Transformers for In-Context PDE Solving and Inverse Problems

A self-contained finite-dimensional theory of encoder recovery, decoder preconditioning, generalization certificates, and replica order parameters

MMNN transformer/PDE notes

July 6, 2026

Abstract

We rewrite the PDE and inverse-problem in-context-learning task as finite least squares. One task is one latent vector z; one prompt is many pairs (u_i, f_i) produced by that same z. The encoder estimates z from weak equations $G_m z \simeq b_m$. The decoder solves the query system by a learned preconditioned Richardson iteration. This gives exact optimization identities for encoder-only, decoder-only, and stop-gradient encoder–decoder training. Generalization is certified either by exact Gaussian/Wishart risk formulas in the low-rank model or by clipped empirical Bernstein bounds on held-out tasks. The replica framework of low-rank linear-attention ICL applies exactly to the linear-attention finite least-squares reduction; for nonlinear or learnable-kernel attention it gives the correct order-parameter target, but the global scalar closure is not yet proved. The required order parameters are not only spectra: they include spectral overlaps between the learned preconditioner, the task Hessian, and the source covariance.

Contents

1 One paragraph reading guide **1**

2 From a PDE prompt to finite least squares **1**
2.1 Task, prompt, and query . . . 1
2.2 Weak equations for the encoder . . . 2
2.3 Finite-rank GP/KRR recast . . . 2

3 Task diversity and task difficulty **3**

4 Encoder-only theory **4**
4.1 Population model . . . 4
4.2 Encoder generalization . . . 4

5 Decoder-only theory **5**
5.1 Frozen decoder is Richardson . . . 5
5.2 Linear-attention pretraining as preconditioner learning . . . 5
5.3 Decoder verification . . . 6

6 Encoder–decoder composition **6**
6.1 Error decomposition . . . 6
6.2 Perturbation from latent error . . . 6
6.3 Stop-gradient triangular dynamics . . . 7

7 Exact correspondence with low-rank linear-attention ICL **8**
7.1 Paper model . . . 8
7.2 Our exact mapping . . . 8
7.3 Decoder as the algorithmic component . . . 8
7.4 Flexformer and learnable attention kernels . . . 9

8 Replica order parameters **9**
8.1 Quenched free energy . . . 9
8.2 Which overlaps are needed . . . 10
8.3 Replica diagnostic plots . . . 11

9 Scaling laws **11**
9.1 Prompt size . . . 11
9.2 Training task count . . . 11
9.3 Task rank and diversity . . . 12

10 Optimization and generalization plots **13**

11 What is proved and what is not **14**

12 Conclusion **14**

A Figure index **14**

B References **15**

1 One paragraph reading guide

The problem has two clocks. Outer time s is pretraining: the attention matrices, encoder dictionary, or preconditioner change. Inner time ℓ is in-context solving at frozen parameters: one decoder layer is one solver step. The exact identities are:

$$G_s - G_\star = (G_0 - G_\star)(I - \eta\Sigma_z)^s, \qquad e_{\ell+1} = (I - P_s H)e_\ell, \qquad e_L = (I - P_s H)^L e_0.$$

The statistical question is how many prompt equations, training tasks, and latent task directions are needed for these identities to be useful on new tasks. The replica question is which macroscopic overlaps close the typical test risk in the high-dimensional limit.

2 From a PDE prompt to finite least squares

2.1 Task, prompt, and query

Let $z \in \mathbb{R}^r$ be the hidden task coordinate,

$$z \sim \mathcal{N}(0, \Sigma_z).$$

Let f be a Gaussian process in a Hilbert space $\mathcal{H}$,

$$f \sim \mathrm{GP}(0, k_f),$$

and let the elliptic operator be affine in z:

$$A(z) = A_0 + \sum_{k=1}^{r} z_k A_k.$$

For one task z, the prompt contains m source-solution pairs

$$\Pi_z = \{(u_i, f_i)\}_{i=1}^m, \qquad A(z)u_i = f_i, \qquad f_i \overset{\text{iid}}{\sim} \mathrm{GP}(0, k_f).$$

The query is a fresh $f_\star$ from the same source law, and the desired answer is

$$u_\star = A(z)^{-1} f_\star.$$

Thus:

$$\boxed{\text{one task} = \text{one latent } z = \text{one prompt with many pairs } (u_i, f_i).}$$

2.2 Weak equations for the encoder

Choose test functions $v_1, \dots, v_q$. Since

$$f_i - A_0 u_i = \sum_{k=1}^{r} z_k A_k u_i,$$

testing against v_a gives

$$\langle v_a, f_i - A_0 u_i \rangle = \sum_{k=1}^{r} z_k \langle v_a, A_k u_i \rangle.$$

Stack rows indexed by (i, a):

$$G_{(i,a),k} = \langle v_a, A_k u_i \rangle, \qquad b_{(i,a)} = \langle v_a, f_i - A_0 u_i \rangle.$$

Then the prompt is the finite inverse problem

$$G_m z = b_m, \qquad G_m \in \mathbb{R}^{mq \times r}.$$

With noise or model error,

$$b_m = G_m z + \xi, \qquad \mathbb{E}[\xi \xi^\top] = \sigma^2 I.$$

The ridge encoder is

$$\widehat{z}_\lambda = (G_m^\top G_m + \lambda I)^{-1} G_m^\top b_m.$$

2.3 Finite-rank GP/KRR recast

If the GP has a finite feature representation

$$f(x) = \phi(x)^\top z, \qquad z \sim \mathcal{N}(0, \Lambda),$$

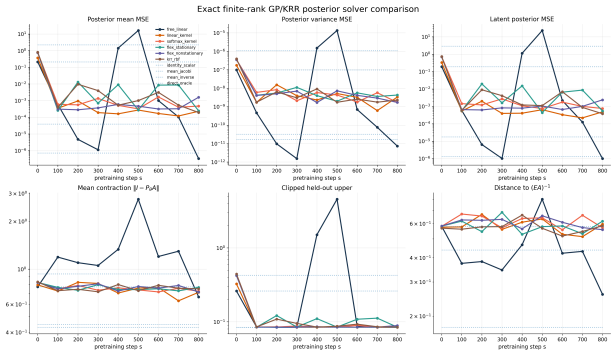
then observations $y = \Phi z + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, give

$$A_{\text{post}} = \Lambda^{-1} + \sigma^{-2} \Phi^\top \Phi, \qquad d_{\text{post}} = \sigma^{-2} \Phi^\top y, \qquad z_{\text{post}} = A_{\text{post}}^{-1} d_{\text{post}}.$$

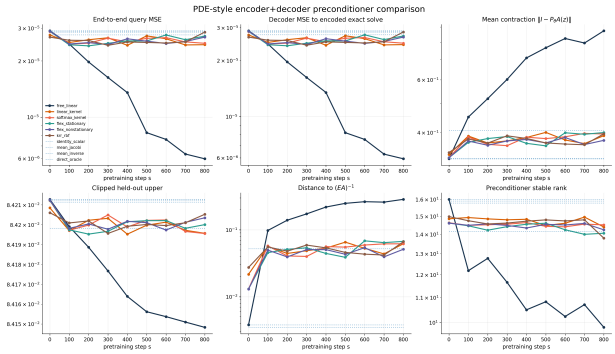
The dual KRR predictor is exactly the same:

$$\phi(x)^\top z_{\text{post}} = k_x^\top (K + \sigma^2 I)^{-1} y, \qquad k(x, x') = \phi(x)^\top \Lambda \phi(x').$$

This is the finite-dimensional bridge used by the GP/KRR experiments.



(a) Finite-rank GP/KRR posterior solvers.



(b) PDE encoder–decoder recast.

Figure 1: The GP and PDE tasks reduce to latent least-squares systems. The reported dual–primal GP gap is 1.443×10^{-14}; in the PDE K=4 recast the encoder has z-MSE 6.136×10^{-8}.

3 Task diversity and task difficulty

The task diversity is the effective dimension of Σ_z:

$$d_{\text{task}} = \frac{\operatorname{Tr}(\Sigma_z)^2}{\operatorname{Tr}(\Sigma_z^2)}.$$

If Σ_z is rank r isotropic, then $d_{\text{task}} = r$. If only a few latent directions have large variance, then pretraining sees a smaller task family even if the ambient parameter vector is long.

The prompt visibility matrix is

$$F_m = \frac{1}{m} G_m^\top G_m.$$

The easy regime is

$$\lambda_{\min}(F_m) \gg 0, \qquad \kappa(F_m + \lambda I) \text{ moderate.}$$

The hard regime is when some latent directions are invisible in the prompt:

$$\lambda_{\min}(F_m) \simeq 0.$$

The decoder difficulty for a known task is

$$\kappa(H(z)) = \frac{\lambda_{\max}(H(z))}{\lambda_{\min}(H(z))},$$

where $H(z)$ is the query Galerkin or latent least-squares Hessian.

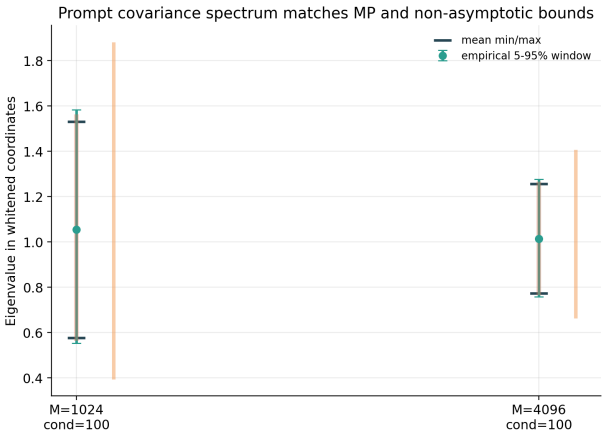
If the whitened prompt rows are iid Gaussian and $n = mq$, then

$$F_m \sim \frac{1}{n} X^\top X, \qquad X_{ij} \sim \mathcal{N}(0,1).$$

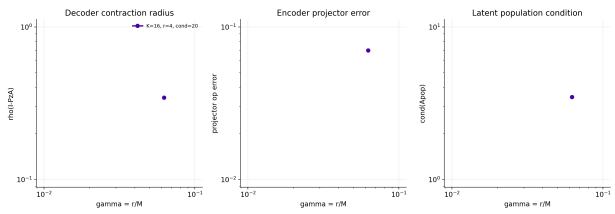
For $r, n \to \infty$ with $\gamma = r/n$, the empirical spectrum converges to Marchenko–Pastur:

$$d\mu_{\text{MP},\gamma}(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma\lambda} \mathbf{1}_{\lambda \in [\lambda_-, \lambda_+]}\, d\lambda, \qquad \lambda_\pm = (1 \pm \sqrt{\gamma})^2.$$

Thus the prompt-size scaling is controlled by $\gamma = r/(mq)$.



(a) Prompt spectrum and MP window.



(b) Prompt aspect-ratio scaling.

Figure 2: Prompt-size laws. Increasing the number of prompt pairs m or weak tests q reduces $\gamma = r/(mq)$, opens the smallest eigenvalue, and improves encoder recovery.

4 Encoder-only theory

4.1 Population model

The encoder-only low-rank recovery model is

$$b = G_\star z + \xi, \qquad z \sim \mathcal{N}(0, \Sigma_z), \qquad \xi \sim \mathcal{N}(0, \sigma^2 I).$$

The learned matrix G minimizes

$$\mathcal{L}_{\mathrm{enc}}(G) = \frac{1}{2}\mathbb{E}\|Gz - b\|_2^2.$$

Since

$$\nabla \mathcal{L}_{\mathrm{enc}}(G) = (G - G_\star)\Sigma_z,$$

population gradient descent with step size η satisfies

$$\boxed{G_s - G_\star = (G_0 - G_\star)(I - \eta\Sigma_z)^s.}$$

Theorem 1 (Encoder optimization curve)**.** *If* $0 < \eta < 2/\lambda_{\max}(\Sigma_z)$*, then*

$$R_{\mathrm{enc}}(s) = \frac{1}{K}\mathbb{E}\|G_s z - b\|_2^2 = \sigma^2 + \frac{1}{K}\operatorname{Tr}\left[(G_s - G_\star)\Sigma_z(G_s - G_\star)^\top\right],$$

and

$$R_{\mathrm{enc}}(s) - \sigma^2 \le \|I - \eta\Sigma_z\|_{\mathrm{op}}^{2s}\frac{1}{K}\operatorname{Tr}\left[(G_0 - G_\star)\Sigma_z(G_0 - G_\star)^\top\right].$$

Proof. The gradient formula gives the linear recursion above. Insert it into the quadratic population risk and diagonalize Σ_z. □

4.2 Encoder generalization

For a held-out sample of n tasks and clipped loss

$$\ell_c = \min(\ell, c), \qquad 0 \le \ell_c \le c,$$

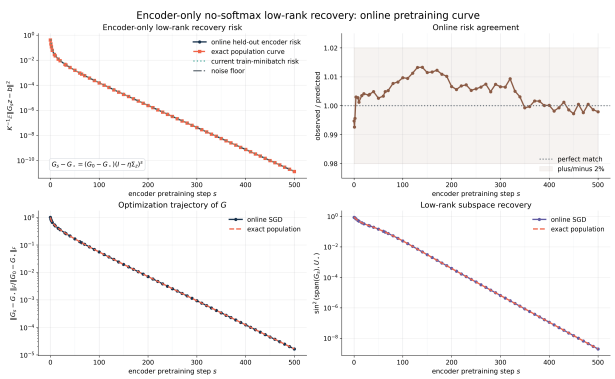
the empirical Bernstein certificate used throughout is

$$\boxed{R_c \le \widehat{R}_c + \sqrt{\frac{2\widehat{V}_c \log(2/\delta)}{n}} + \frac{7c\log(2/\delta)}{3(n-1)}.}$$

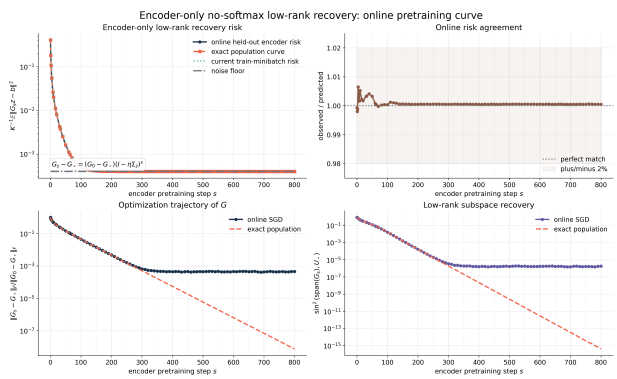
In the Gaussian encoder model the exact population curve above is stronger than a bound. The experiment verifies

$$\widehat{R}_{\mathrm{enc}}(800) = 4.0020466789 \times 10^{-4}, \qquad R_{\mathrm{enc}}(800) = 4.0000000000 \times 10^{-4},$$

$$\widehat{R}_{\mathrm{enc}}(800)/R_{\mathrm{enc}}(800) = 1.0005116697.$$



(a) Encoder online curve.



(b) Final encoder checkpoint.

Figure 3: Encoder-only risk is predicted by the exact population trajectory.

5 Decoder-only theory

5.1 Frozen decoder is Richardson

For one task, let the query system be

$$Hx_\star = d, \qquad H \succ 0.$$

At frozen pretrained parameters s, the decoder applies

$$x_{\ell+1} = x_\ell + P_s(d - Hx_\ell).$$

With $e_\ell = x_\ell - x_\star$,

$$\boxed{e_{\ell+1} = (I - P_sH)e_\ell, \qquad e_L = (I - P_sH)^L e_0.}$$

Consequently

$$\|e_L\|_2 \le \|I - P_sH\|_2^L \|e_0\|_2.$$

The exact depth risk for source covariance $S = \mathbb{E}[x_\star x_\star^\top]$ is

$$R_L(P_s) = \frac{1}{K} \operatorname{Tr}\left[(I - P_sH)^L S (I - P_sH)^{L\top}\right].$$

5.2 Linear-attention pretraining as preconditioner learning

The free linear-attention decoder model is

$$\widehat{x} = Pd, \qquad \mathcal{L}_{\mathrm{dec}}(P) = \frac{1}{2K}\mathbb{E}\|Pd - x_\star\|_2^2.$$

Let

$$C = \mathbb{E}[dd^\top], \qquad B = \mathbb{E}[x_\star d^\top].$$

Then

$$\nabla\mathcal{L}_{\mathrm{dec}}(P) = \frac{1}{K}(PC - B), \qquad P_\star = BC^\dagger.$$

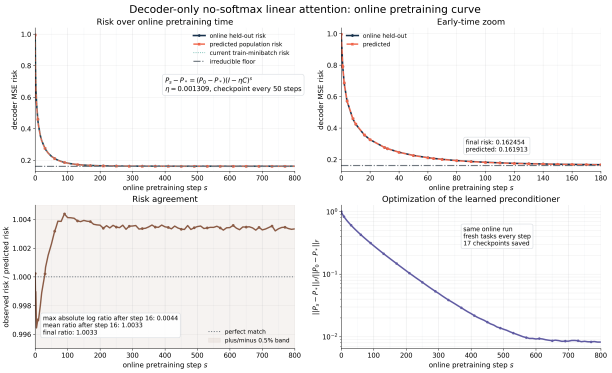
With the rescaled gradient step

$$P_{s+1} = P_s - \eta(P_sC - B),$$

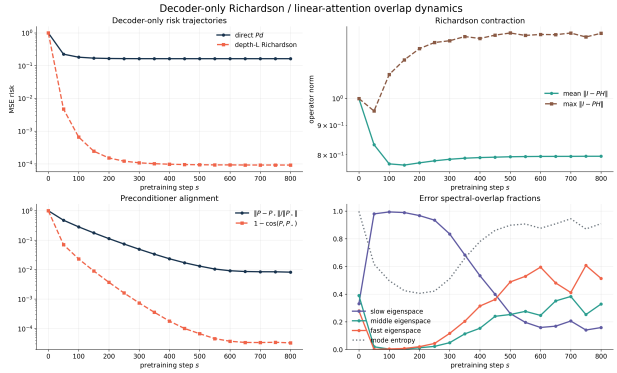
we get

$$\boxed{P_s - P_\star = (P_0 - P_\star)(I - \eta C)^s.}$$

Thus the outer training curve is again a closed covariance-filtered curve.



(a) Decoder online pretraining.



(b) Richardson overlap diagnostics.

Figure 4: Decoder-only training learns a preconditioner; evaluation at frozen P_s is exactly Richardson.

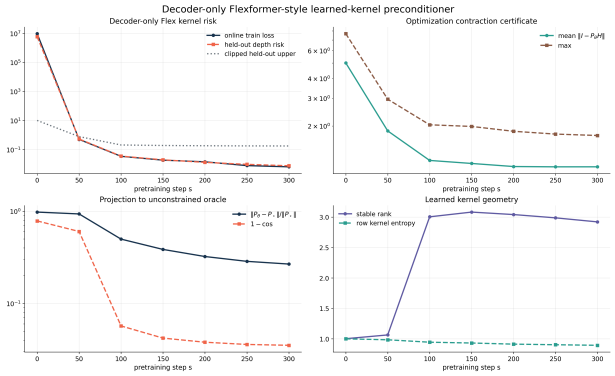
5.3 Decoder verification

At the final decoder-only Flexformer checkpoint:

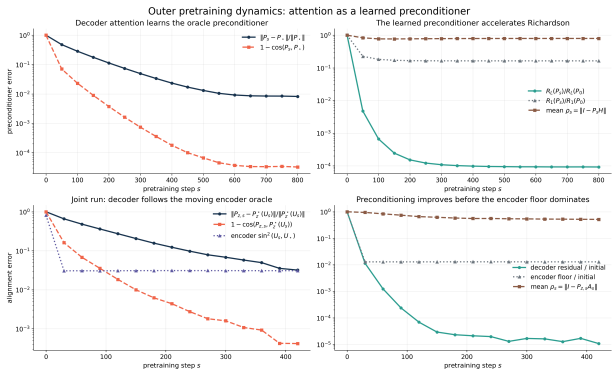
$$\widehat{R}_{\mathrm{dec}} = 7.674719 \times 10^{-3}, \qquad U_{\mathrm{EB}} = 1.771005 \times 10^{-1},$$

$$\Pr\{\|e_L\|_2 > \|I - P_\theta H\|_2^L \|e_0\|_2\} = 0.$$

The error/bound ratio is 2.2833×10^{-2}, the relative distance to the oracle $P_\star$ is 2.6788×10^{-1}, and the stable rank of P_θ is 2.9198.



(a) Decoder generalization and bound.



(b) Preconditioner strategy.

Figure 5: The decoder value is not that it beats an exact direct solver. Its value is that pretraining learns a reusable preconditioner that lowers the number of Richardson-like layers/iterations needed on new prompts.

6 Encoder–decoder composition

6.1 Error decomposition

Let $\widehat{z}$ be the encoder output, and let

$$x^\dagger(\widehat{z}) = H(\widehat{z})^{-1} d(\widehat{z})$$

be the exact solution of the encoded query system. The model output after L decoder layers is $\widehat{x}_L$. Then

$$\widehat{x}_L - x_\star = \underbrace{\widehat{x}_L - x^\dagger(\widehat{z})}_{\text{decoder}} + \underbrace{x^\dagger(\widehat{z}) - x_\star}_{\text{encoder}}.$$

Hence

$$\|\widehat{x}_L - x_\star\|_2^2 \le 2\|\widehat{x}_L - x^\dagger(\widehat{z})\|_2^2 + 2\|x^\dagger(\widehat{z}) - x_\star\|_2^2.$$

The first term is exact Richardson:

$$\widehat{x}_L - x^\dagger(\widehat{z}) = (I - P_\theta H(\widehat{z}))^L (\widehat{x}_0 - x^\dagger(\widehat{z})).$$

6.2 Perturbation from latent error

Assume

$$H(z) = H_0 + \sum_{k=1}^{r} z_k H_k, \qquad d(z) = d_0 + \sum_{k=1}^{r} z_k d_k,$$

and $H(z) \succeq \alpha I$ in a neighborhood of z. Then

$$x^\dagger(z) = H(z)^{-1} d(z).$$

For $\Delta z = \widehat{z} - z$,

$$x^\dagger(\widehat{z}) - x^\dagger(z) = H(z)^{-1}\left[\sum_k \Delta z_k (d_k - H_k x^\dagger(z))\right] + O(\|\Delta z\|^2).$$

Thus, locally,

$$\|x^\dagger(\widehat{z}) - x_\star\|_2 \le \frac{1}{\alpha}\left\|\sum_k \Delta z_k (d_k - H_k x_\star)\right\|_2 + O(\|\Delta z\|^2).$$

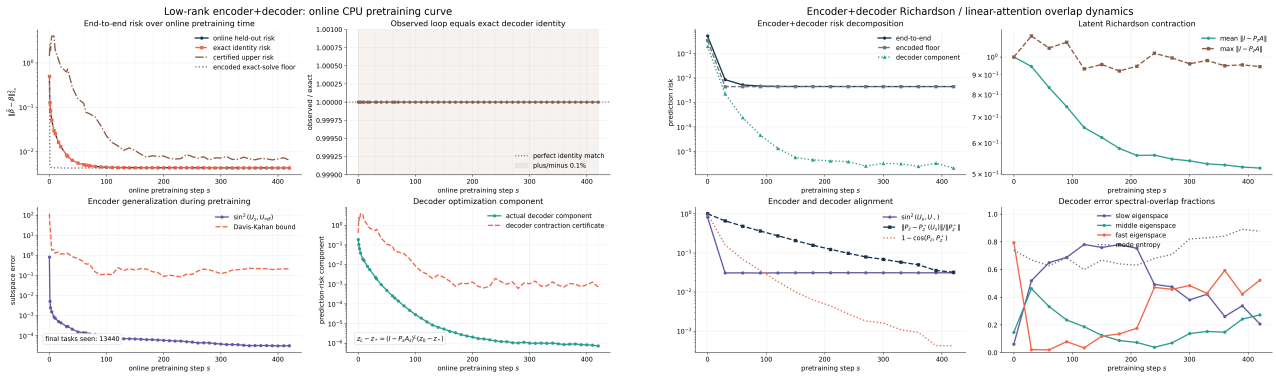
6.3 Stop-gradient triangular dynamics

If the decoder is stopped from backpropagating into the encoder and the encoder has its own loss, then the coupled training is triangular:

$$G_{s+1} - G_\star = (G_s - G_\star)(I - \eta_e \Sigma_z),$$

$$P_{s+1} - P_\star(G_s) = (P_s - P_\star(G_s))(I - \eta_d C(G_s)) + \big(P_\star(G_s) - P_\star(G_{s+1})\big).$$

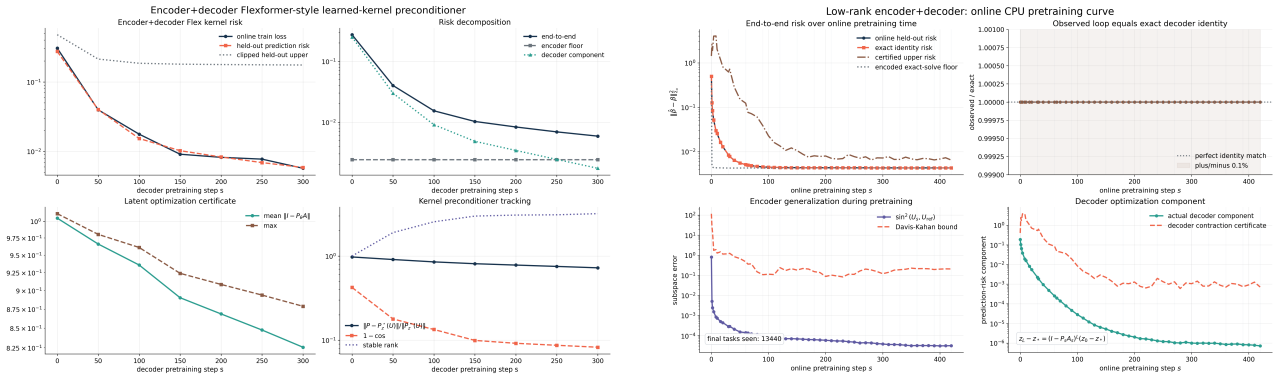
If the decoder is fast relative to the encoder, P_s tracks the slowly moving $P_\star(G_s)$. If the decoder starts before the encoder has signal, the first iterations train on noisy or poorly identified systems; this is the transient seen in joint runs.



(a) Online encoder–decoder.

(b) Composed overlap dynamics.

Figure 6: With stop-gradient and explicit encoder loss, the dynamics separate into encoder recovery plus decoder preconditioner adaptation.



(a) Flexformer encoder–decoder certificate.

(b) Low-rank encoder–decoder final run.

Figure 7: Encoder–decoder experiments verify the identity decomposition and zero pointwise Richardson-bound violations in the reported runs.

7 Exact correspondence with low-rank linear-attention ICL

7.1 Paper model

The low-rank ICL replica framework studies tasks

$$w^\mu = \sqrt{D/r}\, A v^\mu, \qquad v^\mu \sim \mathcal{N}(0, I_r),$$

and prompt pairs

$$y_i^\mu = \langle w^\mu, x_i^\mu \rangle + \varepsilon_i^\mu.$$

The one-layer linear-attention architecture has the form

$$\mathrm{Att}_\Theta(C) = C + \frac{1}{L} V C (KC)^\top (QC),$$

with a linear readout. The resulting query prediction can be written as a linear predictor on a prompt-derived feature matrix H^μ:

$$\widehat{y}^\mu = \mathrm{Tr}(W H^{\mu\top}).$$

7.2 Our exact mapping

Our low-rank linear regression task is the same model after renaming:

$$w^\mu \longleftrightarrow \beta(z^\mu) = U_\star z^\mu, \qquad A \longleftrightarrow U_\star, \qquad v^\mu \longleftrightarrow z^\mu.$$

For PDE inverse problems, the task vector is not directly a regression weight; it is the coefficient vector z in the weak system:

$$G(\Pi_z) z = b(\Pi_z).$$

After Galerkin testing, this is still a low-rank linear task. The context length in the ICL paper corresponds to the number of weak prompt equations:

$$L_{\text{paper}} \longleftrightarrow n = mq.$$

The task rank is

$$r_{\text{paper}} \longleftrightarrow \mathrm{rank}(\Sigma_z).$$

The pretraining task diversity M_0 in the paper corresponds to the number of distinct latent directions/tasks covered by pretraining, i.e. the empirical rank and conditioning of the sampled z-cloud.

7.3 Decoder as the algorithmic component

The paper decomposes prediction into an algorithmic component and a noise component. In our notation, the algorithmic component is the learned preconditioner/solver:

$$x_{\ell+1} = x_\ell + P_\theta (d - H x_\ell).$$

The noise component is any part of the learned map not aligned with the prompt least-squares geometry. In the exact linear-attention/free-P reduction, pretraining drives

$$P_s \to P_\star = B C^\dagger,$$

so the algorithmic component is precisely the covariance-optimal linear preconditioner.

7.4 Flexformer and learnable attention kernels

For kernelized linear attention,

$$A_\theta(i,j) = \frac{k_\omega(q_i, k_j)}{\sum_a k_\omega(q_i, k_a)}, \qquad P_\theta = D_a A_\theta D_v + D_s.$$

For fixed θ, all decoder certificates remain exact because they only use P_θ. For training θ, the finite-set recursion is exact:

$$z_{i,\ell+1} = z_{i,\ell} + P_\theta r_{i,\ell}, \qquad r_{i,\ell} = d_i - H_i z_{i,\ell},$$

$$\lambda_{i,L} = \frac{z_{i,L} - z_i^\star}{NK}, \qquad \lambda_{i,\ell} = (I - P_\theta H_i)^\top \lambda_{i,\ell+1},$$

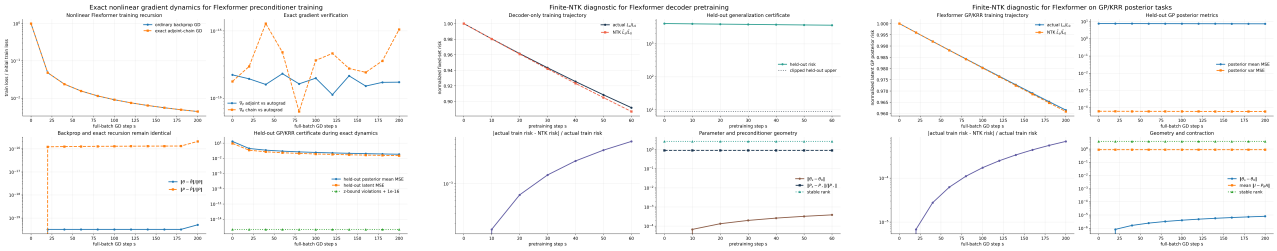
$$\nabla_P \mathcal{L}_N(P_\theta) = \sum_{i=1}^{N} \sum_{\ell=0}^{L-1} \lambda_{i,\ell+1} r_{i,\ell}^\top,$$

$$\theta_{s+1} = \theta_s - \eta\, D P_{\theta_s}^* \left[\nabla_P \mathcal{L}_N(P_{\theta_s})\right].$$

The checked relative errors were

$$\mathrm{relerr}(\nabla_P) = 1.727 \times 10^{-16}, \qquad \mathrm{relerr}(\nabla_\theta) = 1.048 \times 10^{-15}.$$

What is not proved is a global scalar replica closure for this nonlinear outer-training law.



(a) Exact finite-set gradients. (b) Decoder NTK local curve. (c) GP/KRR NTK local curve.

Figure 8: For nonlinear or learnable-kernel attention, the finite gradient law is exact and local NTK curves can be accurate. The global replica closure remains an order-parameter problem.

8 Replica order parameters

8.1 Quenched free energy

For a training set $\mathcal{D}_N$, define a Gibbs posterior over effective attention parameters P:

$$Z_N(\beta) = \int \exp\{-\beta N \widehat{\mathcal{R}}_N(P)\}\, d\pi(P).$$

The quenched free energy is

$$\mathcal{F}(\beta) = -\frac{1}{\beta N} \mathbb{E}_{\mathcal{D}_N} \log Z_N(\beta).$$

The replica identity is

$$\mathbb{E} \log Z_N = \lim_{n \to 0} \frac{\mathbb{E} Z_N^n - 1}{n}.$$

For n replicas $P^1, \ldots, P^n$, the calculation closes only after introducing overlaps.

8.2 Which overlaps are needed

Spectra alone are not enough. The decoder risk is

$$R_L(P) = \frac{1}{K} \operatorname{Tr}\left[(I - PH)^L S (I - PH)^{L\top}\right].$$

It depends on the relative eigenvectors of P, H, and S, not merely on eigenvalues of P. Required order parameters include

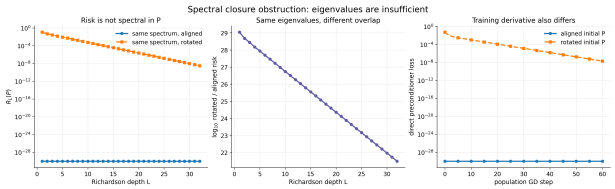
$$Q_{ab} = \frac{1}{K} \operatorname{Tr}(P^a P^{b\top}), \qquad M_a = \frac{1}{K} \operatorname{Tr}(P^a H),$$

$$T^{(1)}_{ab} = \frac{1}{K} \operatorname{Tr}(P^a H P^b H), \qquad T^{(S)}_{ab} = \frac{1}{K} \operatorname{Tr}(P^a H S H^\top P^{b\top}).$$

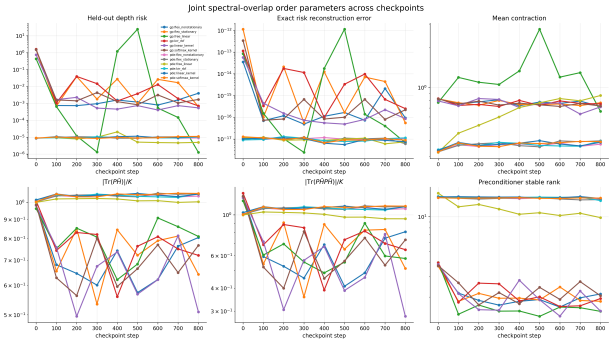
For encoder recovery, the corresponding overlaps are

$$Q^G_{ab} = \frac{1}{r} \operatorname{Tr}\left[(G^a - G_\star)\Sigma_z (G^b - G_\star)^\top\right].$$

For the prompt spectrum, the scalar MP law supplies the marginal spectral measure of $G_m^\top G_m$; the replica calculation supplies how the learned attention aligns to it.



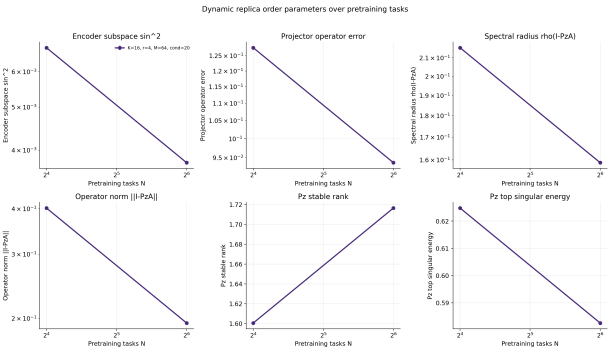
(a) Same spectra, different risk.



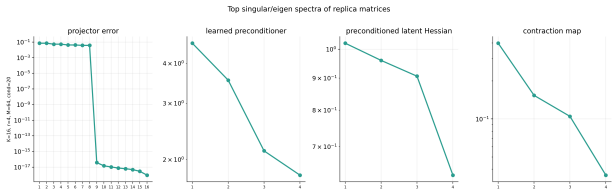
(b) Logged overlap state.

Figure 9: The obstruction experiment fixes eigenvalues and singular values but changes eigenbasis alignment. Depth-1 risks differ by 1.541×10^{-33} versus 1.090×10^{-1}. The logged joint order parameters reconstruct checkpoint risks with max absolute error 1.137×10^{-12}.

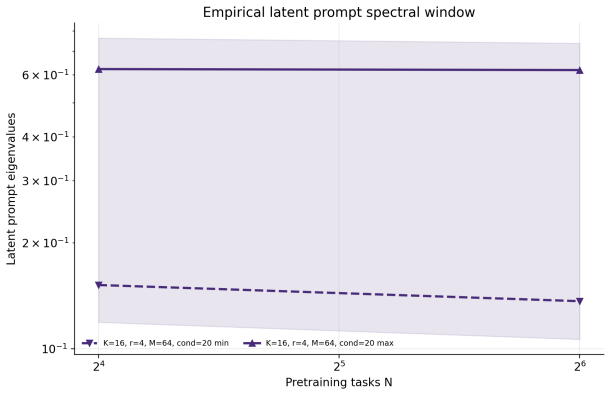
8.3 Replica diagnostic plots



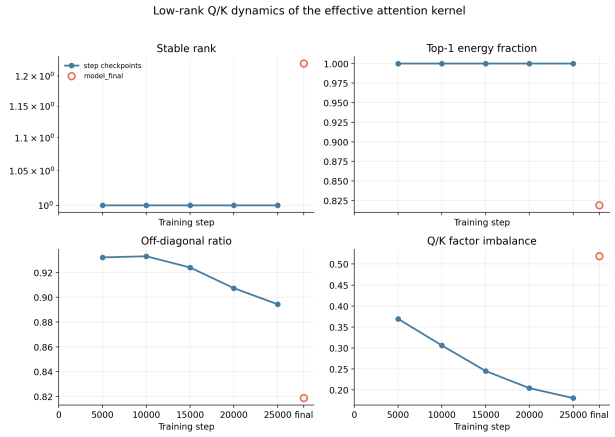
(a) Dynamic order parameters.



(b) Top spectra grid.



(c) Latent prompt spectral window.



(d) Q/K low-rank dynamics.

Figure 10: Replica-style diagnostics: prompt spectra, low-rank Q/K motion, and overlap variables. These are the quantities a closed dynamical replica theory must predict.

9 Scaling laws

9.1 Prompt size

Let $n = mq$ be the number of weak equations. In the whitened Gaussian prompt model with OLS and noise variance σ^2,

$$\mathbb{E}\|\widehat{z} - z\|_2^2 = \sigma^2 \, \mathbb{E} \operatorname{Tr}\left[(G_m^\top G_m)^{-1}\right] = \sigma^2 \frac{r}{n - r - 1}, \qquad n > r + 1.$$

With ridge λ,

$$R_z(\lambda) = \mathbb{E} \operatorname{Tr}\left[\lambda^2 (G_m^\top G_m + \lambda I)^{-2} \Sigma_z + \sigma^2 G_m^\top G_m (G_m^\top G_m + \lambda I)^{-2}\right].$$

In the proportional limit this becomes an MP integral.

9.2 Training task count

Let N be the number of pretraining tasks. For clipped loss, the empirical Bernstein deviation is

$$O\left(\sqrt{\frac{\widehat{V} \log(1/\delta)}{N}} + \frac{c \log(1/\delta)}{N}\right).$$

For linear free-P pretraining with effective covariance C, optimization error at outer time s is exactly

$$\mathcal{R}_{\mathrm{dec}}(P_s) - \mathcal{R}_{\mathrm{dec}}(P_\star) = \frac{1}{K}\,\mathrm{Tr}\{(P_0 - P_\star)(I - \eta C)^s C (I - \eta C)^s (P_0 - P_\star)^\top\}.$$

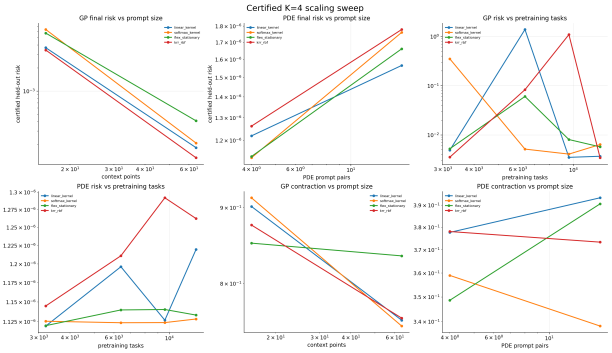
Finite N affects C, B through sample fluctuations; this is the finite pretraining-data regularization captured by the low-rank replica theory.

9.3 Task rank and diversity

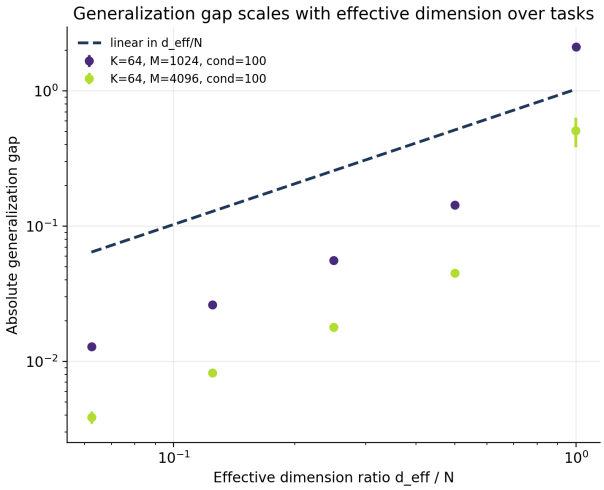
The task rank r appears in three places:

$$\gamma = \frac{r}{mq}, \qquad d_{\mathrm{task}} = \frac{\mathrm{Tr}(\Sigma_z)^2}{\mathrm{Tr}(\Sigma_z^2)}, \qquad \mathrm{rank}(U_\star) = r.$$

Generalization improves when $mq \gg r$, when N covers the effective latent support, and when the decoder preconditioner aligns with the stiff eigenvectors of the query Hessians.



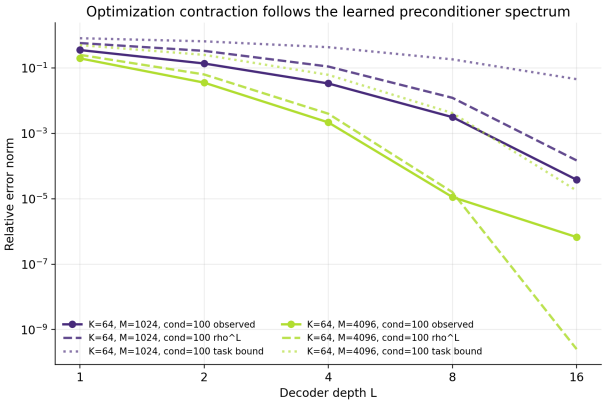
(a) Certified K=4 scaling sweep.



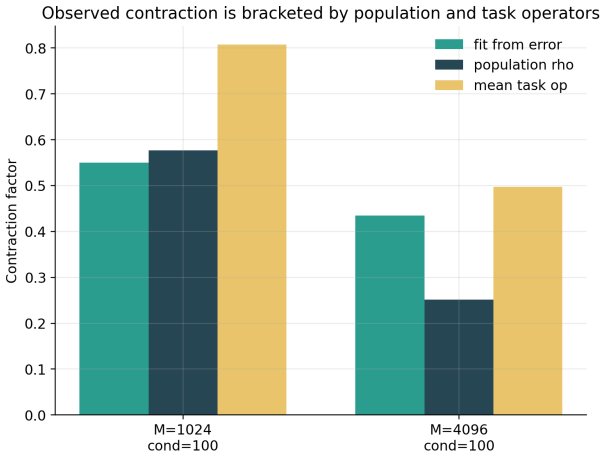
(b) Effective dimension over N.

Figure 11: Scaling evidence. The certified sweep has 96 rows and max reported violation 0.

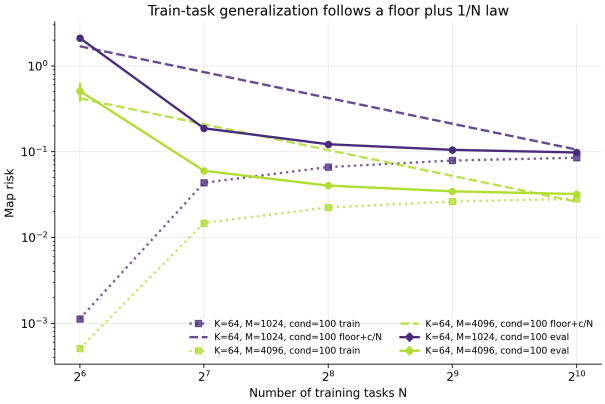
10 Optimization and generalization plots



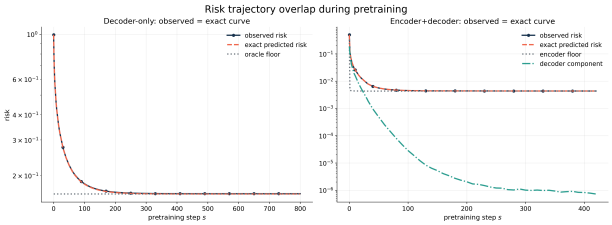
(a) Measured vs predicted contraction.



(b) Contraction factors.

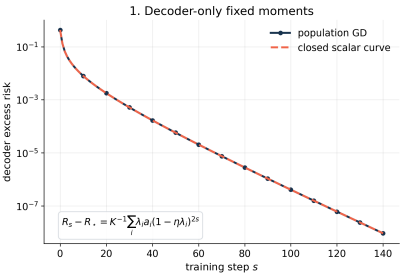


(c) Train/eval risk fit.

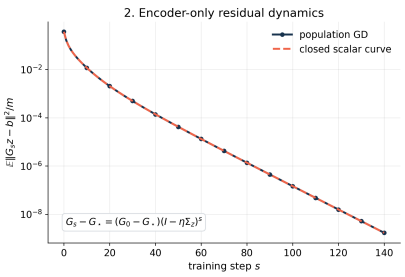


(d) Risk prediction overlap.

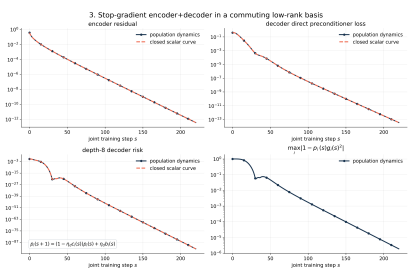
Figure 12: The deterministic optimization curves track contraction factors; the generalization plots compare train and held-out behavior.



(a) Decoder scalar law.



(b) Encoder scalar law.



(c) Coupled scalar law.

Figure 13: Scalar reductions are exact for the deliberately scalarized systems. They explain the shape of the curves but do not replace the overlap theory for full matrices.

11 What is proved and what is not

Claim	Status
PDE prompt $\to G_m z = b_m$	Exact after Galerkin/weak-form discretization.
Finite-rank GP $\leftrightarrow$ KRR	Exact primal–dual identity.
Encoder-only optimization	Exact population GD curve.
Decoder frozen optimization	Exact Richardson identity for every P_θ.
Encoder–decoder with stop-gradient	Exact triangular decomposition plus perturbation bound.
Held-out generalization	Certified by empirical Bernstein after clipping; exact Gaussian risk in encoder model.
Linear-attention low-rank ICL correspondence	Exact after finite least-squares recast and linear attention/readout reduction.
Softmax/Flexformer global training theory	Exact finite recursion and local NTK diagnostics; no proved global scalar replica closure yet.
Replica closure variables	Must include spectra and overlaps; spectra alone are false.

12 Conclusion

The rigorous core is finite-dimensional. The encoder solves a prompt inverse problem; the decoder learns and applies a preconditioner; the composition separates into identification error and solver error. Linear attention fits the low-rank ICL replica framework exactly once the prompt is written as a least-squares feature matrix. The difficult remaining theory is not the frozen optimization bound–that is exact–but the global outer-training law for nonlinear learnable attention kernels. The experiments show which macroscopic quantities that law must track: prompt MP spectra, Q/K low-rank motion, and overlaps among P, H, and S.

A Figure index

All figures below are included so the paper is self-contained as an audit artifact.

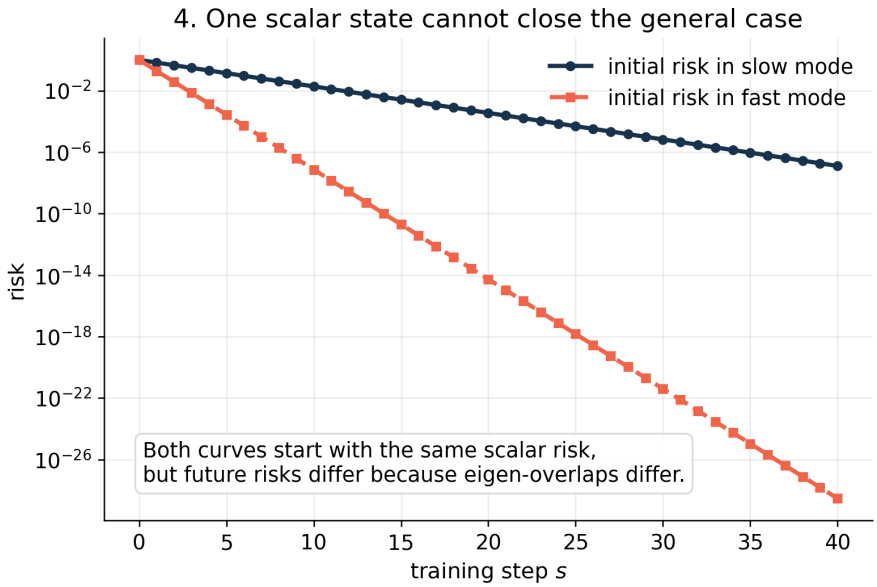
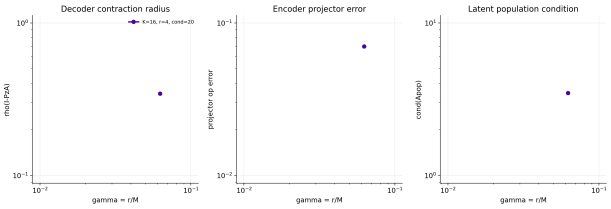
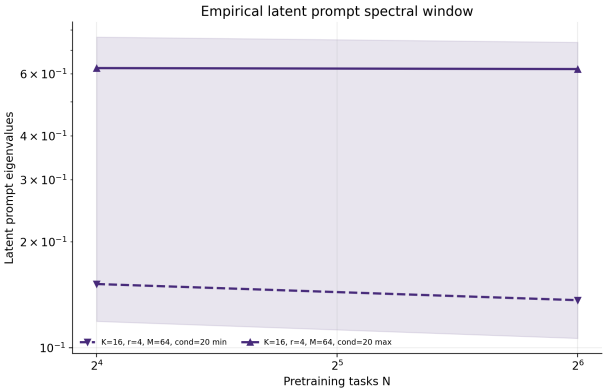


Figure 14: Scalar counterexample: scalar dynamics do not imply a full matrix closure.

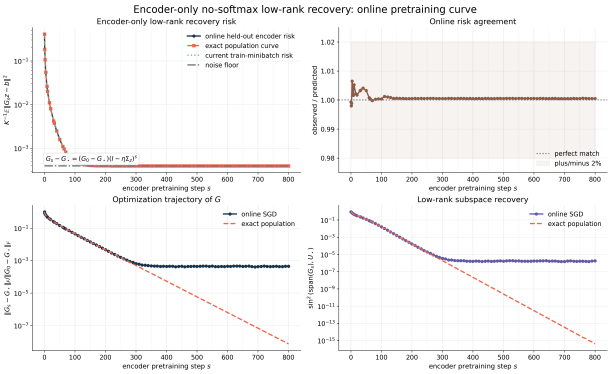


(a) Replica prompt aspect scaling.

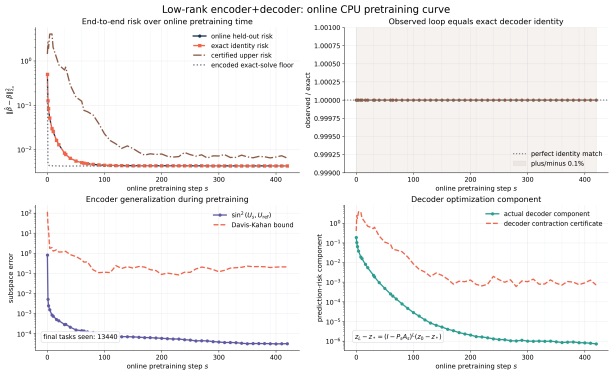


(b) Latent prompt spectral window.

Figure 15: Additional prompt spectral diagnostics.



(a) Encoder final plot.



(b) Encoder–decoder final plot.

Figure 16: Final online runs copied from the optimization/generalization report.

B References

References

[1] K. Takanami, T. Takahashi, and Y. Kabashima. Learning Linear Regression with Low-Rank Tasks In-Context. arXiv:2510.04548, 2026. https://arxiv.org/abs/2510.04548.

[2] Y. Zhang, A. K. Singh, P. E. Latham, and A. Saxe. Training Dynamics of In-Context Learning in Linear Attention. arXiv:2501.16265, 2025. https://arxiv.org/abs/2501.16265.

[3] H. Zhang and F. Zhou. Flexformer: Flexible Linear Transformer with Learnable Attention Kernel. arXiv:2606.27748, 2026. https://arxiv.org/abs/2606.27748.