block threshold for unrestricted Transformers. The result is also not specific to the isotropic raw-monomial prior. The same separation extends to Gaussian coefficient distributions with uniformly well-conditioned covariance in an orthonormal quadratic basis (Corollary F.1).

More generally, our result shows that prediction can require much less intermediate state than a full task representation. In this setting, the model achieves this reduction by applying a task-dependent operator to the query without storing the operator itself. This provides a concrete example of how additional depth can trade sequential computation for token width in nonlinear in-context learning.

AI USE STATEMENT

OpenAI GPT-6 was used for language polishing, manuscript organization, and assistance with literature search. During the research process, it was also used to assist in constructing the explicit two-block label-readable circuit presented in Appendix B.1, which helped isolate the role of the protected-label restriction in the main depth–width separation. Starting from a proof strategy developed by the authors, GPT-6 was further used to help work out technical details of the Gaussian-covariance extension in Corollary F.1, including transferring the shallow spectral-tail lower bound to a well-conditioned coefficient covariance and deriving the corresponding variance bound for the three-block kernel construction. GPT-6 was also used to check mathematical arguments for logical consistency, algebraic manipulations, and boundary cases, and to assist in implementing verification scripts for the exact-moment calculations and end-to-end circuit identities, as well as in figure preparation. All final mathematical arguments, proofs, constructions, code, computational results, figures, and conclusions were independently reviewed and verified by the authors. The authors take full responsibility for all claims and results reported in this work.

REPRODUCIBILITY STATEMENT

Section 2 specifies the task distribution, architecture, parameter quantifiers, and risk. Sections 4–5 give the central proof mechanisms, with complete lower-bound details in Appendix H. Appendices A–F provide the spectrum, construction matrices, exact moments, and proofs of the statistical consequences. Appendix G specifies the executable circuit and exact-moment checks included in the source archive. All stated bounds concern fixed network parameters and the specified task distribution.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45614–45650, 2023. URL https://proceedings.neurips.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 *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Blake Bordelon, Mary Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/24d36eee157559e0d2549455fba28f6a-Abstract-Conference.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 *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8002–8037, 2024. URL https://proceedings.mlr.press/v235/cheng24a.html.

Frank Cole, Yuxuan Zhao, Yulong Lu, and Tianhao Zhang. In-context learning of linear dynamical systems with transformers: Approximation bounds and depth-separation. In *Advances in*

Neural Information Processing Systems, volume 38, pp. 187641–187679, 2025. doi: 10.52202/085713-5641. URL https://papers.nips.cc/paper_files/paper/2025/hash/f7ae4fe91d96f50abc2211f09b6a7e49-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 *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Alexander Hsu, Zhaiming Shen, Wenjing Liao, and Rongjie Lai. Understanding in-context learning for nonlinear regression with transformers: Attention as featurizer. arXiv:2605.05176, 2026. URL https://arxiv.org/abs/2605.05176.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24527–24561, 2024. URL https://proceedings.mlr.press/v235/kim24af.html.

Arvind Mahankali, Tatsunori B. 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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/87f476af4053961667c2c08e9f4b850e-Abstract-Conference.html.

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46554–46585, 2025. URL https://proceedings.mlr.press/v267/nishikawa25a.html.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77316–77365, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8d6c356c82252032a20c749dda36f2d8-Abstract-Conference.html.

Parthe Pandit, Zhichao Wang, and Yizhe Zhu. Universality of kernel random matrices and kernel regression in the quadratic regime. *Journal of Machine Learning Research*, 26(224):1–73, 2025. URL https://jmlr.org/papers/v26/24-1533.html.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the Bayesian prism. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d81cd83e7f6748af351485d73f305483-Abstract-Conference.html.

Zhaiming Shen, Alexander Hsu, Rongjie Lai, and Wenjing Liao. Understanding in-context learning on structured manifolds: Bridging attention to kernel methods. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/45e9827fcd83cfb39285669e13b2a30d-Abstract-Conference.html.

Haoyuan Sun, Ali Jadbabaie, and Navid Azizan. On the role of transformer feed-forward layers in nonlinear in-context learning. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–3, 2026. URL https://proceedings.mlr.press/v313/sun26a.html. Full version: arXiv:2501.18187v3.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Mingsong Yan, Dongyang Li, Charles Kulick, and Sui Tang. Transformers can implement preconditioned Richardson iteration for in-context Gaussian kernel regression. arXiv:2605.08475, 2026. URL https://arxiv.org/abs/2605.08475.

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

A THE RAW QUADRATIC MOMENT SPECTRUM

Order the raw monomials as 1, the d linear terms, the d squares, and the $\binom{d}{2}$ cross terms. Odd Gaussian moments show that the linear and cross terms are orthogonal to all other listed terms and have squared norm one. Thus the eigenvalue one has multiplicity $d + \binom{d}{2} = d(d+1)/2$. The remaining block of G, indexed by $1, x_1^2, \dots, x_d^2$, is

$$\begin{pmatrix} 1 & \mathbf{1}^\top \\ \mathbf{1} & 2I_d + \mathbf{1}\mathbf{1}^\top \end{pmatrix}.$$

On square coefficients orthogonal to $\mathbf{1}$, it has eigenvalue two, with multiplicity $d-1$. On the coefficient directions for 1 and $d^{-1/2}\sum_j x_j^2$, it is

$$\begin{pmatrix} 1 & \sqrt{d} \\ \sqrt{d} & d+2 \end{pmatrix}.$$

Its two eigenvalues are

$$\lambda_\pm = \frac{d+3 \pm \sqrt{d^2+6d+1}}{2}. \tag{19}$$

Their product is two, $\lambda_+ > 2$, and $0 < \lambda_- < 1$. These eigenvalues account for all N directions, and their sum gives $\operatorname{tr} G = N + 2d = V$. Since exactly one eigenvalue is below one, $\tau_m(G) \geq [N - m - 1]_+$. The bound remains valid even though $\lambda_- = \Theta(d^{-1})$; replacing all eigenvalues by one would change the task distribution.

Since $N = \Theta(d^2)$ and $V = N + 2d$, the right-hand side of (5) tends to one whenever $w_d = o(d^2)$, uniformly over the context length n_d. Choosing every attention map as the identity leaves the query label zero and gives the matching upper bound of one.

A.1 A ONE-DIMENSIONAL EXAMPLE OF THE PRIOR DEPENDENCE

When $d = 1$, the raw features are $(1, x, x^2)$ and G has eigenvalues $2+\sqrt{2}, 1, 2-\sqrt{2}$. A one-block width-two network can use $v(x) = (1 + (1+\sqrt{2})x^2, x)$. The first coordinate is a residual bilinear update of the input constant allowed by (2); the second coordinate is unchanged. Attention with feature preconditioner $-(\mathbb{E}vv^\top)^{-1}$ realizes the projection kernel of this two-dimensional space, as in Appendix D. The coefficient vector $(1, 1+\sqrt{2})$ on $1, x^2$ is the top eigenvector of their moment block. As $n \to \infty$, its unnormalized mean squared error therefore tends to $2 - \sqrt{2} < 1 = N - w$. The example shows why dimension counting under the raw coefficient prior must retain spectral weights; it is consistent with the bound $\tau_w(G)$.

B COMPLETE THREE-BLOCK MATRICES

We give the matrices that implement the construction under (2)–(3). The feature order is $(c, x_1, \dots, x_d, t_0, \dots, t_d, \rho, b, \zeta)$, followed by y. For a feature coordinate v, write $(W_0)_{v,:}$ and $(W_1)_{v,:}$ as linear forms in these feature coordinates. All unlisted rows are zero; both extended matrices have zero label input column and output row. A row pair (ℓ, m) implements $v \leftarrow v + \ell m$. The constant coordinate c stays one throughout.

Block	Output coordinate	Row of W_0	Row of W_1
1	$t_j, 1 \le j \le d$	x_j	x_j
2	t_0	c	b
2	$t_j, 1 \le j \le d$	x_j	$b - x_j$
2	ρ	c	$\sum_{j=1}^{d} t_j$
2	ζ	b	$\sum_{j=1}^{d} t_j$
3	$t_j, 1 \le j \le d$	t_j	$x_j - c$

Every right-hand side uses the values at the start of its feed-forward sublayer.

In the second feed-forward sublayer, the rows in the table read the pre-update values $t_j = x_j^2$, giving $t = b\,u(x)$, $\rho = \|x\|^2$, and $\zeta = b\|x\|^2$. These updates preserve the original input coordinates.

The attention matrices are

$$P_1 = e_b e_c^\top, \qquad Q_1 = e_c e_y^\top,$$

$$P_2 = \sum_{j=0}^{d} e_{t_j} e_{t_j}^\top, \qquad Q_2 = e_c e_c^\top + \sum_{j=1}^{d} e_{x_j} e_{x_j}^\top,$$

$$P_3 = e_y e_c^\top, \qquad Q_3 = -\frac{1}{2} e_c \Big(\sum_{j=0}^{d} e_{t_j} \Big)^\top - \frac{d+1}{2} e_b e_c^\top + \frac{1}{2} e_b e_\rho^\top + \frac{1}{2} e_\zeta e_c^\top.$$

The first two attention maps do not change y. Consequently the query label entering the final attention is zero. For any source context token, multiplication by P_3 places its constant coordinate in the label position. The attention score against the query is

$$-\frac{1}{2} c_i \sum_{j=0}^{d} t_{q,j} - \frac{d+1}{2} b_i c_q + \frac{1}{2} b_i \rho_q + \frac{1}{2} \zeta_i c_q.$$

Averaging this score gives the negative of (13), because $c_i = c_q = 1$, $b_i = y_i$, $\zeta_i = y_i \|x_i\|^2$, and $\rho_q = \|q\|^2$. By the prediction convention in (4), the circuit implements (6). The final score does not read the context t_i coordinates. The same matrices also update every context token.

All parameters are task independent and the construction uses no additional tokens. The $1/n$ factor is the normalization already present in the architecture. Scaling Q_3 scales the output, since the query label was zero before that sublayer. These identities verify both the arbitrary-label output formula and the shrinkage rescaling in Theorem 3.1.

B.1 Two blocks when feed-forward sublayers can read labels

We modify only (2) by allowing nonzero label input columns in the two bilinear matrices while keeping their label output rows zero. The attention mask, block order, initial tokens, and task-independent parameters remain unchanged. The following construction has $3d + 5$ feature coordinates,

$$(c, x_1, \ldots, x_d, t_0, \ldots, t_d, s_1, \ldots, s_d, \rho, b, \zeta),$$

followed by y. All feature coordinates except $c = 1$ and x initially vanish. All assignments within a feed-forward sublayer are simultaneous residual bilinear assignments, and all unlisted rows are zero.

In the first feed-forward sublayer, set

$$t_0 \leftarrow t_0 + cy, \quad t_j \leftarrow t_j + x_j y, \quad s_j \leftarrow s_j + x_j^2, \quad b \leftarrow b + cy \qquad (1 \le j \le d).$$

Use attention matrices

$$P_1 = \sum_{j=0}^{d} e_{t_j} e_{t_j}^\top, \qquad Q_1 = e_c e_c^\top + \sum_{j=1}^{d} e_{x_j} e_{x_j}^\top.$$

With the context matrix A from (10), the first attention produces $t_q = Au(q)$ and $t_i = y_i u(x_i) + Au(x_i)$. The s_j, b, c, x, y coordinates are unchanged by attention, and the query label remains zero.

In the second feed-forward sublayer, set

$$t_j \leftarrow t_j + t_j(x_j - c), \quad \rho \leftarrow \rho + c\sum_{j=1}^{d} s_j, \quad \zeta \leftarrow \zeta + y\sum_{j=1}^{d} s_j.$$

Thus $\sum_{j=0}^{d} t_{q,j} = u(q)^\top A u(q)$, $\rho_q = \|q\|^2$, $b_i = y_i$, and $\zeta_i = y_i\|x_i\|^2$. Use

$$P_2 = e_y e_c^\top, \qquad Q_2 = -\frac{1}{2}e_c\Big(\sum_{j=0}^{d} e_{t_j}\Big)^\top - \frac{d+1}{2}e_b e_c^\top + \frac{1}{2}e_b e_\rho^\top + \frac{1}{2}e_\zeta e_c^\top.$$

The final attention reads only the preserved context coordinates c, b, ζ and writes the negative of (13), hence implements (6), for arbitrary real inputs and labels. These identities include all context updates. Under (1), the resulting estimator has relative risk α_d/n, and scaling Q_2 gives (9). The minimum block count and width constant in this modified architecture remain undetermined.

C Projection kernels and exact Gaussian moments

For $X \sim \mathcal{N}(0, I_d)$, the functions $1, x_j, (x_j^2 - 1)/\sqrt{2}, x_j x_k$ for $j < k$ form an orthonormal basis for the quadratic polynomial space. Orthonormality follows from independence, vanishing odd moments, and $\mathbb{E}X_j^4 = 3$. They have the same span as the raw monomials, so their projection kernel is $\phi(x)^\top G^{-1}\phi(q)$. Summing the products gives

$$\begin{aligned} K(x,q) &= 1 + \sum_j x_j q_j + \sum_{j<k} x_j x_k q_j q_k + \frac{1}{2}\sum_j (x_j^2 - 1)(q_j^2 - 1) \\ &= 1 + x^\top q + \frac{1}{2}\big((x^\top q)^2 - \|x\|^2 - \|q\|^2 + d\big). \end{aligned}$$

Orthonormality proves both identities in (14). For fixed θ, q, the random variables $f_\theta(x_i)K(x_i, q)$ are independent and have mean $f_\theta(q)$. Integrating their conditional variance over q and then θ gives

$$\begin{aligned} nV\mathcal{R}_n &= \mathbb{E}_{\theta,X,q} f_\theta(X)^2 K(X,q)^2 - \mathbb{E}_{\theta,q} f_\theta(q)^2 \\ &= \mathbb{E}_X \|\phi(X)\|^2 K(X,X) - V. \end{aligned}$$

The last equality uses $\mathbb{E}_\theta f_\theta(X)^2 = \|\phi(X)\|^2$ in the raw basis, not independence of the orthonormal-basis coefficients.

Write $\rho = \|X\|^2$ and $s_4 = \sum_j X_j^4$. The diagonal kernel and raw-feature norm are

$$K(X,X) = 1 + d/2 + \rho^2/2, \qquad \|\phi(X)\|^2 = 1 + \rho + (\rho^2 + s_4)/2.$$

To average s_4 at fixed radius, write $X = \sqrt{\rho}\,U$ with U uniform on the sphere and independent of ρ. Since $\mathbb{E}\rho^2 = d(d+2)$ and $\mathbb{E}X_j^4 = 3$, one has $\mathbb{E}U_j^4 = 3/[d(d+2)]$. Hence $\mathbb{E}[s_4 \mid \rho] = 3\rho^2/(d+2)$. The radial moments are $\mathbb{E}\rho^k = \prod_{j=0}^{k-1}(d+2j)$. Substituting these conditional and radial moments into the preceding risk identity yields

$$\begin{aligned} C_d &= \frac{d}{2}V + \frac{1}{2}d(d+2) + \frac{1}{2}d(d+2)(d+4) + \frac{1}{4}d(d+5)(d+4)(d+6) \\ &= \frac{d(d^3 + 18d^2 + 95d + 142)}{4}. \end{aligned}$$

Dividing by V gives $\alpha_d = C_d/V = N + 4d + 7 + (2d-8)/V = N + O(d)$. In particular, $\alpha_d = \Theta(N)$ uniformly for $d \geq 1$.

Finally, let $\widehat{f}_\omega(q)$ denote the kernel prediction in (6). Conditional unbiasedness implies $\mathbb{E}[(\widehat{f}_\omega(q) - f_\theta(q))f_\theta(q)] = 0$. Thus

$$\frac{1}{V}\mathbb{E}(a\widehat{f}_\omega(q) - f_\theta(q))^2 = (1-a)^2 + \frac{a^2\alpha_d}{n}.$$

Minimizing this quadratic proves the scaled-risk formula.

D PARTIAL FEATURES AND FINITE-CONTEXT WIDTH

Corollary 3.2 gives the width order at fixed target risk. The next theorem also describes the dependence on $1-\varepsilon$ as the target approaches the zero-predictor risk.

Theorem D.1 (Nonasymptotic width dependence). *There are universal constants $c, C, C_0 > 0$ such that, for every $d \geq 1$, $0 < \varepsilon < 1$, and $n \geq C_0 N/\varepsilon$,*

$$c\big[d + (1-\varepsilon)N\big] \leq W^\star(d, L, n, \varepsilon) \leq C\big[d + (1-\varepsilon)N\big], \qquad L = 1, 2,$$
$$d + 1 \leq W^\star(d, L, n, \varepsilon) \leq 2d + 5, \qquad L \geq 3.$$

D.1 PARTIAL-FEATURE CONSTRUCTION AND VARIANCE CONTROL

Choose s square monomials and m cross monomials. Let $v(x)$ contain the constant, all d linear features, and these selected monomials, so its width is $w = d + 1 + s + m$. A single bilinear feed-forward sublayer writes the selected quadratic features while preserving the others. The matrix $H = \mathbb{E}v(X)v(X)^\top$ is positive definite because the selected polynomials are linearly independent. Attention with

$$P = e_y e_y^\top, \qquad Q = \begin{pmatrix} -H^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

writes $-n^{-1}\sum_i y_i K_S(x_i, q)$ to the query label, where $K_S(x, q) = v(x)^\top H^{-1} v(q)$ is the projection kernel of the selected space S. All parameters are task independent; the same attention map also updates context labels, which no later nonidentity block reads.

Let P_S be $L^2(\mathcal{N}(0, I_d))$ projection onto S. In the full orthonormal quadratic basis, the constant coefficient of f_θ is $\theta_0 + \sum_j \theta_{jj}$ and has variance $d+1$. Each linear and cross coefficient has variance one, and each centered square coefficient has variance two. Thus the captured task variance is

$$e = \mathbb{E}_\theta \|P_S f_\theta\|_{L^2}^2 = 2d + 1 + 2s + m \geq w.$$

The estimator $\widehat{f}_S = n^{-1}\sum_i f_\theta(x_i) K_S(x_i, q)$ has conditional mean $P_S f_\theta(q)$. Define its finite-context variance coefficient by

$$c_S = \mathbb{E}_X \|\phi(X)\|^2 K_S(X, X) - e \geq 0.$$

Orthogonality and independence of the examples give $\mathbb{E}\langle \widehat{f}_S, f_\theta \rangle_{L^2} = e$ and $\mathbb{E}\|\widehat{f}_S\|_{L^2}^2 = e + c_S/n$. Rescaling the last attention matrix by $a = e/(e + c_S/n)$ therefore gives

$$\frac{1}{V}\mathbb{E}(a\widehat{f}_S(q) - f_\theta(q))^2 = 1 - \frac{e}{V}\frac{n}{n + c_S/e}. \tag{20}$$

Lemma D.2 (Uniform partial-feature variance). *For every selected space S above, $c_S/e \leq C_{\mathrm{var}} N$ for a universal constant C_{var}.*

Proof. Write $K_S(X, X) = \sum_{\psi \in \mathcal{B}_S} \psi(X)^2$ for the selected orthonormal Gaussian quadratic basis $\mathcal{B}_S$. Each raw monomial ϕ_j and each $\psi \in \mathcal{B}_S$ has a fourth moment bounded uniformly in d by Gaussian moments through order eight. Cauchy–Schwarz therefore gives $\mathbb{E}[\phi_j(X)^2 \psi(X)^2] \leq C_{\mathrm{var}}$ for a universal constant. Summing over the N raw monomials and the w selected basis functions yields $c_S \leq C_{\mathrm{var}} N w$. The conclusion follows from $e \geq w$. □

D.2 PROOF OF THEOREM D.1

If a shallow network of width w has relative risk at most ε, (5) and $w \geq d + 1$ imply

$$(1-\varepsilon)V \leq 3w + 2d + 3 \leq 6w.$$

Since $V \geq N$, combining this bound with $w \geq d + 1$ gives $w = \Omega(d + (1-\varepsilon)N)$ with a universal constant.

For the shallow upper bound, retain all squares and choose

$$w = \min\{N, \max\{2d + 1, \lceil 2(1-\varepsilon)N \rceil\}\}, \qquad s = d, \quad m = w - 2d - 1.$$

This is a valid feature selection for every $d \geq 1$, with $w = O(d + (1-\varepsilon)N)$ and $e = w + 2d$. Choose C_0 large enough that Lemma D.2 gives $c_S/(en) \leq \varepsilon/2$. If $w = N$, then $e = V$ and (20) is at most $\varepsilon/2$. Otherwise, $w \geq 2(1-\varepsilon)N$, and

$$e \geq 2(1-\varepsilon)N + 2d \geq (1-\varepsilon)(1+\varepsilon/2)V.$$

The last inequality follows from $1 + \varepsilon/2 \leq 3/2 < 2$ and $(1-\varepsilon)(1+\varepsilon/2) \leq 1$. Substitution in (20) gives risk at most ε. One identity block gives the same upper bound for $L = 2$.

For $L \geq 3$, (8) and $\alpha_d = \Theta(N)$ give a three-block network of width $2d + 5$ and risk at most ε after increasing the universal constant C_0 if necessary. Identity blocks preserve this risk; the input representation requires width at least $d + 1$.

E Bayes risk and risk reduction

For context inputs $X = (x_1, \ldots, x_n)$, let $\Phi_X \in \mathbb{R}^{n \times N}$ have rows $\phi(x_i)^\top$. The posterior covariance of the Gaussian coefficient vector is

$$\Pi_X = I - \Phi_X^\top (\Phi_X \Phi_X^\top)^\dagger \Phi_X, \tag{21}$$

where $\dagger$ denotes the Moore–Penrose pseudoinverse. Averaging the conditional prediction variance gives

$$\mathcal{R}_{\mathrm{Bayes}}(d, n) = \frac{1}{V} \mathbb{E} \operatorname{tr}(G \Pi_X). \tag{22}$$

In particular, every predictor has unnormalized mean squared error at least $\tau_n(G)$. Together with the shallow architectural bound, this gives $V \mathcal{R}_n(\omega) \geq \tau_{\min\{n, 3w+2\}}(G)$ for a two-block network. The argument below establishes the posterior formula, almost-sure rank and trace comparison used in the main text.

Conditional on the context inputs, the labels are the noiseless linear observation $y = \Phi_X \theta$ of a standard Gaussian vector. The task coefficients have posterior mean $\Phi_X^\top (\Phi_X \Phi_X^\top)^\dagger y$ and posterior covariance (21). Conditional on X, y, q, the minimum mean squared prediction error is $\phi(q)^\top \Pi_X \phi(q)$. Averaging over q and X proves (22). For any projection Π_X of rank at least $N - n$, the variational characterization of the smallest eigenvalues gives $\operatorname{tr}(G\Pi_X) \geq \tau_n(G)$. This proves the universal sample lower bound.

To justify the exact rank, note that the evaluation vectors $\phi(x)$ span $\mathbb{R}^N$ as x varies. Otherwise a nonzero coefficient vector would define a polynomial vanishing at every point of $\mathbb{R}^d$, contrary to linear independence of monomials. Therefore there are N points whose evaluation vectors are linearly independent. For $n \leq N$, an $n \times n$ minor of evaluations at some n of these points is nonzero. This minor, as a polynomial in the nd input coordinates, is not identically zero. Its zero set has Lebesgue measure zero and hence Gaussian measure zero. For $n > N$, apply the same argument to the first N points. Thus $\operatorname{rank} \Phi_X = \min(n, N)$ almost surely.

Write $m = (N-n)_+$ and recall $0 < \lambda_- < 1$ from (19). For the positive and negative parts of $G - I$, the spectrum gives $\operatorname{tr}(G-I)_- = 1 - \lambda_-$ and $\operatorname{tr}(G-I)_+ = 2d + 1 - \lambda_-$. Because $0 \preceq \Pi_X \preceq I$ and $\operatorname{tr} \Pi_X = m$ almost surely,

$$\lambda_- - 1 \leq \operatorname{tr}((G-I)\Pi_X) \leq 2d + 1 - \lambda_-.$$

Subtracting m/N from (22) gives

$$\mathcal{R}_{\mathrm{Bayes}}(d, n) - \frac{m}{N} = \frac{\mathbb{E} \operatorname{tr}((G-I)\Pi_X) - 2dm/N}{V}.$$

Since $0 \leq m/N \leq 1$, the absolute value of this difference is at most $(2d+1-\lambda_-)/V < (2d+1)/V$, proving (16). If $n \geq N$, the projection vanishes exactly, so Bayes risk is zero rather than merely $O(d^{-1})$.

For any d, n, the Bayes lower bound and the constructed predictor give

$$\mathcal{R}_{\mathrm{Bayes}}(d, n) \leq \mathcal{R}_3^\star(d, n) \leq \frac{\alpha_d}{n + \alpha_d}.$$

In the regime $d \ll n \ll N$, the Bayes lower bound gives $1 - \mathcal{R}_3^\star(d, n) \leq n/N + O(d^{-1}) \sim n/N$. The construction gives $1 - \mathcal{R}_3^\star(d, n) \geq n/(n+\alpha_d) \sim n/N$, because $\alpha_d = N + O(d)$ and $n/N \to 0$. Combining these two estimates proves (17). When $n/N \to \gamma > 0$, taking lower and upper limits instead proves (18).

F GAUSSIAN COEFFICIENT COVARIANCE EXTENSION

The task distribution in this appendix uses Gaussian coefficients in an orthonormal basis rather than the raw monomial basis. Let $\psi(x)$ list the Gaussian quadratic features from Section 5 and put $f_\beta(x) = \beta^\top \psi(x)$ with $\beta \sim \mathcal{N}(0, \Lambda)$. The task variance is $V_\Lambda = \operatorname{tr} \Lambda$, and $\mathcal{R}_n^\Lambda$ denotes relative risk normalized by V_Λ. The input law and architecture remain unchanged.

Corollary F.1 (Gaussian coefficient covariances)**.** *If $0 < \underline{\lambda} I \preceq \Lambda \preceq \overline{\lambda} I$ and $\kappa = \overline{\lambda}/\underline{\lambda}$, every one- or two-block network satisfies*

$$\mathcal{R}_n^\Lambda \geq \frac{\tau_{3w+2}(\Lambda)}{\operatorname{tr} \Lambda} \geq \left[1 - \frac{\kappa(3w+2)}{N}\right]_+ .$$

The same three-block kernel circuit of width $2d + 5$ satisfies

$$\mathcal{R}_n^\Lambda \leq \frac{\kappa \delta_d}{n}, \qquad \delta_d = N + 4d + 7 - \frac{8}{d+1} = N + O(d).$$

For the isotropic Hermite prior $\Lambda = I$, the lower bound is $[N - 3w - 2]_+/N$ and the circuit risk is exactly δ_d/n.

We use the covariance bounds in Corollary F.1 throughout the proof. Since $\mathbb{E}\psi(q)\psi(q)^\top = I_N$, the task variance is $V_\Lambda = \operatorname{tr} \Lambda$.

For the lower bound, write $\beta = \Lambda^{1/2}\theta$, where $\theta \sim \mathcal{N}(0, I_N)$, and define $T\theta = \theta^\top \Lambda^{1/2}\psi$. Then $T^*T = \Lambda$. For fixed context inputs, the first feature vector is independent of θ, and $g = n^{-1}\sum_i v(x_i) f_\beta(x_i)$ remains a linear observation of θ of rank at most w. Lemma 4.1 therefore supplies the same dimension bound $2w + 2$, and Lemma H.1 gives $V_\Lambda \mathcal{R}_n^\Lambda \geq \tau_{3w+2}(\Lambda)$. The one-block bound $\tau_w(\Lambda)$ is stronger. Since removing m eigenvalues removes at most $m\overline{\lambda}$ energy and $\operatorname{tr} \Lambda \geq \underline{\lambda} N$,

$$\frac{\tau_m(\Lambda)}{\operatorname{tr} \Lambda} \geq \left[1 - \frac{m\overline{\lambda}}{\operatorname{tr} \Lambda}\right]_+ \geq [1 - \kappa m/N]_+ .$$

The circuit implements the conditionally unbiased estimator $\widehat{f}_\omega(q) = n^{-1}\sum_i f_\beta(x_i) K(x_i, q)$ under the prediction convention in (4). Define $\Sigma_K = \mathbb{E}_x[K(x,x)\psi(x)\psi(x)^\top] - I_N$. For every deterministic $b \in \mathbb{R}^N$, put $f_b(x) = b^\top \psi(x)$. Then $b^\top \Sigma_K b$ is the integrated variance of the one-example estimate $f_b(x)K(x,q)$, so $\Sigma_K \succeq 0$. Averaging over β and using the covariance bounds gives

$$nV_\Lambda \mathcal{R}_n^\Lambda = \operatorname{tr}(\Lambda \Sigma_K) \leq \overline{\lambda} \operatorname{tr} \Sigma_K, \qquad \mathcal{R}_n^\Lambda \leq \frac{\overline{\lambda}}{\underline{\lambda} N} \frac{\operatorname{tr} \Sigma_K}{n}.$$

For $\Lambda = I$, the variance comparison is an equality: $\mathcal{R}_n^I = \operatorname{tr} \Sigma_K/(nN)$. Because $\|\psi(x)\|^2 = K(x,x) = 1 + d/2 + \rho^2/2$, with $\rho = \|x\|^2 \sim \chi_d^2$,

$$\begin{aligned}
\operatorname{tr} \Sigma_K &= \mathbb{E}K(x,x)^2 - N \\
&= \left(1 + \frac{d}{2}\right)^2 + \left(1 + \frac{d}{2}\right) d(d+2) + \frac{1}{4} d(d+2)(d+4)(d+6) - N \\
&= \frac{1}{4} d(d+2)(d+3)(d+9).
\end{aligned}$$

Dividing by the task variance $N = (d+1)(d+2)/2$ gives the stated $\delta_d = \Theta(N)$. If κ is bounded independently of d, the shallow bound tends to one uniformly in n whenever $w = o(d^2)$, while the same three-block circuit reaches relative risk ε using $O(\kappa N/\varepsilon)$ examples. The input representation requires width at least $d + 1$, so the constructive width order is linear. A full-feature one-block estimator also has $O(d^2)$ width and the same projection-kernel risk, establishing matching shallow width order for long enough contexts.

G REPRODUCIBLE ALGEBRAIC CHECKS

The accompanying script `verification/checks.py` performs end-to-end checks of both circuits across multiple dimensions and context lengths, exact symbolic checks of the moment identities, and finite-dimensional checks of the lower-bound expansion. All checks pass. The exact grids,

seeds, tolerances, and test counts are documented in `verification/README.md` and encoded in the script. These computational checks validate the implemented algebraic identities; the proofs establish the claims for all dimensions.

H COMPLETE LOWER-BOUND PROOFS

This section supplies the operator and coordinate arguments used in Section 4. The first lemma separates the information in linear task observations from the dimension of a prediction space. The second subsection verifies the query-space restriction for the stated architecture.

H.1 GAUSSIAN OBSERVATIONS AND THE RESIDUAL PREDICTION ERROR

Lemma H.1 (Gaussian observations and adaptive subspaces)**.** *Let $T : \mathbb{R}^N \to \mathcal{H}$ be linear, where $\mathcal{H}$ is a real Hilbert space, and write $G = T^*T$. Suppose $\theta \sim \mathcal{N}(0, I_N)$ and $g = C\theta$, with $\operatorname{rank} C = r$. For each g, let $\mathcal{V}_g \subset \mathcal{H}$ be a measurable subspace of dimension at most b. Every measurable predictor $F_\theta \in \mathcal{V}_{C\theta}$ satisfies*

$$\mathbb{E}\|F_\theta - T\theta\|_{\mathcal{H}}^2 \geq \tau_{r+b}(G), \tag{23}$$

even if its coefficients depend on the full vector θ.

Proof. Let Π be the orthogonal projection onto $\ker C$. Conditional Gaussian regression gives $\theta = \mu(g) + \xi$, where $\xi \sim \mathcal{N}(0, \Pi)$ is independent of g. Let P_g project onto $\mathcal{V}_g$. Projecting the target onto this subspace bounds the conditional error below by

$$\mathbb{E}[\|(I - P_g)T\theta\|^2 \mid g] = \|(I - P_g)T\mu(g)\|^2 + \|(I - P_g)T\Pi\|_{\mathrm{HS}}^2,$$

where $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert–Schmidt norm. The residual task directions have covariance Π after observing g. The operator $D_g = T(I - \Pi) + P_g T\Pi$ combines the observed component and the chosen prediction space. Its rank is at most $r + b$, since its two summands have ranks at most r and b. Its residual is $T - D_g = (I - P_g)T\Pi$. Every rank-at-most-m approximation D to T has squared Hilbert–Schmidt error at least $\tau_m(G)$. To see this, restrict $T - D$ to an $(N - m)$-dimensional subspace of $\ker D$ and minimize $\operatorname{tr}(PG)$ over projections of that rank. Applying this bound for each g and averaging proves the claim. For $m \geq N$ the claim is nonnegativity. □

H.2 COMPLETE PROOF OF LEMMA 4.1

Proof. After the first feed-forward map the tokens are $(v(x_i), y_i)$ and $(v(q), 0)$, with v independent of θ. Set

$$\Sigma = \frac{1}{n}\sum_i v(x_i)v(x_i)^\top, \quad g = \frac{1}{n}\sum_i v(x_i)y_i = C_X\theta, \quad h = \frac{1}{n}\sum_i y_i^2. \tag{24}$$

The first context moment is $S = \left(\begin{smallmatrix} \Sigma & g \\ g^\top & h \end{smallmatrix}\right)$. Let $E : \mathbb{R}^w \to \mathbb{R}^{w+1}$ embed feature coordinates, let e_y select the label, and replace h by zero in S to obtain S_g. Then the first query update is exactly

$$z_1(q) = \nu(q) + hpa(q), \quad \nu = (I + P_1 S_g Q_1)Ev, \quad p = P_1 e_y, \quad a = e_y^\top Q_1 Ev. \tag{25}$$

Conditional on X, g, the functions ν_j and a are fixed and belong to $\operatorname{span}\{v_1, \ldots, v_w\}$.

Let ℓ_j, m_j be row j of the two matrices in the second feed-forward map, viewed as linear functionals; both are zero for $j = w + 1$. Set $F_j = \nu_j + (\ell_j\nu)(m_j\nu)$. The jth query coordinate after this map expands to

$$\mathsf{FF}_{2,j}(z_1(q)) = F_j + hp_j a + ha\big[(\ell_j p)(m_j\nu) + (m_j p)(\ell_j\nu)\big] + h^2a^2(\ell_j p)(m_j p). \tag{26}$$

Since $a^2 \in \operatorname{span}\{av_1, \ldots, av_w\}$, the same function space contains every coordinate for all values of h:

$$\mathcal{V}_{X,g} = \operatorname{span}\{F_1, \ldots, F_{w+1}, a, av_1, \ldots, av_w\}. \tag{27}$$

The separate spanning function a is needed because the first feed-forward map need not preserve the constant function in the span of v. This space has dimension at most $2w + 2$. The Gram matrix of

the displayed spanning functions and its pseudoinverse define a measurable orthogonal projection, including at rank-deficient parameters.

All context updates are independent of q, since its column is excluded by M. Conditional on the full context and task, the second attention left-multiplies the query vector by a fixed matrix. Thus its label output remains in (27). □