What can linear attention learn from nonlinear teachers in-context?

Research group discussion | approximately 25 minutes

Paper by Mary Letey, Arman Rysmakhanov, Yue M. Lu,
Cengiz Pehlevan, and Jacob A. Zavatone-Veth

arXiv:2610.10761v1 | 29 October 2026

Accepted at NeurIPS 2026 PriGM **Workshop**, per author publication list.
Code forthcoming: author-linked repository contains only a README.
Status checked 2026-10-29; no individual proceedings or decision page verified.

Only the first Hermite mode becomes signal

Question

What can a reduced linear-attention predictor learn from $y = f(\boldsymbol{x}^\top \boldsymbol{w}) + \varepsilon$?

$$f(g) = \underbrace{a_1 g}_{\text{learnable projection}} + \underbrace{\sum_{m \neq 1} a_m h_m(g)}_{\text{residual modes}}, \qquad g \sim \mathcal{N}(0,1).$$

- In the paper's high-dimensional ridgeless limit, residual energy acts as additional noise.
- More data can improve estimation of $a_1 g$.
- The remaining error floor is fixed by the predictor's expressivity.

Paper, Sec. 3, Result 1, Eqs. (8)–(11), PDF pp. 3–4.

Each prompt hides one task direction

$$\underbrace{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)}_{\text{context}} \quad + \quad \underbrace{\boldsymbol{x}_{\text{test}}}_{\text{query}} \quad \longrightarrow \quad \widehat{y}_{\text{test}}$$

$$y_i = f(\boldsymbol{x}_i^\top \boldsymbol{w}) + \varepsilon_i, \qquad \boldsymbol{x}_i, \boldsymbol{x}_{\text{test}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d/d), \quad \varepsilon_i \sim \mathcal{N}(0, \rho).$$

- Shared $\boldsymbol{w}$ within one prompt; fixed link function f across tasks.
- Pretraining: k tasks from $\mathcal{N}(0, C_{\text{train}})$, spread uniformly over n prompts.
- Testing: a fresh task from $\mathcal{N}(0, C_{\text{test}})$; parameters stay fixed.

Paper, Sec. 2, Eq. (1), PDF p. 2. Independent Gaussian label noise.

Three ratios describe different data bottlenecks

Ratio	Resource	What it controls
$\alpha = \ell/d$	Context length	Per-task statistics
$\kappa = k/d$	Task diversity	Coverage of task directions
$\tau = n/d^2$	Pretraining prompts	Parameter estimation

Proportional limit

$d \to \infty$ with α, κ, τ of order one.
The reduced predictor has $d(d+1) \approx d^2$ parameters.

Repeating the same tasks increases τ without increasing κ.

Paper, Sec. 2, Eqs. (3)–(4), PDF pp. 2–3.

The query enters through a linear map

$$\boldsymbol{u} = \frac{d}{\ell}\sum_i y_i \boldsymbol{x}_i, \qquad s = \frac{1}{\ell}\sum_i y_i^2, \qquad H = \boldsymbol{x}_{\text{test}}[\boldsymbol{u}^\top, s].$$

$$\Gamma = [A, \boldsymbol{b}] \in \mathbb{R}^{d\times(d+1)} \quad \Longrightarrow \quad \boxed{\widehat{y}_{\text{test}} = \boldsymbol{x}_{\text{test}}^\top (A\boldsymbol{u} + s\boldsymbol{b})}.$$

- $\boldsymbol{u}$: empirical input–label correlation; s: label second moment.
- Context changes the prediction direction.
- Squared labels do not create a nonlinear query readout or an intercept.

Paper, Eqs. (4)–(5), PDF p. 3. The $[A, \boldsymbol{b}]$ expansion is a teaching derivation.

Pretraining is ridge regression across prompts

$$\gamma = \mathrm{vec}(\Gamma), \qquad \boldsymbol{h}_\mu = \mathrm{vec}(H^\mu),$$

$$\gamma^* = \left(\frac{n}{d}\lambda I + \sum_{\mu=1}^{n} \boldsymbol{h}_\mu \boldsymbol{h}_\mu^\top \right)^{-1} \sum_{\mu=1}^{n} y_{\mathrm{test}}^\mu \boldsymbol{h}_\mu.$$

Pretraining

Fit one shared parameter matrix Γ^* from labeled prompts.

In-context inference

Compute $\boldsymbol{u}, s, H$ on a new task; keep Γ^* fixed.

Metric: fresh-task mean squared error, including query-label noise.

Paper, Eqs. (6)–(7), PDF p. 3. Main result concerns $\lambda \to 0$.

Hermite projection is different from a Taylor term

$$f(g) = \sum_{m \geq 0} a_m h_m(g), \quad a_m = \mathbb{E}[f(g)h_m(g)], \quad \mathbb{E}[h_m h_j] = \delta_{mj}.$$

$$h_0 = 1, \quad h_1 = g, \quad h_2 = \frac{g^2 - 1}{\sqrt{2}}, \qquad a_1 = \mathbb{E}[gf(g)].$$

Link	First coefficient	Residual energy
$f(g) = g$	$a_1 = 1$	0
$f(g) = g^2 - 1$	$a_1 = 0$	2
$f(g) = g^3$	$a_1 = 3$	6

A cubic target contains the learnable projection $3g$.

Definition: paper, Eq. (2), PDF p. 2. Table: teaching examples derived from Gaussian moments.

Signal energy and residual energy play different roles

$$\underbrace{\rho_f}_{\text{total label second moment}} = \rho + \sum_{m \geq 0} a_m^2 = \underbrace{a_1^2}_{\text{linear signal}} + \underbrace{\rho_{\text{nonlinear}}}_{\text{effective noise}},$$

$$\boxed{\rho_{\text{nonlinear}} = \rho + \sum_{m \neq 1} a_m^2}.$$

- The residual includes a_0^2: this predictor has no query intercept.
- ρ_f is a second moment, not a centered variance when $a_0 \neq 0$.
- "Effective noise" describes the asymptotic risk contribution; the residual need not be independent of the input.

Paper, Eq. (8), PDF p. 3; predictor in Eq. (5).

Task-norm concentration makes Hermite analysis valid

$$\boldsymbol{x}^\top \boldsymbol{w} \mid \boldsymbol{w} \sim \mathcal{N}(0, \sigma_{\boldsymbol{w}}), \quad \sigma_{\boldsymbol{w}} = \|\boldsymbol{w}\|^2/d,$$

$$\mathbb{E}[\sigma_{\boldsymbol{w}}] = \frac{\operatorname{tr}(C)}{d}, \qquad \operatorname{Var}(\sigma_{\boldsymbol{w}}) = \frac{2\operatorname{tr}(C^2)}{d^2}.$$

Assumptions on both training and test tasks

$\operatorname{tr}(C) = d$ and $\operatorname{tr}(C^2) = o(d^2)$ imply $\sigma_{\boldsymbol{w}} \to 1$.
The effective rank $d^2/\operatorname{tr}(C^2)$ must diverge.

The link has at most polynomial growth. A fixed-rank covariance carrying all the signal is outside these assumptions.

Paper, Result 1, PDF p. 3; Appendix A, PDF p. 7.

The context correlation extracts exactly one mode

$$\mathbb{E}[y\boldsymbol{x} \mid \boldsymbol{w}] \approx \frac{\boldsymbol{w}}{d}\mathbb{E}[gf(g)] = \frac{a_1 \boldsymbol{w}}{d} \quad \Longrightarrow \quad \mathbb{E}[\boldsymbol{u} \mid \boldsymbol{w}] \approx a_1 \boldsymbol{w}.$$

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top \mid \boldsymbol{w}] \approx \underbrace{a_1^2 \boldsymbol{w}\boldsymbol{w}^\top}_{\text{task structure}} + \underbrace{\frac{\rho_f}{\alpha} I_d}_{\text{context fluctuations}} .$$

- Orthogonality removes higher Hermite modes from the mean correlation.
- All label energy still contributes to finite-context fluctuations.
- Matching the leading moments allows reuse of the linear risk theory.

Paper, Appendix B, Lemma 1, PDF pp. 8–9; displayed at $\|\boldsymbol{w}\|^2/d \to 1$.

The main result rescales signal and replaces noise

For $a_1 \neq 0$, in the high-dimensional ridgeless limit:

$$\boxed{e_{\mathrm{ICL}}^{\mathrm{nonlinear}}(C_{\mathrm{train}}, C_{\mathrm{test}}, \rho) = e_{\mathrm{ICL}}^{\mathrm{linear}}(a_1^2 C_{\mathrm{train}}, a_1^2 C_{\mathrm{test}}, \rho_{\mathrm{nonlinear}})}.$$

Equivalent linear teacher

$\widetilde{\boldsymbol{w}} = a_1 \boldsymbol{w}$

$\mathrm{Var}(\widetilde{\varepsilon}) = \rho_{\mathrm{nonlinear}}$

What is preserved

Training–test task geometry; context, diversity, and sample ratios.

Increasing only the noise, while leaving signal strength unchanged, misses part of the transformation.

Paper, Result 1, Eq. (9), PDF p. 3; Appendix C, PDF p. 10.

Without a linear mode, interpolation amplifies noise

If $a_1 = 0$:

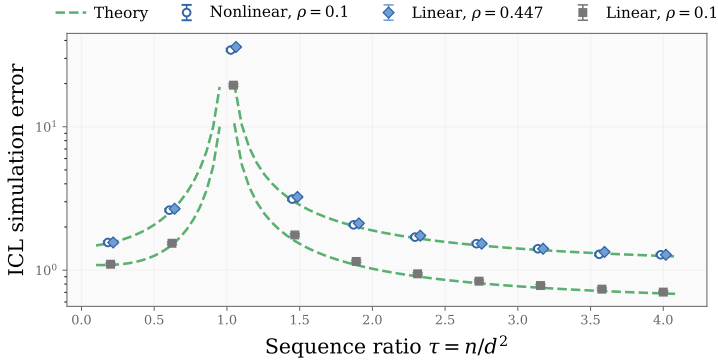
$$e_{\mathrm{ICL}}^{\mathrm{nonlinear}} = \rho_{\mathrm{nonlinear}} \begin{cases} (1-\tau)^{-1}, & \tau < 1, \\ \tau/(\tau-1), & \tau > 1. \end{cases}$$

- No dependence on α or κ in this limiting formula.
- Positive residual energy produces a singularity at $\tau = 1$.
- More pretraining eventually returns the error to $\rho_{\mathrm{nonlinear}}$.

Formula examples: $\tau = 0.5, 2 \Rightarrow 2\rho_{\mathrm{nonlinear}}$; $\tau = 0.9 \Rightarrow 10\rho_{\mathrm{nonlinear}}$.

Paper, Eq. (10), PDF p. 3; Appendix C, Eq. (15), PDF p. 10. Examples are derived, not simulation results.

Figure 1 displays the predicted noise equivalence



$d = 64$, $\alpha = \kappa = 2$, $C_{\text{train}} = C_{\text{test}} = I_d$; sine versus linear teachers.

Nonlinear and adjusted-noise markers nearly coincide; both show an interpolation peak. Their scaling needs care (next slide).

Paper, Fig. 1, PDF p. 4. Original figure. Vertical axis is logarithmic; lower error is better.

Figure 1 leaves an error-normalization ambiguity

For $f(g) = \sin g$ and $\rho = 0.1$:

$$a_1^2 = e^{-1} \approx 0.367879, \qquad \rho_{\mathrm{nonlinear}} = 0.1 + \frac{1 - e^{-2}}{2} - e^{-1} \approx 0.164453.$$

Unit-signal noise variance

$\rho_{\mathrm{nonlinear}}/a_1^2 \approx 0.447029$ matches the linear-noise legend.

- The prose instead gives $(a_1^2 + \rho_{\mathrm{nonlinear}})/a_1^2 \approx 1.447029$: the normalized *total* second moment.
- Rescaling labels by $1/a_1$ also rescales MSE by $1/a_1^2$; the axis/caption does not specify this.

Reading analysis of paper, Fig. 1 discussion, PDF p. 3, and legend, p. 4. Do not read overlap as equality of unscaled MSE.

Task diversity removes coverage error, not residual energy

With $\alpha, \tau \to \infty$ and $\tau \propto \alpha$, the inherited task-diversity threshold is

$$\kappa_c = \frac{\mathrm{rank}(C_{\mathrm{train}})}{d}.$$

Isotropic case: $C_{\mathrm{train}} = C_{\mathrm{test}} = I_d$

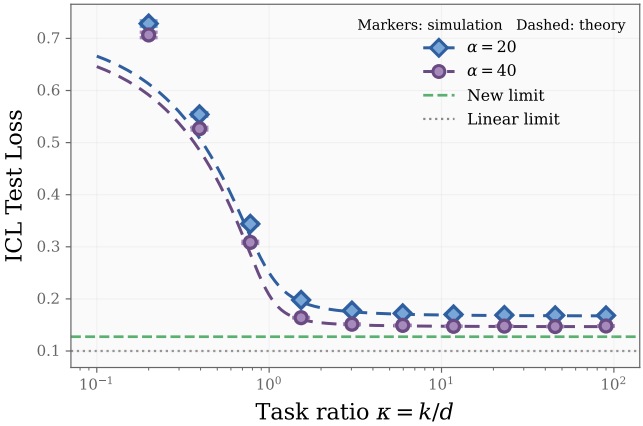
If $a_1 \neq 0$ and $\kappa > 1$, then

$$e_{\mathrm{ICL}} \longrightarrow \rho_{\mathrm{nonlinear}} = \rho + \sum_{m \neq 1} a_m^2.$$

Few tasks leave directions poorly covered. Sufficient diversity supports transfer of the linear rule to fresh tasks; the nonlinear residual remains.

Paper, Sec. 3, task-diversity paragraph and Eq. (11), PDF pp. 3–4.

Figure 2 approaches a raised error floor



$f(g) = \tanh g$
$d = 32$, $\rho = 0.1$
$\tau = \alpha/2$

Error falls around $\kappa \approx 1$.

Finite-resource curves remain above the new limiting error.

Paper, Fig. 2 and Eq. (11), PDF p. 4. Original figure; task-ratio axis is logarithmic.

For tanh, the extra limiting MSE is about 0.0274

$$a_1 \approx 0.605706, \quad \mathbb{E}[\tanh^2 g] \approx 0.394294,$$

$$\boxed{\rho_{\text{nonlinear}} = 0.1 + 0.394294 - 0.605706^2 \approx 0.127415}.$$

α	Context ℓ	τ	Prompts n
20	640	10	10,240
40	1,280	20	20,480

The two curves increase context and pretraining size together. Their difference does not isolate a context-length effect.

Settings: paper, Fig. 2, PDF p. 4. Table derived from ratios; moments computed from Eq. (8), not reported experimental values.

The conclusion is specific to this predictor and limit

- **Architecture:** a reduced predictor with a linear query readout. Nonlinear readouts or multilayer nonlinear transformations change the function class.
- **Distribution:** independent Gaussian inputs and Gaussian tasks with concentrating norms.
- **Asymptotics:** proportional dimensions and ridgeless risk; the derivation keeps leading moments and uses prior resolvent theory.
- **Evidence:** sine at $d = 64$ and tanh at $d = 32$ illustrate the mechanism; Figure 1 has a normalization ambiguity.

Paper, Secs. 1–3 and Appendices A–C, PDF pp. 1–4, 7–10; normalization issue is reading analysis.

The useful distinction is estimation versus expressivity

1. **Compute the projection.** $a_1 = \mathbb{E}[gf(g)]$ identifies the accessible signal.
2. **Account for residual energy.** $\rho_{\mathrm{nonlinear}}$ predicts a floor that remains after sufficient task coverage.
3. **Match scales.** Signal covariance and error units matter when comparing teachers.

Discussion

Which additional Hermite modes would become learnable if we added a nonlinear query readout?

Paper, Result 1 and Conclusion, PDF pp. 3–4. Discussion question is an extension motivated by the analysis.

References and paper resources

- Letey et al. (2026). *What can linear attention learn from nonlinear teachers in-context?* arXiv:2610.10761v1.
- Letey et al. (2026). *Pretrain-test task alignment governs generalization in in-context learning*. ICLR paper cited by the derivation.
- Lu et al. (2025). *Asymptotic theory of in-context learning by linear attention.* PNAS 122(28): e2502599122.

Author publication list: PriGM Workshop at NeurIPS 2026.
Author-linked code repository: README only as of 2026-10-29.

Background references are those used by the paper, PDF pp. 5–6. Figures in this deck are the authors' original assets.

Backup: leading moments match a linear teacher

At $\|\boldsymbol{w}\|^2/d \to 1$, with the paper's vectorization convention,

$$\mathbb{E}[y^2 \mid \boldsymbol{w}] \approx \rho_f, \qquad \mathbb{E}[y_{\text{test}} H \mid \boldsymbol{w}] \approx \frac{1}{d}[a_1^2 \boldsymbol{w}\boldsymbol{w}^\top,\ a_1\rho_f \boldsymbol{w}],$$

$$\mathbb{E}[\boldsymbol{h}\boldsymbol{h}^\top \mid \boldsymbol{w}] \approx \frac{1}{d} I_d \otimes \begin{bmatrix} \frac{\rho_f}{\alpha} I_d + a_1^2 \boldsymbol{w}\boldsymbol{w}^\top & a_1\rho_f \boldsymbol{w} \\ a_1\rho_f \boldsymbol{w}^\top & \rho_f^2 \end{bmatrix}.$$

- Substitute $\widetilde{\boldsymbol{w}} = a_1 \boldsymbol{w}$ and noise $\rho_{\text{nonlinear}}$.
- Total second moment stays $a_1^2 + \rho_{\text{nonlinear}} = \rho_f$.
- Appendix B drops finite-ℓ corrections to these leading blocks.

Paper, Lemma 1, Eqs. (12)–(14), PDF pp. 8–9; Appendix C, p. 10.

Backup: the no-signal branch has a scalar equation

When $a_1 = 0$, task randomness disappears from the leading moments. With $q = \rho_{\text{nonlinear}}/\alpha$:

$$\chi_0 \approx \frac{q}{q\tau/(1+\chi_0) + \tau\lambda}.$$

$$\widetilde{\lambda} = \lambda(1+\chi_0) \longrightarrow \begin{cases} q(1-\tau)/\tau, & \tau < 1, \\ 0, & \tau > 1. \end{cases}$$

Substitution into the inherited linear risk expression yields

$$e_{\text{ICL}} = \rho_{\text{nonlinear}}/(1-\tau) \quad \text{or} \quad e_{\text{ICL}} = \rho_{\text{nonlinear}}\tau/(\tau-1).$$

Paper, Appendix C, Eq. (15), PDF p. 10.

Backup: a projection identity explains the error floor

Teaching derivation. Fix $\|\boldsymbol{w}\|^2 = d$ and a context-dependent vector v independent of the fresh query. Then $g = \boldsymbol{x}^\top \boldsymbol{w} \sim \mathcal{N}(0,1)$.

$$\mathbb{E}[(f(g) + \varepsilon - \boldsymbol{x}^\top v)^2] = \rho + \mathbb{E}[f(g)^2] - \frac{2a_1}{d}\boldsymbol{w}^\top v + \frac{\|v\|^2}{d}$$

$$= \boxed{\rho_{\mathrm{nonlinear}} + \frac{\|v - a_1\boldsymbol{w}\|^2}{d}}.$$

- The best linear query prediction uses $v = a_1\boldsymbol{w}$.
- Estimation error is the second term; representational residual is $\rho_{\mathrm{nonlinear}}$.
- The paper computes how pretraining and context control the estimation term asymptotically.

Derived explanation based on paper, Eqs. (2), (5), (8). Normalized fixed task used for this identity.