RETHINKING ALGORITHMIC EXPLANATIONS OF TRANSFORMER IN-CONTEXT LEARNING: A SYSTEMATIC STUDY OF NOISY LINEAR REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer in-context learning (ICL) is increasingly being used to learn statistical procedures, making otherwise costly or impractical inference accessible. Their reuse across datasets makes it important to establish how faithfully they approximate the intended statistical procedure as the inputs change. Nevertheless, constructions showing that Transformers can implement familiar algorithms, together with agreement on selected distributions and isolated mechanistic tests, do not establish that trained models inherit those algorithms' generalisation properties. We address this gap by systematically studying Transformers trained to perform noisy linear regression in-context. We identify the main features organising their generalisation and how they respond to changes in pretraining, finding substantial competence beyond the training curriculum, but also significant sensitivity to input changes that leave the optimal prediction unchanged. Comparisons with the algorithms proposed in recent literature, supported by theory and targeted interventions, show that the trained Transformers violate structural properties those algorithms satisfy. The resulting characterisation supplies constraints for explanations of statistical ICL and insight for designing models that preserve the properties their intended use requires.

1 INTRODUCTION

Large language models can use a sequence of examples in their input to learn a corresponding task without updating their parameters (Brown et al., 2020). This ability, known as in-context learning (ICL), is being used to learn reusable alternatives to costly statistical inference. Transformer-based prior-data fitted networks (PFNs), for example, learn Bayesian prediction from simulated datasets, enabling fast approximate inference for models whose posterior computation would otherwise be costly or impractical (Müller et al., 2022). Indeed, PFNs have been shown to outperform established methods for prediction on small tabular datasets, while requiring substantially less computation per dataset after pretraining (Hollmann et al., 2025).

Reliable use of these models for statistical inference requires understanding the conditions under which they generalise. Following Garg et al. (2022), linear regression has become a widely used setting for investigating the learning procedures acquired by Transformers (Akyürek et al., 2023; von Oswald et al., 2023), since its well-understood solutions provide explicit references for assessing their generalisation. In the noisy setting considered here, each task is specified by a coefficient vector $\beta \sim \mathcal{N}(0, I_d)$. The Transformer receives a context of n examples (x_i, y_i), with covariates x_i independent of β, where $y_i = x_i^\top \beta + \varepsilon_i$ and the observation noises are independent draws from $\mathcal{N}(0, \sigma^2)$. For a new target input $x_\star$, also referred to as the query, the Transformer is trained to predict the underlying noiseless response $x_\star^\top \beta$. Under these assumptions, the posterior mean $\mathbb{E}[x_\star^\top \beta \mid X, y, x_\star]$, which minimises the expected squared prediction error, is given by the ridge predictor

$$f_R(X, y, x_\star) = x_\star^\top (X^\top X + \sigma^2 I_d)^{-1} X^\top y, \tag{1}$$

where $X \in \mathbb{R}^{n \times d}$ is the matrix whose rows are $x_i^\top$ and $y \in \mathbb{R}^n$ are the observed responses. We keep the coefficient prior and noise law fixed, and draw input covariates independently of β. We can therefore vary the input distribution and the number of examples while retaining the same optimal prediction rule in Eq. (1).

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.

A substantial literature has sought to explain regression ICL in terms of familiar algorithms. Support for these accounts comes from constructions showing that Transformers can implement the algorithms, analyses of learning in single-layer or looped architectures, and empirical comparisons with trained models. Section 4.1 reviews these proposals in more detail.

Yet generalisation tests already expose discrepancies between trained Transformers and some of these algorithms, both under changes in input scale, mean, or subspace (Garg et al., 2022; Ahuja & Lopez-Paz, 2023; Zhang et al., 2024; Hill et al., 2025) and under adversarially chosen inputs (Anwar et al., 2025). These discrepancies show that the generalisation properties of familiar regression algorithms cannot simply be attributed to trained Transformers on the basis of representability or agreement on selected distributions. Yet identifying particular failures does not establish what governs the range of accurate prediction, how pretraining changes that range, or whether the proposed algorithms account for the failures themselves. Without such an account, these algorithmic interpretations provide limited grounds for anticipating how trained models will behave as their inputs change. We therefore establish the following:

- **Input scale and effective dimension organise broad generalisation beyond pretraining.** Transformers trained on two simple covariance families approximate ridge across much richer input distributions. Their average error is largely organised by input scale $s(M) = \operatorname{tr} M/(n_0 d\sigma^2)$, and effective dimension $d_{\text{eff}}(M) = \operatorname{tr}[M(M+\sigma^2 I)^{-1}]$, where $M = X^\top X$ and $n_0 = 80$ is a fixed reference context count. Richer covariance pretraining does not consistently improve accuracy, while pretraining on larger scales improves accuracy at larger scales at the expense of accuracy on smaller ones.
- **Substantial violations of algorithmic invariances persist despite targeted pretraining.** Contexts of the same length with identical M and $b = X^\top y$ can elicit substantially different Transformer predictions at a fixed query, despite ridge and several proposed algorithms being exactly invariant to these changes. Even after a strong explicit penalty on symmetry violations during pretraining, significant sensitivity remains.
- **Proposed algorithms' query structure provably fails for the trained Transformers.** Earlier work shows that LayerNorm can bound the magnitude of Transformer predictions (Naim et al., 2025a). Using boundedness and continuity, we prove that, for the studied architecture, predictions can be affine in the target input magnitude in every direction and at all positive magnitudes only if they are independent of the target input, in particular ruling out global agreement with the true ridge predictor and with all reviewed algorithmic proposals. Moreover, we show empirically that the failure to be linear in the target inputs is significant even for magnitudes frequently present in the pretraining distribution.
- **Finite iteration alone does not explain observed failure at large input scales.** We prove that the inverse iteration proposed by Fu et al. (2024) as an underlying mechanism for softmax Transformers admits relative error bounds uniform in context scale with a fixed number of iterations. The Transformer's observed errors exceed these bounds, showing that its scale failures are not inevitable consequences of finite iteration. Nevertheless, we also prove that for a wide class of gradient-based algorithms, which have been hypothesised in the ICL literature to be unrolled across Transformer layers, the relative error will diverge with increasing scale whenever the method improves on the zero predictor at some scale.

By combining systematic controlled changes to the inputs and pretraining with theoretical tests of proposed algorithms, we characterise the generalisation of trained Transformers and establish where those algorithms fail to explain it. The resulting constraints give reason to reconsider the simplifications used to explain regression ICL and guide the design of models that can be more reliably reused across datasets.

1.1 Related Work

Alongside the successes and failures under input distribution shifts reviewed above, work on task (β) diversity examines generalisation to regression functions beyond those used in pretraining. Increasing the number of pretraining tasks can induce ridge-like prediction on unseen tasks (Raventós et al., 2023), generalisation also depends on how broadly the pretraining task vectors are spread (Goddard et al., 2025), and on differences between the pretraining and test task distributions (Letey et al., 2026). The related question of which learning procedure is acquired has motivated accounts based

on gradient updates, iterative matrix inversion and algorithm selection. Section 4.1 reviews these proposals and their supporting evidence.

2 EXPERIMENTAL SET-UP

2.1 PARAMETRISING THE CONTEXT

As we saw, for the regression model in Section 1, the posterior mean coefficient vector is

$$w_R = \mathbb{E}[\beta \mid X, y] = (M + \sigma^2 I_d)^{-1} b, \tag{2}$$

so that $f_R(X, y, x_\star) = x_\star^\top w_R$. In our experiments, we consider $d = 40$ and $\sigma^2 = 1$. Further, we call a context and its query inputs a *prompt*, and a distribution over these and the number of context points a *prompt environment*. In particular, even for different prompt environments, the loss-minimising predictor for every realised prompt is always Eq. (1).

The context enters the ridge predictor only through (M, b). To make the remaining degrees of freedom explicit, let $\lambda_1, \dots, \lambda_d > 0$ be the eigenvalues of M/σ^2 for a context of full column rank, and write

$$M = \sigma^2 U \operatorname{diag}(\lambda_1, \dots, \lambda_d) U^\top, \qquad X = \sigma V \operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_d}) U^\top, \tag{3}$$

where $U \in O(d)$ and $V \in \mathbb{R}^{n \times d}$ satisfies $V^\top V = I_d$. Conditional on X, the response admits the decomposition

$$y = \sigma V \operatorname{diag}(\sqrt{1+\lambda_1}, \dots, \sqrt{1+\lambda_d}) z + \varepsilon_\perp, \qquad z \sim \mathcal{N}(0, I_d), \qquad X^\top \varepsilon_\perp = 0. \tag{4}$$

Here z and $\varepsilon_\perp$ are conditionally independent, with $\varepsilon_\perp \mid X \sim \mathcal{N}(0, \sigma^2(I_n - VV^\top))$. With $h_i = \lambda_i/(1+\lambda_i)$, we obtain $f_R(X, y, x_\star) = (U^\top x_\star)^\top \operatorname{diag}(\sqrt{h_1}, \dots, \sqrt{h_d}) z$. Thus the ridge prediction depends on the context and query through the realised spectrum $\{\lambda_i\}$, informative response z, and $U^\top x_\star$, and is independent of V and $\varepsilon_\perp$.

2.2 TRANSFORMER AND COMPACT SPECTRAL CURRICULUM

Our reference model is an eight-layer pre-LayerNorm softmax Transformer of width 128 with four attention heads. Each context pair is represented by one token $[x_i, y_i, 0] \in \mathbb{R}^{d+2}$ and each target by $[x_\star, 0, 1]$, followed by a shared linear embedding. There are no positional embeddings. Context tokens attend bidirectionally within the context; target tokens attend to the context but are never used as keys or values, so target predictions are conditionally independent of the other target tokens. Full architectural and optimisation details are given in Appendix A.

Our baseline curriculum C_0 draws a reference covariance Σ_{ref}. Using the reference count $n_0 = 80$, define $\bar{M} = n_0 \Sigma_{\mathrm{ref}}$, $\Sigma_n = \bar{M}/n$, and sample the n context covariates independently from $\mathcal{N}(0, \Sigma_n)$. Thus $\mathbb{E}[X^\top X \mid \bar{M}, n] = \bar{M}$. Each training batch uses a count n sampled uniformly from $\{40, 80, 160, 320\}$.

Let $\rho_1, \dots, \rho_d$ denote the eigenvalues of $\Sigma_{\mathrm{ref}}/\sigma^2$, and set $s_0 = d^{-1} \sum_i \rho_i = s(\bar{M})$. The curriculum draws

$$s_0 \sim \mathrm{LogUniform}[0.5, 500], \qquad \sum_{i=1}^{d} \frac{\rho_i}{1+\rho_i} \;\Big|\; s_0 \sim \mathrm{Uniform}\left[1, \frac{d s_0}{1+s_0}\right], \tag{5}$$

where the second co-ordinate is the effective dimension of the reference covariance. We then select one of two spectral shapes with equal probability. The *smooth* family has a log-linear spectrum $\rho_i \propto \exp[-\ell(i-1)/(d-1)]$, while the *step* family has two eigenvalue levels, with the upper multiplicity sampled uniformly from its feasible values. In both cases the free parameters are chosen to match the sampled scale and reference-covariance effective dimension. The eigenbasis of Σ_{ref} is sampled independently from the Haar distribution on $O(d)$. Thus the curriculum varies scale, effective dimension, context count and feature orientation broadly while restricting residual spectral shape to two simple families.

For each of eight target queries we draw $x_\star = au$ with $u \sim \mathcal{N}(0, I_d)$ and $\log_2 a \sim \mathrm{Uniform}[-1, 1]$, and the target label is noiseless, so $y_\star = x_\star^\top \beta$. We train for 500,000 optimiser steps using target-only squared error dividing each target error, normalised by $da^2 = \mathbb{E}[y_\star^2 \mid a]$. This prevents larger query scales from being overweighted and, since the normalisation depends only on the sampled query scale, does not change the population minimiser. Section 3 evaluates the resulting model on realised spectra substantially richer than the two training spectral families.

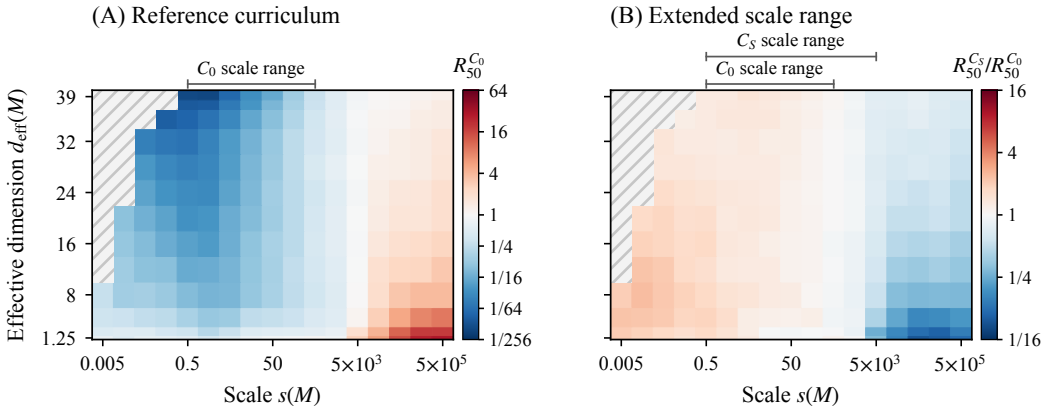


Figure 1: **Agreement with ridge across context spectra.** (A) Median relative risk R_{50} for C_0 over 4,096 spectra, with averaging within each spectrum. (B) Ratio of median risks for C_S and C_0 on matched prompts. Both use $n = 80$. Brackets mark sampled training scales s_0. Hatching marks infeasible coordinates; colour scales are logarithmic (Appendix B.1).

2.3 Evaluation

For the spectral comparisons, we sample realised spectra at fixed scale and effective dimension, using the coordinates defined in Section 1. At a feasible pair (s, d_{eff}), draw q uniformly from $\{q \in \mathbb{R}^d : \sum_i q_i = 0,\ \|q\|_2 = 1\}$ and set

$$\lambda_i(q) = n_0 d s \frac{\exp(\alpha q_i)}{\sum_j \exp(\alpha q_j)}, \qquad \sum_{i=1}^{d} \frac{\lambda_i(q)}{1 + \lambda_i(q)} = d_{\text{eff}}, \tag{6}$$

where the second equation determines $\alpha \geq 0$. At each feasible (s, d_{eff}), Eq. (6) defines our reference spectral law. We call draws from this law reference spectra.

Write $\Lambda = (\lambda_1, \ldots, \lambda_d)$ for the sampled spectrum. At a fixed context count, we construct X using Eq. (3) with independent Haar-distributed U, V, draw responses using Eq. (4), and draw queries as in Section 2.2. For each spectrum, we average squared prediction discrepancies and squared ridge predictions over these draws, then divide the two averages. The median of these per-spectrum relative risks is

$$R_{50}(s, d_{\text{eff}}) = \operatorname*{median}_{\Lambda} \left[\frac{\mathbb{E}[(f_\theta - f_R)^2 \mid \Lambda]}{\mathbb{E}[f_R^2 \mid \Lambda]} \right], \tag{7}$$

where the median is over the reference law above, with the model and context count held fixed. The normalisation gives relative risk zero for ridge and one for the zero predictor.

3 What does a standard Transformer learn?

3.1 Agreement with ridge across spectra

After pretraining on the two covariance families in C_0, agreement with ridge across reference spectra is strongest for lower scales, especially at high effective dimension (Figure 1(A)). At $s = 0.5$, median relative risk is 0.568, 0.035 and 0.0066 at $d_{\text{eff}} = 1.25$, 20 and 39, respectively. At $d_{\text{eff}} = 20$, it rises to 0.385 at $s = 500$ and 2.036 at $s = 500{,}000$, so that deterioration is therefore already substantial at scales comparable to the upper end of the pretraining range. Larger relative risks at small d_{eff} partly reflect the normalisation: $\mathbb{E}[f_R^2 \mid \Lambda] = \mathbb{E}[a^2] d_{\text{eff}}$, so a fixed mean squared discrepancy gives larger relative risk at smaller d_{eff}.

Fixing (s, d_{eff}) generally leaves 38 spectral degrees of freedom. We maximise conditional relative risk over these using multi-start gradient ascent, including separate searches over feasible rank-deficient spectra. Searches start from the reference spectra at each (s, d_{eff}), with candidates validated on fresh common draws (Appendix B.1). At every point in Figure 1(A), the estimated maximum is at

most 3.26 times the matched reference median, whereas medians span more than three orders of magnitude. Scale and effective dimension therefore continue to organise agreement when spectral shape is optimised to increase error.

Extending the training scale range from $[0.5, 500]$ to $[0.5, 5{,}000]$ (with this new environment denoted by C_S), at the same training budget, improves large-scale agreement at a cost within the original range (Figure 1(B)). At $d_{\text{eff}} = 20$, median risk falls from 0.998 to 0.680 at $s = 5{,}000$, but rises from 0.035 to 0.057 at $s = 0.5$. Related effects of training range on extrapolation and accuracy have been reported for noiseless univariate tasks (Naim et al., 2025a;b).

3.2 Does limited spectral diversity explain the remaining errors?

The remaining dependence on spectral shape could reflect the limited variety of covariance spectra used in pretraining. However, errors are already large on a family closely related to pretraining. Let $\mathcal{H}$ denote the two-level spectra with ridge gains $h_H \geq 0.95$ and $0.10 \leq h_L \leq 0.80$. Across evaluated points with $0.5 \leq s \leq 500$ and $8 \leq d_{\text{eff}} \leq 32$, the median ratio of C_0's median risk on $\mathcal{H}$ to its estimated maximum is 0.90. Moreover, 27.3% of expected Gram matrices sampled from C_0 within these bounds belong to $\mathcal{H}$. Thus, this family on which the Transformer achieves close to its worst found error is also strongly represented in pretraining, suggesting that it is not a simple lack of exposure to other spectra that causes errors beyond those organised by (s, d_{eff}).

To further investigate this, we test whether increasing spectral diversity improves agreement by pretraining a separate model under a new environment C_D defined by modifying C_0 as follows: each sampled spectrum of the expected Gram matrix is instead replaced by a draw from the reference law described in Eq. (6) at the same scale and effective dimension. We do not find a systematic improvement, but rather changes that depend strongly on the scale and effective dimension (Figure 5(A)). The median ratio $R_{50}^{C_D}/R_{50}^{C_0}$ across evaluated points is 0.60 for $d_{\text{eff}} \leq 2$ and $0.5 \leq s \leq 5{,}000$, but 0.978 in the region considered above. At $s \geq 50{,}000$, risk increases at all but one evaluated point, with median ratio 1.33. Repeating the spectral optimisation independently for C_D gives a similar pattern for estimated maximum risks: improvements in the region of low effective dimension, but no consistent improvement for intermediate scales and effective dimensions, and degradation at large scales (Figure 6(A)). This gives further evidence that it is not the restricted spectral shape of C_0 that is primarily responsible for the remaining error beyond (s, d_{eff}).

3.3 Sensitivity to equivalent representations

The preceding results average over context representations at each spectrum. We now ask whether predictions remain consistent under transformations of the context that leave the ridge solution invariant. We consider three such transformations: joint rotations of the context and target covariates, joint rotations of the context covariates and responses, and changes to the orthogonal noise $\varepsilon_\perp$ that preserve X and the informative response component.

We evaluate 64 spectra sampled uniformly without replacement from the reference spectra across the 49 evaluated points with $0.5 \leq s \leq 500$ and $8 \leq d_{\text{eff}} \leq 32$. For each spectrum, we draw 16 independent contexts (X, y), with eight fixed queries for each (see Appendix C for details). For a prompt $Z = (X, y, x_\star)$, let Z', Z'' be two independently sampled representations obtained by applying the same type of transformation G to Z. We measure

$$D_G = \frac{\frac{1}{2}\mathbb{E}_Z\mathbb{E}[(f_\theta(Z') - f_\theta(Z''))^2 \mid Z]}{\mathbb{E}_Z[f_R(Z)^2]}. \tag{8}$$

The outer expectation averages uniformly over these states and their queries. Figure 2(A) also reports percentiles of the individual squared differences divided by the same $2\mathbb{E}[f_R(Z)^2]$, before averaging.

For covariate rotations under C_0, the mean is 0.0342 and the 99th percentile is 0.380. The latter corresponds to a squared difference between predictions on equivalent prompts of 76% of the mean squared ridge prediction. Context rotations produce comparable discrepancies, while changes to orthogonal noise have a smaller but non-zero effect.

To test whether directly penalising violations of these equivalences reduces the discrepancies, we pretrain the same Transformer architecture on the same prompt environment, but with a modified loss. Each sampled prompt is paired with a randomly transformed version obtained by jointly rotating

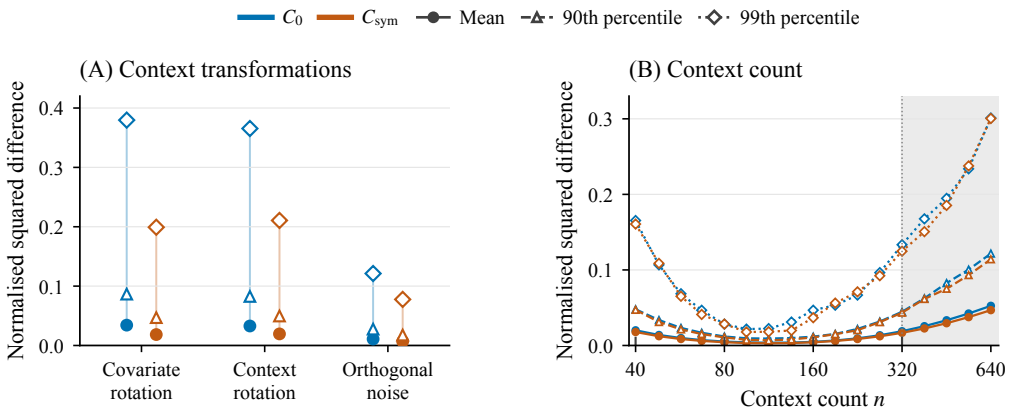


Figure 2: **Sensitivity to context representation.** (A) At $n = 80$, means estimate D_G; percentiles describe individual normalised squared differences. (B) Means and percentiles of Eq. (9) across states, after averaging representations. Base ticks mark training counts; shading marks larger counts. Both panels use 1,024 states from 64 reference spectra (Appendix C).

the context and target covariates and jointly rotating the context covariates and responses, with the loss also penalising the difference between predictions for these paired prompts. We call this new curriculum C_{sym} (see Appendix A.1). This substantially reduces sensitivity but leaves appreciable differences between equivalent prompts. For covariate rotations, the mean discrepancy falls from 0.0342 to 0.0184 and the 99th percentile from 0.380 to 0.199. Thus, even after explicitly penalising prediction differences, the squared difference at the 99th percentile remains approximately 40% of the mean squared ridge prediction. Context rotations show similar residual sensitivity, as does orthogonal noise to a more limited extent. Median agreement with ridge neither improves nor degrades consistently under C_{sym} (see Figure 7).

In a separate experiment, we vary context count while preserving M, b and eight fixed queries. The number of target queries cannot affect an individual prediction by construction (Section 2.2). Let $\bar{f}_{\theta,j}(n)$ be the prediction for query j, averaged over 128 context realisations with the same M, b at count n (Appendix C). We measure its squared departure from its average over the four pretraining counts:

$$\frac{\frac{1}{8}\sum_{j=1}^{8}\left[\bar{f}_{\theta,j}(n) - \frac{1}{4}\sum_{k\in\{40,80,160,320\}}\bar{f}_{\theta,j}(k)\right]^2}{\mathbb{E}_Z[f_R(Z)^2]}. \tag{9}$$

Figure 2(B) reports means and percentiles across states. We find significant discrepancies even at the pretraining counts, especially smaller and larger ones, and these only grow as n increases beyond their range. At $n = 640$, C_0 has mean 0.0525 and 99th percentile 0.301.

3.4 Dependence on the response and query

The preceding transformations preserve the ridge prediction. We now vary the informative response and query, on which that prediction depends. Both experiments evaluate C_0 at $n = 80$ on the same 64 reference spectra used in Section 3.3. For each spectrum, we sample uniform unit directions v and vary $z = r_z v$, reusing the directions across radii. At each fixed spectrum and z, we average over $U, V, \varepsilon_\perp$ and ordinary queries. Since the ridge prediction vanishes as $r_z \to 0$, we normalise by the conditional second moment of the noiseless target:

$$\widetilde{R}_\theta = \frac{\mathbb{E}[(f_\theta - f_R)^2 \mid \Lambda, z]}{\mathbb{E}[y_\star^2 \mid \Lambda, z]}. \tag{10}$$

At each radius, we compute percentiles of these conditional ratios over response directions within each spectrum. Figure 3(A) reports the median of each percentile across the 64 spectra. Between $r_z/\sqrt{d} = 1$ and 4, the 50th-percentile curve rises from 0.032 to 0.251, and the 99th-percentile curve from 0.049 to 0.517. The median over response directions increases between these radii within

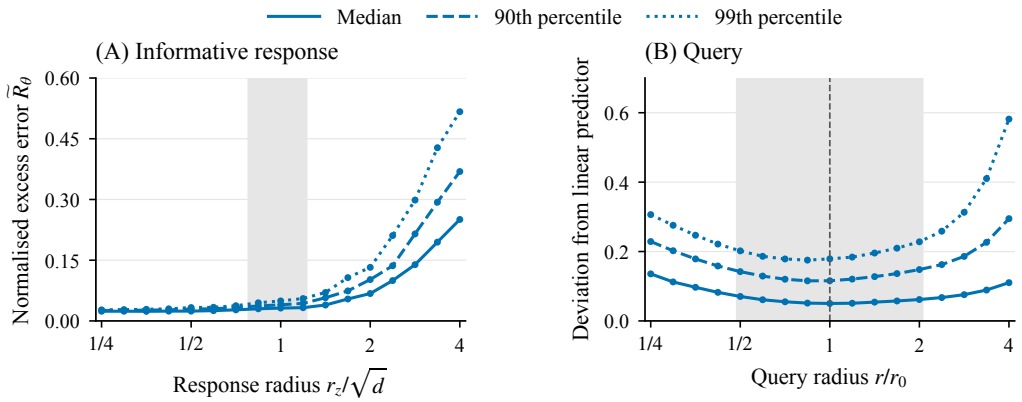


Figure 3: **Response magnitude and query dependence.** Both panels evaluate C_0 at $n = 80$ on the 64 reference spectra of Figure 2. (A) Within-spectrum percentiles over response directions of Eq. (10), summarised by medians across spectra. (B) Within-context percentiles over query directions of Eq. (11), summarised by medians across contexts. Shading marks the central 95% of pretraining radii; the dashed line marks the median pretraining query radius r_0.

every reference spectrum. We therefore find that generalisation degrades as the magnitude of the informative response is increased beyond values typically seen in pretraining.

For query dependence, we reuse the 1,024 fixed contexts from Section 3.3. In each context, we calibrate a coefficient $\widehat{w}_\theta(X, y)$ by querying the model at $r_0\phi_i$ and $-r_0\phi_i$ for each vector ϕ_i of a randomly sampled orthonormal basis, where r_0 is the median pretraining query radius (Appendix D). This calibration recovers the coefficient exactly for a predictor linear in the query. Holding it fixed, we evaluate fresh uniform unit directions u and measure

$$\frac{|f_\theta(X, y, ru) - ru^\top \widehat{w}_\theta(X, y)|}{r\|\widehat{w}_\theta(X, y)\|_2/\sqrt{d}}. \tag{11}$$

The denominator is the square root of the calibrated predictor's mean squared output over directions. Figure 3(B) shows medians across contexts of the directional percentiles computed within each context. The 90th-percentile curve rises from 0.116 at r_0 to 0.295 at $4r_0$. Thus, despite the curriculum deliberately using a wider range of magnitudes by adding a log-uniform prefactor a (as defined in Section 2.2), the Transformer displays significant non-linearity in the query, including at query magnitudes typical during pretraining.

4 Theoretical tests of algorithmic explanations

We now examine whether the algorithms proposed to explain regression ICL account for the generalisation patterns established in Section 3.

4.1 Candidate regression algorithms

Many algorithmic accounts propose that the Transformer approximates a computation of the form $x_\star^\top w_K(X, y)$ where w_K is a coefficient evaluated after K iterative updates which depends only on the context. Akyürek et al. (2023) began by proposing two candidates for w_K, the first of which uses the formula $w_{k+1} = w_k + \eta(x_{k+1}(y_{k+1} - x_{k+1}^\top w_k) - \lambda_{\text{reg}} w_k)$, with regularisation coefficient $\lambda_{\text{reg}} \geq 0$, while the second uses Sherman–Morrison updates. In parallel, von Oswald et al. (2023) connect a linear-attention layer to a gradient step using the whole context, so that $w_{k+1} = w_k + A_k(b - Mw_k)$, $A_k = (\eta_k/n)I_d$. Their deeper models outperform this plain gradient descent, motivating GD++ which matches their performance more closely, and at every step also updates the context and target covariates by $x \mapsto (I_d - \gamma M_k)x$ where γ is a tuned scalar, and M_k is the Gram matrix recomputed from the updated context at step k.

A complementary development identifies one-step GD or preconditioned-GD predictors at population optima of restricted linear-attention models, where preconditioned-GD replaces A_k by a learned

matrix that is fixed during inference (Ahn et al., 2023; Mahankali et al., 2024). Zhang et al. (2024) additionally prove that, under suitable initialisation, a single-layer linear-attention model converges under gradient flow to a preconditioned-GD step, with the preconditioner determined by the training distribution. Gatmiry et al. (2024) extend this optimality analysis to multiple steps in a looped linear-attention model, and show that under their assumptions population optima implement preconditioned-GD with a shared matrix.

Fu et al. (2024) instead argue that the faster than GD convergence of the Transformer's predictions if read-out after each layer is better explained by a second-order method, with Newton–Schulz inverse iteration as their principal candidate. This maintains an estimate B_k of the inverse context Gram matrix and updates it by $B_{k+1} = 2B_k - B_k M B_k,\ w_K = B_K b$. Alternatively, Bai et al. (2023) propose that Transformers can select among candidate algorithms by forming a convex combination of their predictions, with the weights depending on the context.

The gradient explanation has also been developed for softmax attention. He et al. (2025) find that a trained single layer of multi-head softmax attention approximates one debiased gradient step, obtained by replacing each context input x_i by $x_i - \bar{x}$, where $\bar{x} = n^{-1}\sum_i x_i$. Here, while the softmax attention mechanism means that the hypothesised form of the single layer is not exactly linear in query $x_\star$, they approximate gradient updates which are. Nevertheless, we also find in Ahn et al. (2023) that a restricted single-layer ReLU attention model leads to an optimal predictor that is no longer query-linear, but remains affine in the query magnitude.

4.2 Context invariance and query dependence

At fixed context count and query, full-context GD, preconditioned GD, GD++, and Newton–Schulz algorithm analysed below, depend on the context only through M, b. Their predictions are therefore invariant at every iteration to the context rotations and orthogonal-noise changes tested in Section 3.3, whereas the trained Transformers remain sensitive to both, even after a strong pretraining penalty.

Now, to examine query dependence, fix the context (X, y) and write the query as $x_\star = ru$, with $\|u\|_2 = 1$ and $r > 0$. We call a predictor radially affine if, for every direction u,

$$f(X, y, ru) = r\alpha_{X,y}(u) + \beta_{X,y}(u), \qquad r > 0. \tag{R}$$

Then, we see that the proposed algorithms above are all radially affine. For the architecture we have been considering, however, we can in fact prove that radial affinity cannot be globally satisfied. Naim et al. (2025a) have related failures to extrapolate linear functions to bounded predictions caused by LayerNorm, and similarly for our architecture, the final LayerNorm and affine readout likewise bound the prediction, giving:

Proposition 1 *Let f_θ be the model of Section 2.2, with finite parameters and positive LayerNorm epsilon. For every fixed nonempty context (X, y), $f_\theta(X, y, \cdot)$ satisfies* (R) *for all positive radii and directions if and only if it is independent of the query. Moreover, for every unit direction u with $u^\top w_R \neq 0$,*

$$\lim_{r\to\infty} \frac{|f_\theta(X, y, ru) - f_R(X, y, ru)|}{r} = |u^\top w_R|,$$

so that the absolute prediction error grows linearly in the query radius.

The proof is given in Appendix F.1. Moreover, Figure 3(B) shows appreciable departures from a context-calibrated linear predictor even at query magnitudes typical of pretraining.

4.3 Context scale and finite computation

We next test whether finite iteration of the proposed algorithms explains the deterioration at large context scales observed in Section 3. To isolate a change in overall scale, fix a full-column-rank design X_0 and consider $X_t = \sqrt{t}X_0,\ y_t = X_t\beta + \varepsilon,\ t \geq 1$, with fixed n, d and the unit-prior, unit-noise regression law and independent query distribution of Section 2. Write $M_t = X_t^\top X_t$, $b_t = X_t^\top y_t$, and let e_t denote this prompt environment. For an algorithm $\mathcal{A}$, write $R_{\mathcal{A}}(e_t)$ for its relative risk under this environment, using the same normalisation as in Section 2.3. Here X_0 is fixed and the expectations are over β, ε and $x_\star$.

Consider first the inverse iteration implemented in Fu et al. (2024)'s released code, applied to the unit-ridge matrix $M_t + I_d$.

$$B_0 = \frac{M_t + I_d}{\|(M_t + I_d)^2\|_F}, \qquad B_{k+1} = 2B_k - B_k(M_t + I_d)B_k, \qquad \widehat{w}_K = B_K b_t.$$

The prediction is then $x_\star^\top \widehat{w}_K$. Denote this algorithm after K updates by $\mathcal{A}_K$, and let $\kappa_0 = \lambda_{\max}(X_0^\top X_0)/\lambda_{\min}(X_0^\top X_0)$ be the condition number of the fixed design's Gram matrix.

Theorem 1 *Under the regression law above, for $d \geq 2$ and for every fixed integer $K \geq 0$,*

$$\sup_{t \geq 1} R_{\mathcal{A}_K}(e_t) \leq \min\left\{1 - \frac{1}{d},\ \exp\left[-\frac{2^{K+1}}{\sqrt{d}\,\kappa_0^2}\right]\right\}.$$

The proof is given in Appendix F.2. Immediately, we see that the bound $1 - 1/d$ is independent of both scale and condition number, yet the Transformer exceeds it. At $s = 500{,}000$ and $d_{\text{eff}} = 20$, its median relative risk is 2.0362 (Figure 1(A)). Furthermore, arbitrary accuracy can be attained for all $t \geq 1$ by choosing a sufficiently large number of updates, independent of t.

On the other hand, several of the algorithms proposed in Section 4.1 can themselves produce even unbounded relative error as context scale increases. We establish that, for a broad class of finite computations, this divergence is unavoidable once the algorithm achieves relative error below one at any scale. Consider an algorithm that computes $w_K(X, y)$ through a fixed finite sequence of additions and multiplications of the entries of X, y, and fixed constants or matrices, and predicts $x_\star^\top w_K(X, y)$. We call this a finite polynomial context computation.

Theorem 2 *Let $\mathcal{A}$ be a finite polynomial context computation, with all parameters and its computation budget fixed across the dilation. For every full-column-rank X_0, under the regression law above, either $R_{\mathcal{A}}(e_t) \sim c\,t^D$ as $t \to \infty$ for some constant $c > 0$ and integer $D \geq 1$, or $R_{\mathcal{A}}(e_t) \geq 1$ for every $t \geq 1$.*

The proof is given in Appendix F.3. This class includes a large portion of the proposed algorithms from Section 4.1, including the gradient-based updates in Akyürek et al. (2023), GD, GD++, and preconditioned-GD updates, as well as He et al. (2025)'s debiased gradient step.

We also pretrain a variant that satisfies (R) by construction, and find that it retains substantial sensitivity to equivalent representations, and its median relative risk at large context scales still exceeds the uniform bound for finite inverse iteration in Theorem 1 (see Appendix E for details). Thus, enforcing (R) does not resolve either the violations of context invariance or the scale discrepancy with inverse iteration.

5 Discussion

Assessing our findings overall, the discrepancies between trained Transformers and proposed algorithms give reason to reconsider the simplifications used to explain regression ICL. Single-layer and looped models may exclude interactions between independently parametrised layers that matter for generalisation. A potentially fruitful research direction would consider architectures with simpler computations within each layer while retaining multiple layers with distinct parameters, and analyse their joint training to understand how the behaviours observed here arise.

More broadly, our findings that the trained models remain sensitive to input changes that preserve ridge predictions, even when symmetry violations are explicitly penalised or radial affinity is enforced, and that training on more varied covariance matrices does not consistently improve accuracy, suggest that finite training of a model as flexible as a multi-layer Transformer is likely to produce accurate approximations adapted to the pretraining distribution without learning the intended procedure's invariances. To realise the promise of Transformer ICL in more complex and practically relevant settings, for example through PFNs, robust generalisation beyond the pretraining distribution is key. By systematically identifying where that generalisation succeeds and fails for linear regression, our study provides empirical constraints that can guide more robust architectures in this simple setting, and a pathway to robust neural learning of more complex Bayesian procedures.

AI USE STATEMENT

In this work, we used generative AI tools to aid with and polish writing, using some of their suggestions iteratively while drafting the text, and using them to record implementation details in the appendices. They were also used to help with literature reviews, and as aids when developing research plans for certain experimental investigations that led to this work, as well as helping to choose the right metrics to plot in Sections 3.3 and 3.4. They were used to investigate and review a range of mathematical claims related to softmax Transformer regression ICL which led to the choice of the specific results presented in Section 4, and were used to prove them. Most importantly for the development of the work, generative AI tools were used to implement and package the code for the experiments carried out, including those important to its development but not included in it, and including scripts for figure generation. The authors verified every citation and reviewed AI generated code and mathematical arguments. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Details of all constructions, training procedures, and experiments are provided in the appendices so as to be reproducible, as well as all proofs. Moreover, the code used for training the architectures in this work, as well as the code for evaluation, figure generation, or other experiments, and the trained Transformers themselves, are all found in the supplementary material.

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.

Kartik Ahuja and David Lopez-Paz. A closer look at in-context learning under distribution shifts. In *Workshop on Efficient Systems for Foundation Models at ICML*, 2023.

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.

Usman Anwar, Johannes von Oswald, Louis Kirsch, David Krueger, and Spencer Frei. Understanding in-context learning of linear models in Transformers through an adversarial lens. *Transactions on Machine Learning Research*, 2025.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98675–98716, 2024.

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.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15130–15152. PMLR, 2024.

Chase Goddard, Lindsay M. Smith, Vudtiwat Ngampruetikorn, and David J. Schwab. When can in-context learning generalize out of task distribution? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19585–19599. PMLR, 2025.

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22686–22742. PMLR, 2025.

Joshua Hill, Benjamin Eyre, and Elliot Creager. Transformers don't in-context learn least squares regression. In *Workshop on Reliable and Responsible Foundation Models at ICML*, 2025.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Mary I. Letey, Jacob A. Zavatone-Veth, Yue M. Lu, and Cengiz Pehlevan. Pretrain–test task alignment governs generalization in in-context learning. In *International Conference on Learning Representations*, 2026.

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.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Omar Naim, Jérôme Bolte, and Nicholas Asher. Analyzing limits for in-context learning. In *NeurIPS 2025 Workshop on What Can('t) Transformers Do?*, 2025a.

Omar Naim, Guilhem Fouilhé, and Nicholas Asher. Re-examining learning linear functions in context. In *KI 2025: Advances in Artificial Intelligence*, volume 15956 of *Lecture Notes in Computer Science*, pp. 104–117. Springer, 2025b.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14228–14246, 2023.

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. PMLR, 2023.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained Transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

A IMPLEMENTATION AND TRAINING

Architecture and optimisation. Each of the eight pre-LayerNorm blocks has four softmax attention heads and a two-layer feedforward map of width 256 with GELU activation. The residual width is 128. A shared affine map embeds the $d + 2 = 42$ input coordinates, and a final LayerNorm and affine scalar readout produce the prediction. Attention projections have no bias. There are no positional embeddings or dropout. The attention mask excludes every target token from the keys and values for all tokens. Each target therefore depends on its own input and the context, but not on other targets in the same forward pass.

The models C_0, C_S, C_D and C_{sym} use AdamW with learning rate 5×10^{-5}, coefficients $(0.9, 0.999)$, numerical constant 10^{-8} and weight decay 10^{-4}. The learning rate increases linearly for 1,000 steps and then follows a cosine schedule to 10^{-5} at step 500,000. We use no gradient clipping. Data are generated in double precision and the network is evaluated and trained in single precision. Every comparison uses the step-500,000 checkpoint from one training run per model, with seed 0. Evaluation intervals describe sampling variability conditional on these trained models, not variability across training runs. Appendix E gives the architectural changes for $C_0^{(R)}$, which uses the same optimisation procedure.

Training loss and curricula. For a query $x_\star = au$, the supervised loss is $(f_\theta - y_\star)^2/(da^2)$, averaged over eight queries per prompt and 32 prompts per step. Under the independent standard Gaussian laws for u and β, $\mathbb{E}[y_\star^2 \mid a] = da^2$. Evaluation instead uses the unweighted squared discrepancy from ridge, normalised as described in Section 2.3.

C_S changes only the upper endpoint of the sampled s_0 range to 5,000. For C_D, first draw $\bar{M}$ under C_0. Draw a fresh uniform unit direction $q \in \mathbf{1}^\perp$, then use the construction in Eq. (6) with $s = s(\bar{M})$ and effective dimension $d_{\mathrm{eff}}(\bar{M})$ to replace the eigenvalues of $\bar{M}/\sigma^2$. This preserves both quantities exactly before the finite context is sampled. The other task, query, count, architecture and optimisation choices are unchanged.

A.1 SYMMETRY TRAINING

C_{sym} uses 16 independent prompts and 16 transformed partners per step, keeping the forward batch size at 32. For each partner, draw independent Haar matrices $A \in O(n)$ and $Q \in O(d)$ and transform the context and every query by

$$(X, y, x_\star) \longmapsto (AXQ, Ay, Q^\top x_\star).$$

Both views retain the same noiseless target. Let p_{ij}, p'_{ij} be their predictions on query j of prompt i, whose target and query scale are t_{ij}, a_{ij}. With $B = 16$ and $m = 8$, the loss is

$$\frac{1}{2Bm} \sum_{i,j} \frac{(p_{ij} - t_{ij})^2 + (p'_{ij} - t_{ij})^2}{da_{ij}^2} + \frac{3}{4Bm} \sum_{i,j} \frac{(p_{ij} - p'_{ij})^2}{da_{ij}^2}. \tag{12}$$

This intervention penalises a combined covariate/context transformation. It does not separately penalise orthogonal noise changes or changes in context count.

B SPECTRAL EVALUATION AND CURRICULUM COMPARISONS

B.1 REFERENCE EVALUATION AND CONSTRAINED SPECTRAL OPTIMISATION

Reference evaluation. The spectral maps evaluate 201 feasible pairs of realised scale and effective dimension, with $n = 80$. At each pair we draw 4,096 spectra from Eq. (6). Each spectrum uses 256 draws of the remaining context variables and eight queries per draw, sampled as in Section 2.2. We pool squared errors and squared ridge predictions within a spectrum, then take the median of their ratios across spectra. Within each comparison, the models receive identical prompts. Scale brackets in the figures mark the sampled training scale $s_0 = s(\bar{M})$.

Objective and constraints. We maximise relative risk averaged over the remaining prompt variables, holding (s, d_{eff}) fixed. At rank k, apply Eq. (6) to the k positive eigenvalues, with $q \in \mathbb{R}^k$, $\mathbf{1}^\top q = 0$ and $\|q\|_2 = 1$, and set the other $d - k$ eigenvalues to zero. The softmax parametrisation fixes $\sum_i \lambda_i = n_0 ds$. Solving for α fixes $\sum_i \lambda_i/(1+\lambda_i) = d_{\text{eff}}$, and we differentiate through this constraint by implicit differentiation.

At rank k, the context construction uses the first k columns of Haar $U \in O(d)$, an independent uniform $n \times k$ orthonormal frame V, and $z \sim \mathcal{N}(0, I_k)$, with $\varepsilon_\perp = (I - VV^\top)\eta$ for independent $\eta \sim \mathcal{N}(0, \sigma^2 I_n)$. For each model, cell and rank, draw $B = 128$ sets of these variables and $m = 8$ ordinary queries per set, and hold the draws fixed throughout optimisation. Let $Z_{\ell j}(q)$ be the resulting prompt for draw ℓ and query j. The ascent objective is

$$\frac{1}{Bm\,\mathbb{E}[a^2]d_{\text{eff}}} \sum_{\ell=1}^{B} \sum_{j=1}^{m} \big[f_\theta(Z_{\ell j}(q)) - f_R(Z_{\ell j}(q))\big]^2. \tag{13}$$

Both predictors are differentiated with respect to the spectrum. Under the response and query laws of Section 2, $\mathbb{E}[f_R^2 \mid \Lambda] = \mathbb{E}[a^2]d_{\text{eff}}$, so this analytic denominator is constant within the constrained search. Final risk estimates use the ratio of empirical averages described in Section 2.3.

Multi-start gradient ascent. We rank the $4{,}096$ full-rank reference spectra by the same objective, using 256 context draws, and initialise ascent at the 16 highest-scoring spectra. For each feasible rank $k < d$, we screen 128 random spectra with 128 context draws and initialise ascent at the highest-scoring one. A unique flat or rank-two spectrum is screened once; flat spectra require no optimisation. Each update projects the gradient onto $\{v : \mathbf{1}^\top v = q^\top v = 0\}$, takes an Adam step with learning rate 0.04, coefficients $(0.9, 0.999)$ and numerical constant 10^{-8}, then centres, normalises and sorts q. We retain the highest-scoring iterate. Runs use at most 100 steps at full rank and 75 at lower rank, stopping after 10 consecutive steps without a relative improvement greater than 10^{-6}. Final spectra satisfy both constraints to tolerance 10^{-8}. Eight full-rank descent starts and one descent start per non-flat boundary stratum additionally explore the lower-risk range.

Independent optimisation check. At 24 prespecified cells per model, we also run a derivative-free ascent with independent context and query draws. It starts from the best screened spectrum at full rank and at the three boundary ranks with the highest screened scores, repeating the best boundary rank if fewer than three are available. Each proposal adds a random unit tangent displacement to q and then centres and normalises it; only improvements are accepted. The displacement magnitude starts at 0.2, increases by a factor 1.05 after acceptance (up to 0.5), and decreases by 0.92 otherwise (down to 10^{-4}). Proposal signs alternate. The budgets are 100 proposals at full rank and 75 at each selected boundary rank.

Candidate selection and fresh validation. The saved candidate set contains the five highest-scoring full-rank reference spectra, five highest-scoring spectra from the boundary screens and optimisations, three highest-scoring full-rank ascent results, and two lowest-scoring spectra from these primary searches. We add any derivative-free result whose score exceeds the best primary optimised score, then deduplicate spectra. The searches for C_0 and C_D are independent.

After freezing these sets, we re-evaluate the full $4{,}096$-spectrum reference bank and saved candidates using $1{,}024$ fresh context draws with eight queries each. These draws are independent of every search stage and shared across the two models. Write $R_{\max}$ for the largest validated risk over a model's reference/candidate union; the matching R_{50} uses the reference spectra under the same fresh validation law. Figure 4 shows both $R_{\max}/R_{50}$ and $R_{\max}$; the former reaches 3.2555 for C_0. Multiple starts, explicit boundary searches and independent validation strengthen the estimate, but do not certify its distance to the global maximum. The maximum is over spectrum-averaged risks, not individual-prompt errors.

B.2 Two-level spectra and their training exposure

For a two-level full-rank spectrum with multiplicities m and $d - m$, write its eigenvalues as $\lambda_H > \lambda_L > 0$ and gains as $h_H = \lambda_H/(1+\lambda_H)$ and $h_L = \lambda_L/(1+\lambda_L)$. The family $\mathcal{H}$ consists of such spectra with $h_H \geq 0.95$ and $0.10 \leq h_L \leq 0.80$. These thresholds were motivated by earlier

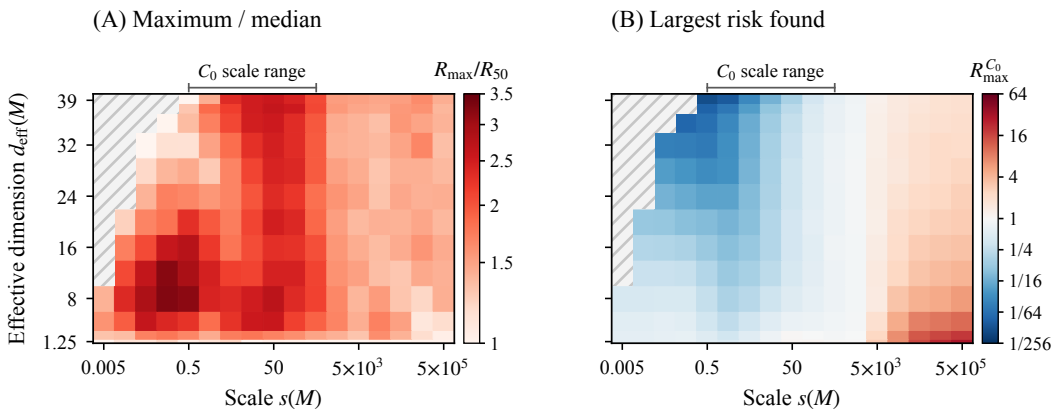


Figure 4: **Variation across spectra at fixed scale and effective dimension.** (A) Estimated maximum relative risk $R_{\max}$ divided by the reference median R_{50}. (B) The same $R_{\max}$ without division by the median. The candidate set includes 4,096 reference spectra and the constrained optimisation results from Appendix B.1, including rank-deficient spectra. Both statistics use fresh common validation draws, with 1,024 prompts per spectrum. The maximum concerns risks averaged within each spectrum; it is neither a bound over all spectra nor a maximum over individual prompts. Brackets show the training scale range in s_0.

exploratory evaluations. The present bank fixes 849 spectra across the 49 central cells, with between four and 31 selected multiplicities per cell, without using the current models' errors. Each spectrum is evaluated with 1,024 context draws and eight queries per draw. Table 1 compares the median risk over these spectra with the C_0 reference-law median from the paired curriculum evaluation in Figure 5(A). That reference median uses 4,096 spectra and 256 context draws per spectrum, with different prompt draws from Figure 1(A). The ratios compare two spectral populations at matched coordinates, not paired spectra.

We also compare the median risk on $\mathcal{H}$ with the estimated maximum $R_{\max}$ at the same coordinates. Across the 49 evaluated points, the median ratio is 0.903, with range $[0.671, 1.078]$. The two estimates use separate evaluation draws, and the $\mathcal{H}$ spectra were not included in the search candidate sets. At seven points, the estimated median on $\mathcal{H}$ exceeds $R_{\max}$; this may reflect sampling variation, spectra missed by the search, or both.

To assess exposure during training, we draw 200,000 reference covariances from the C_0 sampler and apply the same gain criterion to $\bar{M} = 80\Sigma_{\text{ref}}$. Of these draws, 51,606 lie in $0.5 \le s(\bar{M}) \le 500$, $8 \le d_{\text{eff}}(\bar{M}) \le 32$, including 14,266 step-family draws. The proportions in $\mathcal{H}$ are 98.7% among these step draws, 27.3% among all draws in the band, and 8.8% across the full sampler. Exact two-level membership excludes the smooth family. Sampling a finite context changes its eigenvalues, so these proportions describe expected Gram matrices rather than realised training contexts.

B.3 Effects of changing the curriculum

Figure 5(A) gives the complete C_D/C_0 median-risk comparison. At $d_{\text{eff}} \le 2$ and $0.5 \le s \le 5{,}000$, all 18 evaluated cells improve, with median ratio 0.600 and range $[0.417, 0.811]$. Across the 49 central cells, 33 improve, with median 0.978 and range $[0.822, 1.204]$. At $s \ge 50{,}000$, 38 of 39 cells worsen, with median 1.334 and range $[0.911, 1.903]$.

Figure 5(B) compares maximum-to-median ratios, while Figure 6(A) compares the estimated maximum risks themselves. For $d_{\text{eff}} \le 2$ and $0.5 \le s \le 5{,}000$, the maximum decreases at all 18 evaluated points, with median ratio $R_{\max}^{C_D}/R_{\max}^{C_0} = 0.597$. In the region $8 \le d_{\text{eff}} \le 32$ and $0.5 \le s \le 500$, it increases at 32 of 49 points, with median ratio 1.077. At $s \ge 50{,}000$, it increases at all 39 evaluated points, with median ratio 1.281. Across the complete grid, 103 maxima decrease and 98 increase; the median ratio is 0.993.

Table 1: Median risk on $\mathcal{H}$ divided by the reference-law median for C_0, at matched realised coordinates. Columns give d_{eff}. The median of the 49 cellwise ratios is 1.96; 48 exceed one. The two medians use different spectral banks and evaluation budgets, as described in Appendix B.2.

$s(M)$	8	12	16	20	24	28	32
0.5	2.88	2.47	2.67	2.05	2.00	1.25	1.24
1.58	2.82	2.11	1.94	1.61	1.71	1.16	0.94
5	2.31	1.99	1.96	1.62	1.71	1.28	1.29
15.8	2.64	2.25	2.10	1.88	1.77	1.68	1.59
50	2.49	2.32	2.32	2.22	2.20	2.26	1.98
158	2.09	2.10	2.29	2.24	1.75	1.83	1.70
500	2.24	1.94	1.67	1.45	1.40	1.30	1.33

The symmetry intervention also has a mixed effect on agreement with ridge (Figure 7). Across all 201 cells, 108 median risks improve and 93 worsen. In the central band the median ratio is 1.016, with 24 of 49 cells improving. At $s \geq 50{,}000$, 32 of 39 worsen and the median ratio is 1.190. Reduced representation sensitivity therefore need not coincide with improved prediction accuracy at the same training budget.

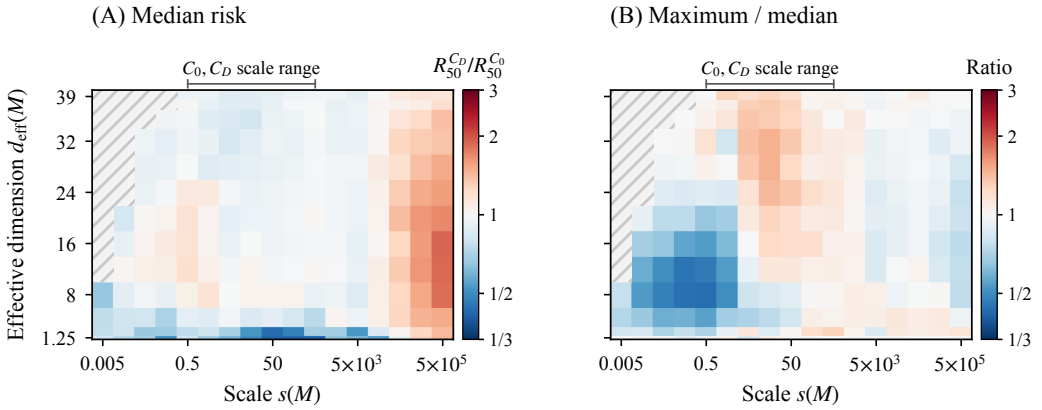


Figure 5: **Effect of a broader spectral curriculum.** (A) Ratio of median relative risks for C_D and C_0, using the same reference spectra and 256 prompt draws per spectrum. (B) Ratio of their maximum-to-median risk ratios, $(R_{\max}/R_{50})^{C_D}/(R_{\max}/R_{50})^{C_0}$. The searches are performed separately for each model and validated with 1,024 fresh common prompt draws per spectrum. C_D changes spectral shape while preserving the scale and effective dimension of the expected Gram matrix M.

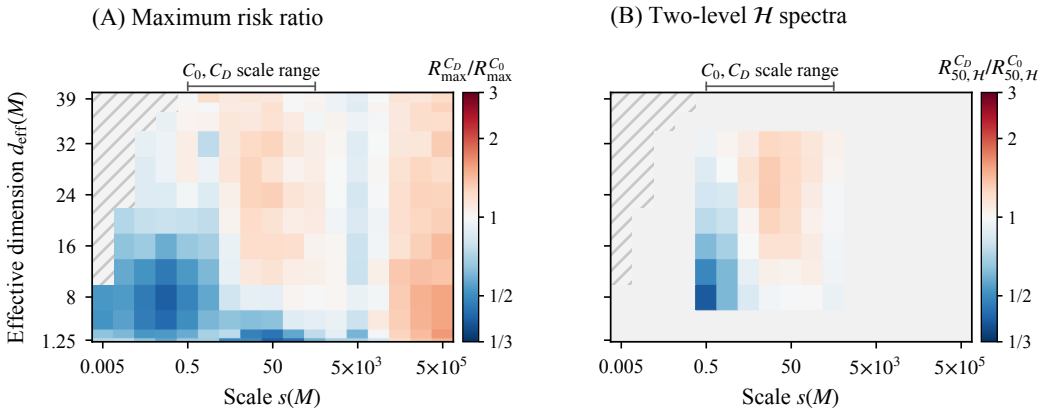


Figure 6: **Maximum risk and agreement on two-level spectra.** (A) Ratio of the largest risks found for C_D and C_0, using the validation in Figure 5(B). (B) Ratio of median risks on $\mathcal{H}$. These spectra are selected independently of model errors, within $0.5 \leq s \leq 500$ and $8 \leq d_{\text{eff}} \leq 32$, and evaluated with 1,024 prompts per spectrum. Grey tiles are outside this evaluation set; hatched tiles are infeasible.

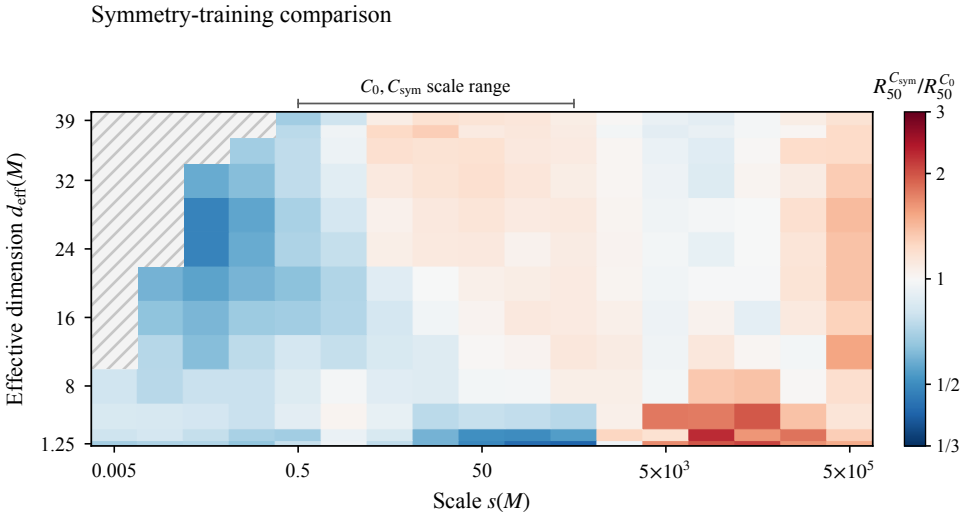


Figure 7: **Effect of symmetry training on agreement with ridge.** Ratio $R^{C_{\text{sym}}}_{50}/R^{C_0}_{50}$ on the same 201 evaluated scale/effective-dimension pairs, with 4,096 spectra per pair and 256 matched prompt draws per spectrum.

C EQUIVALENT REPRESENTATIONS AND CONTEXT COUNT

Evaluation states. We use two fixed banks of 64 spectra within $0.5 \le s \le 500$ and $8 \le d_{\text{eff}} \le 32$. The reference bank is sampled uniformly without replacement from the pool of $49 \times 4{,}096$ reference spectra at the 49 evaluated coordinate pairs in this region. The $\mathcal{H}$ bank is sampled uniformly without replacement from its 849 spectra. These selections are fixed before the representation evaluations. They occupy 35 and 33 coordinate pairs, respectively. Each selected spectrum receives equal weight; the coordinate pairs need not have equal frequencies in either sample.

For each spectrum, we draw 16 independent base contexts at $n = 80$ using the construction in Section 2.1, together with eight queries per context. Thus each bank contains 1,024 states. The same states and queries are used for C_0 and C_{sym} in the transformation and context-count experiments. Within each bank, Z in Eq. (8) is uniform over these states and their eight queries. The denominator is the mean of the 8,192 squared ridge predictions, shared by all models and transformations on that bank.

Transformations and percentiles. The three transformation laws in Figure 2(A) are

$$\begin{aligned}
&(X, y, x_\star) \mapsto (XQ, y, Q^\top x_\star), && Q \sim \mathrm{Haar}(O(d)),\\
&(X, y, x_\star) \mapsto (AX, Ay, x_\star), && A \sim \mathrm{Haar}(O(n)),\\
&(X, y, x_\star) \mapsto (X, VV^\top y + (I - VV^\top)\eta, x_\star), && \eta \sim \mathcal{N}(0, \sigma^2 I_n).
\end{aligned}$$

Every transformation preserves f_R exactly.

For each state and law, draw 64 independent pairs of transformed contexts, with eight queries per pair. Conditional independence gives

$$\tfrac{1}{2}\mathbb{E}[(f_\theta(Z') - f_\theta(Z''))^2 \mid Z] = \mathrm{Var}(f_\theta(Z') \mid Z).$$

Consequently, D_G is an average conditional variance divided by the ridge second moment. The 90th and 99th percentiles are computed over all $1{,}024 \times 64 \times 8$ individual squared differences divided by $2\mathbb{E}_Z[f_R^2]$, before averaging within states. The mean of these normalised differences estimates D_G.

Figures 8 and 9 report means and individual percentiles on both banks. In units of the bank's RMS ridge prediction, the RMS change is $\sqrt{2D_G}$. For reference-spectrum covariate rotations, it falls from 0.262 to 0.192 under symmetry training, while the 99th percentile of the absolute change falls from 0.871 to 0.631. These are changes between equivalent representations, not errors against the target.

Uncertainty of individual percentiles. Table 2 gives 95% intervals from 2,000 hierarchical bootstrap replicates. Within each of the 64 fixed spectra, we resample the 16 states, then resample the 64 pairs within each selected state while retaining the eight-query blocks. Model comparisons use the same bootstrap weights, and each replicate recomputes the common ridge denominator. This accounts for dependence within states and pairs, conditional on the selected spectra. The mean-only intervals in Figure 8 use 5,000 bootstrap replicates that resample whole states.

Changing the number of context points. For each full-rank state, hold $U, \{\lambda_i\}, z$ and all eight queries fixed. At count n, sample V_n uniformly on $\{V \in \mathbb{R}^{n\times d} : V^\top V = I_d\}$ and form X_n, y_n using Eqs. (3)–(4), with $\varepsilon_\perp = (I - V_n V_n^\top)\eta_n$ and $\eta_n \sim \mathcal{N}(0, \sigma^2 I_n)$. This preserves M, b and the ridge predictions across counts. We evaluate

$$n \in \{40, 48, 57, 67, 80, 95, 113, 135, 160, 190, 226, 269, 320, 381, 453, 538, 640\}.$$

There are 128 independent replicates per state. Within a replicate, counts are coupled: take the first n rows of one $640 \times d$ standard Gaussian matrix and orthonormalise by QR with positive diagonal, and take the first n entries of one shared Gaussian noise vector before projecting. The same replicates are used for both models.

We average predictions over representations, centre on their mean over the four training counts, square, and average over the eight queries, as in Eq. (9). Means and percentiles are then taken over the 1,024 states. These summaries describe changes in the estimated representation-averaged prediction across counts, whereas Figure 2(A) describes individual transformed-prompt pairs. The bootstrap holds these averages fixed, so its intervals exclude uncertainty from the 128 sampled representations.

Figure 10 gives mean count discrepancies on both banks. The 99th percentile at $n = 640$ on the reference bank is 0.301 for C_0 (interval $[0.266, 0.346]$) and 0.300 for C_{sym} ($[0.250, 0.346]$). These intervals use 3,000 bootstrap replicates, resampling states within each fixed spectrum.

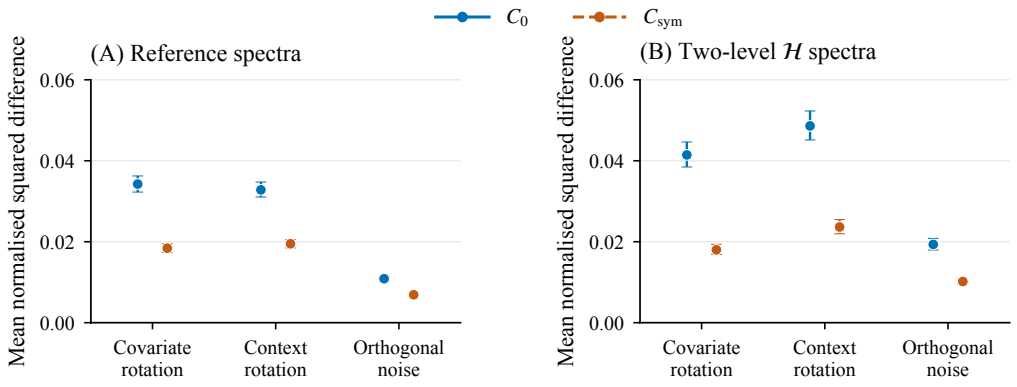


Figure 8: **Mean sensitivity to context representation.** Normalised mean squared discrepancies D_G for C_0 and C_{sym} on the reference-spectrum and two-level $\mathcal{H}$ evaluation sets. Each set contains 1,024 states, with 64 independent transformation pairs and eight target queries per state. Error bars are 95% state-bootstrap intervals.

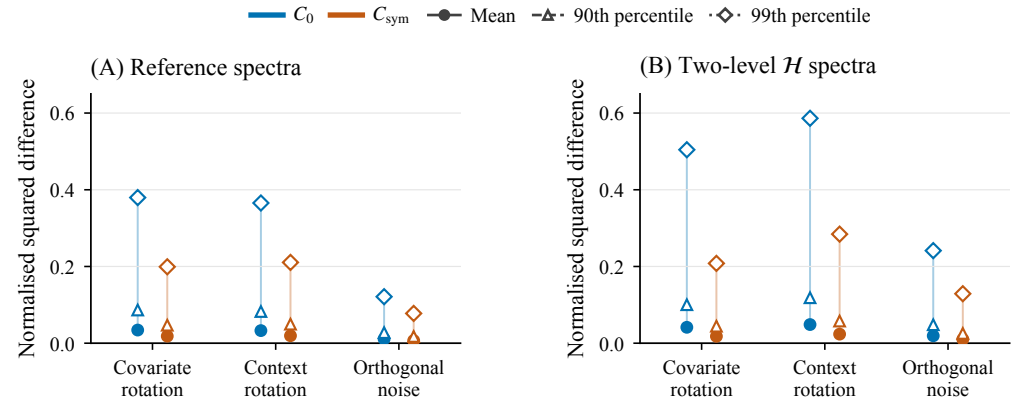


Figure 9: **Individual discrepancies between equivalent representations.** Mean, 90th and 99th percentiles of individual paired squared prediction differences on both evaluation sets, using 1,024 states, 64 transformation pairs per state and eight targets per pair. Each difference is normalised by twice the pooled mean squared ridge prediction in the corresponding set. Means reproduce D_G in Figure 8; percentiles are computed before averaging within states.

Table 2: Individual paired squared prediction differences divided by twice the bank-wide ridge second moment. The bootstrap retains state and pair dependence and conditions on the fixed spectra.

Bank	Transformation	Model	Mean	99th percentile (95% interval)
Reference	Covariate rotation	C_0	0.0342	0.380 [0.362, 0.401]
Reference	Covariate rotation	C_{sym}	0.0184	0.199 [0.189, 0.209]
Reference	Context rotation	C_0	0.0328	0.365 [0.348, 0.385]
Reference	Context rotation	C_{sym}	0.0195	0.211 [0.200, 0.222]
Reference	Orthogonal noise	C_0	0.0109	0.121 [0.115, 0.128]
Reference	Orthogonal noise	C_{sym}	0.0069	0.078 [0.073, 0.082]
$\mathcal{H}$	Covariate rotation	C_0	0.0414	0.504 [0.474, 0.534]
$\mathcal{H}$	Covariate rotation	C_{sym}	0.0180	0.208 [0.196, 0.220]
$\mathcal{H}$	Context rotation	C_0	0.0486	0.586 [0.552, 0.623]
$\mathcal{H}$	Context rotation	C_{sym}	0.0236	0.284 [0.269, 0.301]
$\mathcal{H}$	Orthogonal noise	C_0	0.0193	0.241 [0.226, 0.255]
$\mathcal{H}$	Orthogonal noise	C_{sym}	0.0102	0.129 [0.122, 0.136]

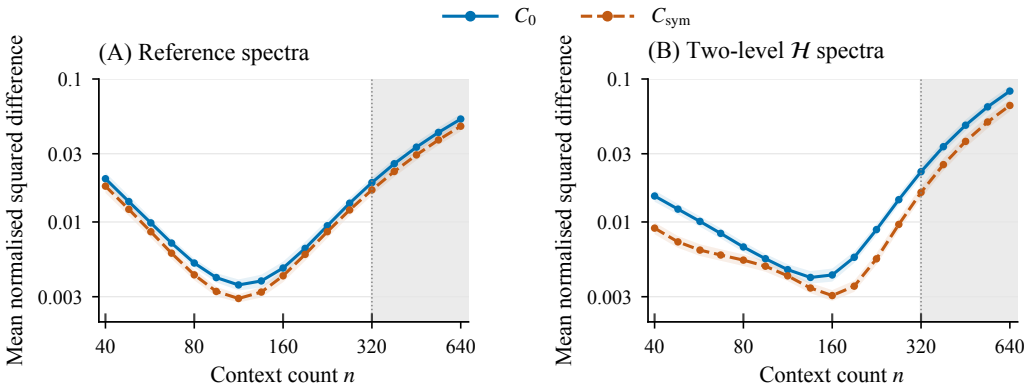


Figure 10: **Changing context count while holding ridge predictions fixed.** For each state, M, b and the target queries are fixed across counts. Predictions are averaged over 128 representations at each count, then compared with their mean over $n \in \{40, 80, 160, 320\}$. The pooled squared difference is normalised by the pooled mean squared ridge prediction. Ticks mark training counts; shading marks $n > 320$. Bands are 95% state-bootstrap intervals. Both panels use the same logarithmic vertical scale.

D Conditional response and query experiments

Spectra and contexts. Both experiments evaluate C_0 at $n = 80$ on the 64 reference spectra and 64 spectra from $\mathcal{H}$ specified in Appendix C. Figure 3 shows the reference bank, and this appendix reports both banks. Their empirical distributions of scale and effective dimension differ, so comparing the aggregate curves does not isolate spectral shape at matched coordinates.

Response conditioning and normalisation. For each spectrum, draw 256 uniform unit directions v and set $z = r_z v$. At each direction and radius, evaluate 64 independent draws of $U, V, \varepsilon_\perp$ and eight ordinary queries per draw. Within each spectrum, these draws are shared across response directions and radii, and each direction is reused at every radius. For each spectrum–direction pair, we average squared prediction discrepancies and squared ridge predictions over these draws. The two normalisations below use these same averages.

Conditional on $X, y, x_\star$, the noiseless target has posterior mean f_R and variance $x_\star^\top (I + M/\sigma^2)^{-1} x_\star$. With all following expectations conditional on Λ, z,

$$\mathbb{E}[(f_\theta - y_\star)^2] - \mathbb{E}[(f_R - y_\star)^2] = \mathbb{E}[(f_\theta - f_R)^2], \tag{14}$$

$$\mathbb{E}[y_\star^2] = \mathbb{E}[f_R^2] + \mathbb{E}[a^2](d - d_{\text{eff}}), \tag{15}$$

$$\mathbb{E}[a^2] = \frac{4 - 1/4}{4 \log 2}. \tag{16}$$

For Eq. (10), we estimate the ridge second moment on the same draws as the numerator and add the analytic posterior-variance term. The effective dimension in this term is that of the individual spectrum. At each radius, we compute the 50th, 90th and 99th percentiles over the 256 response directions within each spectrum, then report the median of each percentile across the 64 spectra in that bank, weighting spectra equally.

Figure 11 uses the same estimates with the ridge second moment alone in the denominator. Figure 13 reports the target-normalised quantity over the full radius range. The latter remains well behaved when the ridge signal tends to zero at small response radii. It measures excess error over ridge, not total prediction error.

Query calibration and evaluation. We reuse the 16 fixed context/response states per spectrum from the representation experiments, giving 1,024 contexts per bank. In each context draw an independent Haar orthonormal basis $\{\phi_i\}_{i=1}^d$ and define

$$\widehat{w}_\theta(X, y) = \frac{1}{2r_0} \sum_{i=1}^{d} \big[f_\theta(X, y, r_0 \phi_i) - f_\theta(X, y, -r_0 \phi_i) \big] \phi_i, \tag{17}$$

where $r_0 = 6.24534$ is the median of the pretraining query-radius distribution $a\chi_{40}$. The sampled basis is implicit in this notation. This construction recovers the coefficient exactly for any predictor linear in the query. Holding the coefficient fixed, we evaluate 1,024 fresh uniform unit directions in each context, using an independent random stream and reusing the directions across radii. Since

$$\mathbb{E}_u[(r u^\top \widehat{w}_\theta)^2] = r^2 \|\widehat{w}_\theta\|_2^2 / d,$$

the denominator in Eq. (11) is the square root of the calibrated predictor's mean squared output over directions. We compute directional percentiles within each context, then take the median of each percentile across the 1,024 contexts in that bank. Figure 12 reports both banks over the full radius range.

On $\mathcal{H}$, the median across contexts of the directional 90th percentile rises from 0.114 at r_0 to 0.382 at $4r_0$. For both banks, the corresponding summary increases between these radii within every spectrum.

Radius ranges and sampling. Both experiments are plotted at normalised radii $2^{k/4}$ for $k = -16, \ldots, 16$. The main figure shows the 17 radii in $[1/4, 4]$; the appendix shows all 33 radii in $[1/16, 16]$. All plotted points are measured values, connected by straight segments. Percentiles use linear interpolation between adjacent order statistics.

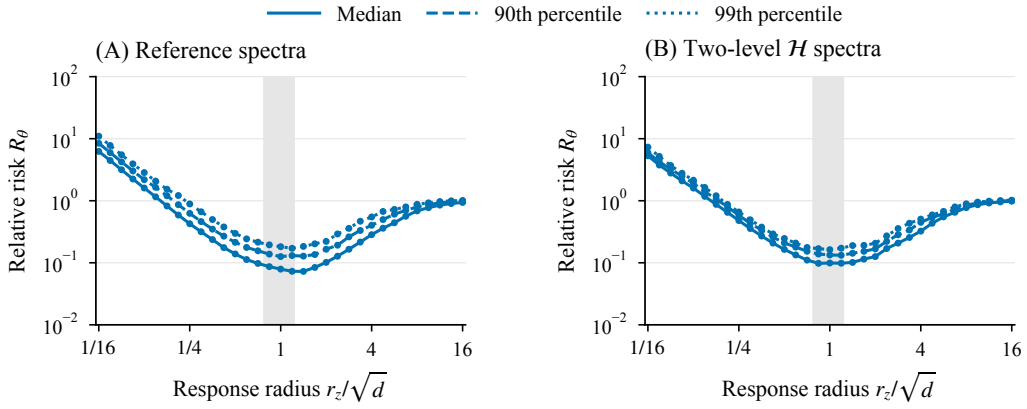


Figure 11: **Relative risk conditional on informative response.** Curves show medians across spectra of percentiles over 256 response directions within each spectrum, with ratios formed after averaging representations and queries within each spectrum–direction pair. The denominator is the conditional ridge second moment. Shading marks the central 95% of the χ_{40} response-radius distribution. Both axes are logarithmic.

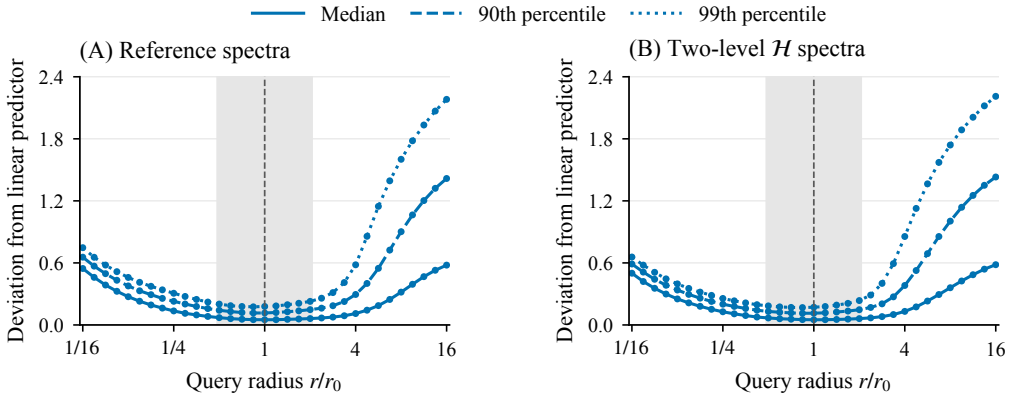


Figure 12: **Query dependence over the full measured radius range.** For each context, the calibrated coefficient is fixed across radii; directional percentiles use 1,024 fresh query directions. Curves show the median of each percentile across contexts. Shading marks the central 95% of pretraining query radii; the dashed line marks calibration at r_0.

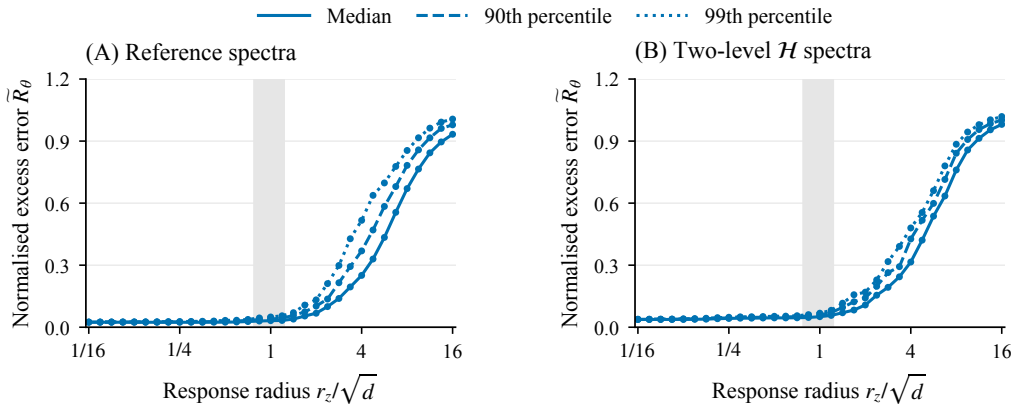


Figure 13: **Normalised excess error over the full response-radius range.** Medians across spectra of within-spectrum directional percentiles of Eq. (10) for the reference and $\mathcal{H}$ banks, each with 64 spectra and 256 directions per spectrum. Each conditional ratio uses the target second moment, obtained by adding $\mathbb{E}[a^2](d - d_{\text{eff}})$ to the ridge second moment at that spectrum and response direction. Shading marks the central 95% of pretraining response radii.

E Enforcing radial affinity

Architecture. We modify the reference Transformer to satisfy (R). For a fixed context (X, y), write $x_\star = ru$ with $r > 0$ and $\|u\|_2 = 1$. The Transformer receives the unchanged context and the query $r_0 u$, where $r_0 = 6.24534$ is the median pretraining query radius used in Appendix D. Two affine scalar heads on the final target representation produce $c_\theta(X, y, u)$ and $v_\theta(X, y, u)$, and the prediction is

$$f_\theta(X, y, ru) = c_\theta(X, y, u) + \left(\frac{r}{r_0} - 1\right) v_\theta(X, y, u). \tag{18}$$

Neither head receives r, so (R) holds for every parameter value with $\alpha_{X,y}(u) = v_\theta(X, y, u)/r_0$ and $\beta_{X,y}(u) = c_\theta(X, y, u) - v_\theta(X, y, u)$. The final LayerNorm is retained, but multiplication by the original radius after the readout removes the boundedness restriction of Proposition 1. Radial affinity alone does not impose linearity in the query or invariance to equivalent context representations. For $x_\star = 0$, the Transformer receives a zero query and the prediction is $c_\theta - v_\theta$, evaluated at that input. Continuity at zero is not imposed.

Training and implementation check. We denote the variant by $C_0^{(R)}$. Apart from the query normalisation and readout in Eq. (18), it uses the architecture and C_0 training procedure in Appendix A, including the curriculum, loss, optimiser, batch sizes and random streams for the training prompts. The additional head adds 129 parameters, giving 1,061,762 in total. Shared parameters have the same initial values as C_0 for the same seed. The new head initially copies the original head, so $c_\theta = v_\theta$ and the initial prediction is positively homogeneous along each ray. We train from scratch for 500,000 updates with seed 0, without a symmetry penalty. The reference radius is fixed before training.

For a numerical check on the trained model, we use 32 fixed context/response states from Appendix C, equally divided between the reference and $\mathcal{H}$ banks, with eight query directions per state. We evaluate $r/r_0 \in \{1/4, 1, 4, 16\}$ and compare the middle prediction with linear interpolation between the endpoints of each successive triple. All 512 interpolation residuals satisfy $|\delta| \le 2 \times 10^{-5}(1 + \max_i |f_i|)$, where the maximum is over the three predictions in each triple. The largest absolute residual in single precision is 3.05×10^{-6}.

Context scale. We evaluate both models on the same 201 feasible scale/effective-dimension pairs and 4,096 reference spectra per pair. Each spectrum uses 256 matched prompt draws at $n = 80$, with eight queries per context drawn from the pretraining query law. We pool squared prediction errors against ridge and squared ridge predictions within each spectrum, then take the median of their ratios

as in Eq. (7). Figure 14 shows $R_{50}^{C_0^{(R)}}$ and $R_{50}^{C_0^{(R)}}/R_{50}^{C_0}$. At $s = 500{,}000$ and $d_{\text{eff}} = 20$, the estimated median risks are 1.6352 for $C_0^{(R)}$ and 2.1164 for C_0. The intervention improves agreement at this point, but substantial error remains. The C_0 value differs from Figure 1(A) because this comparison uses fresh prompt draws for the same reference spectra.

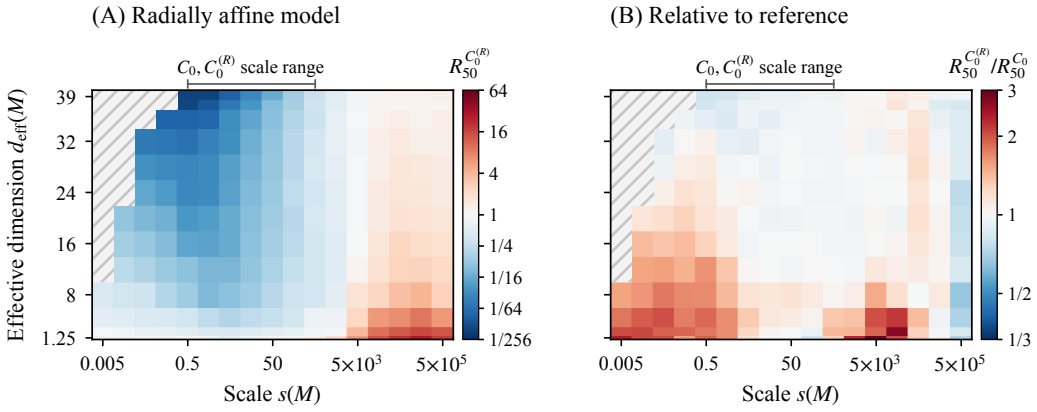


Figure 14: **Context-scale error after enforcing (R).** (A) Estimated median relative risk of $C_0^{(R)}$. (B) Ratio $R_{50}^{C_0^{(R)}}/R_{50}^{C_0}$ on matched prompts. Brackets mark sampled training scales s_0. Hatching marks infeasible coordinates. Colour scales are logarithmic.

Equivalent representations. We use the same fixed states and transformations as Appendix C, with 64 independent pairs per state for each transformation. The two models receive identical pairs. Table 3 reports D_G from Eq. (8), separately for the reference and $\mathcal{H}$ banks. As for the mean-only intervals in Figure 8, we resample whole states, using $5{,}000$ bootstrap replicates.

Table 3: Mean normalised squared differences D_G after enforcing (R), with 95% bootstrap intervals. Each bank contains $1{,}024$ states from 64 spectra.

Bank	Transformation	C_0	$C_0^{(R)}$
Reference	Covariate rotation	$0.0342\ [0.0323, 0.0362]$	$0.0383\ [0.0360, 0.0406]$
Reference	Context rotation	$0.0328\ [0.0310, 0.0347]$	$0.0351\ [0.0332, 0.0372]$
Reference	Orthogonal noise	$0.0109\ [0.0102, 0.0116]$	$0.0122\ [0.0115, 0.0130]$
$\mathcal{H}$	Covariate rotation	$0.0414\ [0.0385, 0.0446]$	$0.0441\ [0.0409, 0.0476]$
$\mathcal{H}$	Context rotation	$0.0486\ [0.0451, 0.0523]$	$0.0519\ [0.0481, 0.0561]$
$\mathcal{H}$	Orthogonal noise	$0.0193\ [0.0180, 0.0208]$	$0.0213\ [0.0197, 0.0230]$

For $C_0^{(R)}$, the six mean discrepancies range from 0.0122 to 0.0519. Enforcing (R) therefore leaves both substantial context-scale error and sensitivity to equivalent representations in this model.

F Proofs of the theoretical results

F.1 The restriction on radial affinity

Proof of Proposition 1. Let p be the residual width, and $h \in \mathbb{R}^p$ be the final residual vector at the target token. Write the final LayerNorm as

$$\mathrm{LN}(h) = \gamma \odot \frac{h - \bar{h}\mathbf{1}}{\sqrt{p^{-1}\|h - \bar{h}\mathbf{1}\|_2^2 + \epsilon}} + \delta, \qquad \bar{h} = p^{-1}\mathbf{1}^\top h, \quad \epsilon > 0,$$

where $\gamma, \delta \in \mathbb{R}^p$ are its learned scale and shift, and $\odot$ denotes elementwise multiplication.

$$\left\| \frac{h - \bar{h}\mathbf{1}}{\sqrt{p^{-1}\|h - \bar{h}\mathbf{1}\|_2^2 + \epsilon}} \right\|_2^2 = \frac{\|h - \bar{h}\mathbf{1}\|_2^2}{p^{-1}\|h - \bar{h}\mathbf{1}\|_2^2 + \epsilon} \le p,$$

so if the final readout is $\ell^\top \mathrm{LN}(h) + c$, then

$$|f_\theta(X, y, x_\star)| \le \|\ell\|_2 \big(\sqrt{p}\,\|\gamma\|_\infty + \|\delta\|_2\big) + |c| =: C_\theta < \infty.$$

This bound is independent of the query. For a fixed nonempty context, the prediction is also continuous in the query since affine maps, attention with its fixed mask, GELU and LayerNorm with positive epsilon are continuous.

Suppose condition (R) holds. For each unit vector u, boundedness of $r\alpha_{X,y}(u) + \beta_{X,y}(u)$ as $r \to \infty$ forces $\alpha_{X,y}(u) = 0$. Continuity at the zero query then gives

$$\beta_{X,y}(u) = \lim_{r \downarrow 0} f_\theta(X, y, ru) = f_\theta(X, y, 0),$$

independently of u. Thus the predictor is independent of the query. Conversely, a query-independent predictor satisfies (R) with $\alpha_{X,y}(u) = 0$ and $\beta_{X,y}(u) = f_\theta(X, y, 0)$.

Finally, for every unit vector u,

$$\frac{|f_\theta(X, y, ru) - f_R(X, y, ru)|}{r} = \left| \frac{f_\theta(X, y, ru)}{r} - u^\top w_R \right| \longrightarrow |u^\top w_R|,$$

since $|f_\theta(X, y, ru)| \le C_\theta$. If $u^\top w_R \neq 0$, the limiting slope is positive, as claimed. ∎

F.2 Uniform accuracy of inverse iteration

Proof of Theorem 1. Write $G = X_0^\top X_0 \succ 0$ and $A_t = I_d + tG$. Let $\mu_i > 0$ be the eigenvalues of G, so $a_i = 1 + t\mu_i$ are those of A_t, and write

$$L_t = \|A_t^2\|_F = \left(\sum_{i=1}^d a_i^4 \right)^{1/2}.$$

The initial matrix $B_0 = A_t / L_t$ commutes with A_t, and so does every subsequent iterate. For $E_k = I_d - B_k A_t$, the inverse-iteration update gives

$$E_{k+1} = E_k^2, \qquad E_K = \left(I_d - \frac{A_t^2}{L_t} \right)^{2^K}, \qquad \widehat{w}_K - w_R = -E_K w_R.$$

Under the stated regression law,

$$b_t = tG\beta + \sqrt{t} X_0^\top \varepsilon, \qquad w_R = A_t^{-1} b_t, \qquad \mathbb{E}[w_R] = 0, \qquad \mathrm{Cov}(w_R) = tG(I_d + tG)^{-1}.$$

The independent query has second-moment matrix $\mathbb{E}[x_\star x_\star^\top] = \mathbb{E}[a^2] I_d$, so its common factor $\mathbb{E}[a^2]$ cancels from the relative risk. Diagonalising in the eigenbasis of G therefore yields

$$R_{\mathcal{A}_K}(e_t) = \frac{\sum_{i=1}^d h_i (1 - v_i)^{2^{K+1}}}{\sum_{i=1}^d h_i}, \qquad h_i = \frac{t\mu_i}{1 + t\mu_i}, \quad v_i = \frac{a_i^2}{L_t}. \tag{19}$$

Here $h_i > 0$, $0 < v_i \le 1$ and $\sum_i v_i^2 = 1$.

Order the μ_i increasingly. Then h_i increases and $(1 - v_i)^2$ decreases. In particular,

$$d \sum_i h_i (1 - v_i)^2 - \left(\sum_i h_i \right) \left(\sum_i (1 - v_i)^2 \right) = \sum_{i<j} (h_i - h_j)\big((1 - v_i)^2 - (1 - v_j)^2 \big) \le 0.$$

Since $2^{K+1} \ge 2$ and $0 \le 1 - v_i \le 1$, Eq. (19) implies

$$R_{\mathcal{A}_K}(e_t) \le \frac{\sum_i h_i (1 - v_i)^2}{\sum_i h_i} \le \frac{1}{d} \sum_i (1 - v_i)^2 = 1 - \frac{2}{d} \sum_i v_i + \frac{1}{d} \le 1 - \frac{1}{d},$$

where the last inequality uses $\sum_i v_i \geq (\sum_i v_i^2)^{1/2} = 1$.

For the second bound, $L_t \leq \sqrt{d}\, a_{\max}^2$ and

$$\frac{a_{\max}}{a_{\min}} = \frac{1 + t\mu_{\max}}{1 + t\mu_{\min}} \leq \frac{\mu_{\max}}{\mu_{\min}} = \kappa_0.$$

Consequently $v_i \geq 1/(\sqrt{d}\,\kappa_0^2)$ for every i. Applying $1 - v \leq e^{-v}$ termwise in Eq. (19) gives

$$R_{\mathcal{A}_K}(e_t) \leq \exp\left[-\frac{2^{K+1}}{\sqrt{d}\,\kappa_0^2}\right].$$

Both bounds hold for every $t \geq 1$ with the same K, proving the stated bound on the supremum. ∎

F.3 The dichotomy for polynomial computations

Proof of Theorem 2. A fixed finite sequence of additions and multiplications makes each coordinate of the computed coefficient a polynomial in the entries of X and y. Denote this vector-valued polynomial by $P(X, y)$ and write $\tau = \sqrt{t}$. Substitution into this fixed polynomial gives

$$P(\tau X_0, \tau X_0\beta + \varepsilon) = \sum_{j=0}^{J} \tau^j c_j(\beta, \varepsilon),$$

where each c_j is a vector-valued polynomial with finite second moment under the Gaussian law of (β, ε).

Set $G = X_0^\top X_0 \succ 0$. The ridge coefficient is

$$w_R(t) = (I_d + tG)^{-1}\big(tG\beta + \sqrt{t}X_0^\top \varepsilon\big).$$

The Gaussian regression model gives

$$\mathbb{E}\|w_R(t) - \beta\|_2^2 = \operatorname{tr}(I_d + tG)^{-1} \longrightarrow 0, \qquad \mathbb{E}\|w_R(t)\|_2^2 = \operatorname{tr}\big(tG(I_d + tG)^{-1}\big) \longrightarrow d.$$

Thus $w_R(t) \to \beta$ in mean square and its second moment remains bounded. Independence and isotropy of the query reduce relative prediction risk to

$$R_{\mathcal{A}}(e_t) = \frac{\mathbb{E}\|P(\sqrt{t}X_0, \sqrt{t}X_0\beta + \varepsilon) - w_R(t)\|_2^2}{\mathbb{E}\|w_R(t)\|_2^2}. \tag{20}$$

If some positive-degree coefficient has positive second moment, let $D \geq 1$ be the largest such degree. Then

$$\tau^{-D}\big(P(\tau X_0, \tau X_0\beta + \varepsilon) - w_R(\tau^2)\big) \longrightarrow c_D(\beta, \varepsilon) \quad \text{in mean square.}$$

Indeed, all lower-degree terms vanish after division by τ^D, all higher-degree terms are zero almost surely, and the bounded-second-moment ridge term vanishes. Using Eq. (20),

$$R_{\mathcal{A}}(e_t) \sim \frac{\mathbb{E}\|c_D(\beta, \varepsilon)\|_2^2}{d}\, t^D.$$

The coefficient is strictly positive, which proves the first alternative.

Otherwise, every c_j with $j \geq 1$ vanishes almost surely. Since these are continuous polynomials and the joint Gaussian law has full support, they vanish identically. The constant coefficient is $c_0(\beta, \varepsilon) = P(0, \varepsilon)$, so, writing $g(y) = P(0, y)$, we have the identity

$$P(\tau X_0, \tau X_0\beta + \varepsilon) = g(\varepsilon) \qquad \text{for all } \tau, \beta, \varepsilon.$$

Taking $\tau = 1$ and $\beta = 0$ first shows that $P(X_0, y) = g(y)$ for every y. Taking $\tau = 1$ with arbitrary β, ε then gives

$$g(\varepsilon + X_0\beta) = g(\varepsilon).$$

Let $\Pi_\perp = I_n - X_0 G^{-1} X_0^\top$ be the orthogonal projection onto the complement of the column space of X_0. The preceding identity implies $g(\varepsilon) = g(\Pi_\perp \varepsilon)$. Under the Gaussian noise law, $\Pi_\perp \varepsilon$ is independent of $X_0^\top \varepsilon$ and of β. Hence $g(\varepsilon)$ is independent of the centred vector $w_R(t)$. Equation (20) now yields

$$R_{\mathcal{A}}(e_t) = 1 + \frac{\mathbb{E}\|g(\Pi_\perp \varepsilon)\|_2^2}{\mathbb{E}\|w_R(t)\|_2^2} \geq 1 \qquad \text{for every } t \geq 1,$$

which proves the second alternative. ∎