WHEN CAN BEHAVIORAL TESTS DISTINGUISH IN-CONTEXT COMPUTATIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning in transformers is often studied by comparing predictions with familiar algorithms. We ask when a benchmark can distinguish these computational explanations. For linear regression, our test compares a model with the best possible output of a specified family of gradient-style updates, allowing all scalar settings to be chosen separately for each prompt. Lower model error excludes the whole family, rather than one optimizer configuration. On a released 12-layer transformer and one independently trained replication, shorter contexts make predictions less accurate but provide stronger risk-based exclusions. With 22 examples in 20 dimensions, both models outperform the best 12-update predictors, including in an exploratory analysis removing the hardest tenth of prompts. With forty examples, the same families are already more accurate than the models. An exact Gaussian residual law explains the comparator's improvement: longer contexts lower its minimum achievable error. A separate symmetry analysis distinguishes useful departures from errors that depend on example mixing and feature coordinates; at forty examples, these dependencies account for 96–97% of error in the recovered regression coefficients. Prediction accuracy and the strength of behavioral evidence are different objectives: an easier prediction task can make a computational explanation harder to exclude.

1 INTRODUCTION

When a transformer trained on regression predicts a new output from examples in its prompt, what learning algorithm has it acquired? One possibility is that its layers run an optimizer, such as gradient descent or Newton's method. Controlled studies investigate this question by comparing predictions and internal representations with familiar algorithms (Garg et al., 2022a; Akyürek et al., 2023; von Oswald et al., 2023; Fu et al., 2024a). But resembling one algorithm more than another does not rule out alternatives, and failure of a particular update schedule need not mean its whole family fails.

We replace that comparison with a stronger one: *give an entire family the best possible output it can produce on each task*. For a fixed update budget, we optimize over every allowed combination of update directions, separately for each prompt and even using the correct regression function to choose it. If the model achieves lower prediction error, no member of that specified family can explain its performance. The measured gap also bounds how closely any member of the family can approximate the model's predictions, strengthening the evidence beyond failure of selected optimizer settings.

With forty examples, the competing family is already more accurate than the model when allowed one update per layer. With fewer examples, the model is less accurate but beats that family (Figure 1). Here, increasing context improves prediction while making the best-in-family risk comparison inconclusive. Departures from a family can also be errors: predictions vary when examples are mixed or feature coordinates rotated, although the regression answer stays unchanged. We separate this harmful sensitivity from useful information. The practical lesson is to establish what the chosen comparison can distinguish and whether the departure helps solve the task.

The relevant *Krylov space* contains the prediction vectors reachable by a fixed number of gradient-style updates from zero, whatever scalar step sizes and mixing weights are used. Our *oracle* chooses the best predictor in that space for each prompt; its error is a floor for the entire family. This covers zero-initialized gradient descent, conjugate gradient, and arbitrary prompt-dependent scalar update

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.

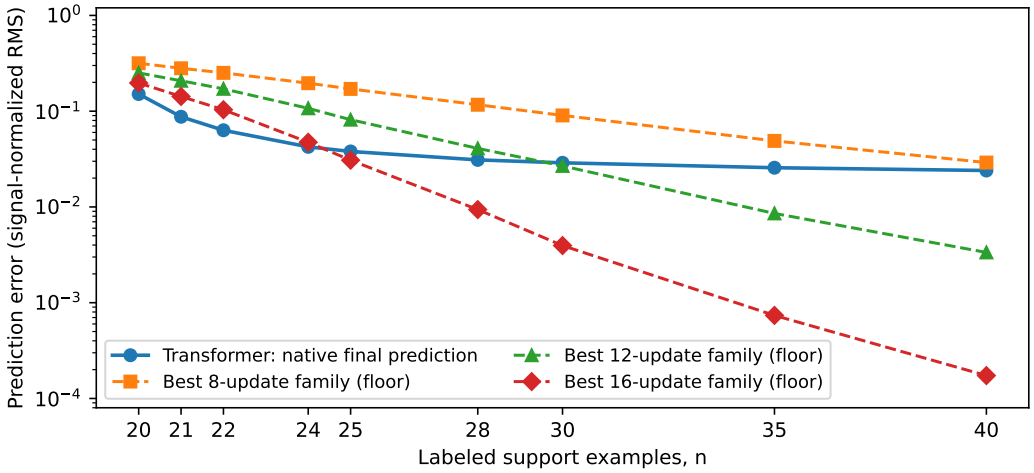

Figure 1: **Worse predictions can give stronger evidence about computation.** Each dashed curve is the lowest error in a family allowing 8, 12, or 16 updates, choosing the best combination of directions for each prompt. *Below a floor, the model outperforms the family; above it, this test is inconclusive.* Points summarize 1,024 paired episodes. The reported exclusions are established by simultaneous paired confidence bounds.

schedules. Testing a budget of one update per layer is a specific computational hypothesis: GD++ feature transformations and Newton-type matrix updates are richer alternatives (von Oswald et al., 2023; Vladymyrov et al., 2024; Fu et al., 2024a). Section 2 defines the class and its preconditioned extension precisely.

At each context length, the *exclusion boundary* is the largest tested update budget for which the mean squared-error difference (oracle minus model) has a positive simultaneous lower confidence bound. For the unpreconditioned family, this boundary coincides across all nine contexts in the two audited models. This observed agreement contrasts with evidence of seed-dependent adversarial responses in larger models (Anwar et al., 2025).

Our contributions are:

- **A test that rules out entire algorithm families, not single baselines.** Adaptive Krylov floors optimize every scalar coefficient, allow a specified preconditioner menu, and quantify separation from approximate family explanations.
- **Evidence that more accurate predictions can be less informative about computation.** At forty examples, the 12-update families outperform the models; shorter contexts reverse that comparison. An exact Gaussian law explains the falling family floor.
- **A way to separate useful departures from symmetry-dependent error.** Crossed row mixing and feature rotation distinguish errors removed by either average from useful information that survives both.

2 Class-relative behavioral separation

2.1 Task, risk, and a permissive rival

Let $D = (X, y)$, with $X \in \mathbb{R}^{n\times d}$ having iid $N(0,1)$ entries, $w \sim N(0, I_d)$ independent of X, and $y = Xw$. A query $q \sim N(0, I_d)$ is independent of the support. We study $d = 20$ and $n \geq d$, so X has full column rank and w is determined by D almost surely. Write

$$S = X^\top X/n, \qquad r = X^\top y/n = Sw, \qquad \mathcal{R}_n(f) = \frac{1}{d}\mathbb{E}_{D,q}\big(f(D,q) - q^\top w\big)^2. \tag{1}$$

Risks are normalized by the population signal energy $\mathbb{E}(q^\top w)^2 = d$; model inputs retain their original scale.

For $k \geq 1$, define

$$\mathcal{K}_k(S, r) = \operatorname{span}\{r, Sr, \ldots, S^{k-1}r\}, \qquad \mathcal{C}_k = \{(D, q) \mapsto q^\top z(D) : z(D) \in \mathcal{K}_k(S, r)\}. \tag{2}$$

The coefficients are arbitrary measurable functions of the support and its labels, fixed with respect to the query. They can adapt separately to each prompt and right-hand side. In exact arithmetic, k zero-initialized gradient updates with scalar linear-span operations lie in this space. Conjugate-gradient iterates are included, even though their coefficients depend on the labels. We count a first update from zero as $k = 1$, after forming r. The hypothesis $H_L : f \in \mathcal{C}_L$ therefore includes one scalar linear-span update per layer, with arbitrarily chosen support-dependent coefficients.

The test is deliberately simple: restricting the rival's output directions makes its risk floor exactly computable, even with arbitrary scalar adaptation. The resulting risk comparison applies to any transformer predictor.

Proposition 1 (Adaptive target oracle)**.** *Let $P_k(D)$ be the Euclidean projector onto $\mathcal{K}_k(S, r)$. Then*

$$F^2_{n,k} := \inf_{h \in \mathcal{C}_k} \mathcal{R}_n(h) = \frac{1}{d}\mathbb{E}_D\|(I - P_k)w\|^2. \tag{3}$$

Consequently, $\Delta_{n,k} := F^2_{n,k} - \mathcal{R}_n(f) > 0$ excludes f from $\mathcal{C}_k$ and shows that its risk is lower than that of every member of the class.

Proof. Conditioning on D, Gaussian query covariance gives $\mathbb{E}_q(q^\top(z - w))^2 = \|z - w\|^2$. Euclidean projection minimizes this quantity over the subspace. The projected target is a measurable function of D in the stated noiseless full-rank setting, so the pointwise minimum is attained within the unrestricted measurable output class. Averaging proves the result. □

The converse is false: failure to beat the floor does not establish membership. Following query-response implicit-weight recovery (Akyürek et al., 2023; Fu et al., 2024a), define its population version

$$v(D) = \mathbb{E}_q[qf(D, q) \mid D], \qquad h(D, q) = f(D, q) - q^\top v(D). \tag{4}$$

Then $\mathbb{E}_q[qh \mid D] = 0$ and

$$\operatorname{dist}(f, \mathcal{C}_k)^2 = \frac{1}{d}\mathbb{E}\|(I - P_k)v\|^2 + \frac{1}{d}\mathbb{E}h^2, \tag{5}$$

where distance uses the risk's normalized L^2 norm and h includes any query-independent offset.

Corollary 2 (Sharp risk-to-distance certificate)**.** *For squared loss with Bayes predictor b, let $r_f = \|f - b\|^2_{L^2}$ and $a_{\mathcal{C}} = \inf_{g \in \mathcal{C}} \|g - b\|^2_{L^2}$. Then*

$$\operatorname{dist}(f, \mathcal{C}) \geq \left(\sqrt{a_{\mathcal{C}}} - \sqrt{r_f}\right)_+.$$

This follows from the triangle inequality and is sharp given only these two risks. For a nonzero closed linear class, risk at or above its floor alone cannot certify nonmembership (Appendix B.9).

2.2 Specified preconditioning and an oracle menu

For a symmetric positive-definite inverse-preconditioner $C = C(D)$ fixed during the putative iteration, define

$$\mathcal{K}^C_k(D) = \operatorname{span}\{Cr, CSCr, \ldots, (CS)^{k-1}Cr\}. \tag{6}$$

Preconditioned conjugate-gradient solution iterates lie in this space. The oracle minimizes Euclidean prediction error after mapping back to solution coordinates.

Our declared menu $\mathcal{M}$ consists of identity, Jacobi, block-Jacobi with consecutive blocks of size two and four, and one global SPD map fitted on independent calibration prompts for each n. The menu oracle is

$$F^2_{n,k,\mathcal{M}} = \frac{1}{d}\mathbb{E}_D \min_{C \in \mathcal{M}} \|(I - P_{k,C})w\|^2. \tag{7}$$

The minimum is inside the expectation: the oracle selects one preconditioner on each prompt and optimizes coefficients within that member's space. Proposition 1 applies to this union. Aggregating predictions or directions across members can leave the union and is outside this class.

2.3 USEFUL DEPARTURES, RATHER THAN ARBITRARY DEPARTURES

Let $v_\perp = (I - P_k)v$ and $w_\perp = (I - P_k)w$. Define the usefulness of the outside linear component by

$$U_{n,k}(v) = \frac{1}{d}\mathbb{E}\left[2\langle w_\perp, v_\perp\rangle - \|v_\perp\|^2\right] = F_{n,k}^2 - \frac{1}{d}\mathbb{E}\|(I - P_k)(v - w)\|^2. \tag{8}$$

Removing only $q^\top v_\perp$ changes risk by $U_{n,k}$; positive values mean the component helps. Using the orthogonality in (4),

$$\Delta_{n,k} = U_{n,k}(v) - \frac{1}{d}\mathbb{E}\|P_k(v - w)\|^2 - \frac{1}{d}\mathbb{E}h^2. \tag{9}$$

A target-risk crossing therefore establishes useful query-linear information outside the class. Query nonlinearity contributes only additional error, making a crossing harder. Outside energy by itself can instead be harmful.

2.4 LONGER GAUSSIAN CONTEXTS SHRINK THE EXCLUSION WINDOW

Write $F_{n,d,k}$ to make dimension explicit. The following consequence of Gaussian MINRES theory explains the benchmark side of Figure 1.

Proposition 3 (Finite-dimensional Gaussian floor). *Fix $d \geq 2$ and $1 \leq k < d$. For $n \geq d$, the relative promptwise oracle error $Z_{n,d,k} = \|(I - P_k)w\|^2/\|w\|^2$ is stochastically decreasing in n, and $F_{n,d,k}^2 = \mathbb{E}Z_{n,d,k}$ is strictly decreasing. Moreover,*

$$F_{n,d,k}^2 \sim \frac{\prod_{i=1}^k (d-i)}{n^k}, \qquad F_{n,d,k}^2 \leq \min\left\{1, \prod_{i=1}^k \frac{d-i}{n-i-1}\right\} \quad (n > k+1). \tag{10}$$

The asymptotic holds for fixed d, k as $n \to \infty$. At $k = d$ the floor is zero.

Appendix G applies the exact Gaussian law of Paquette & Trogdon (2023, Theorem 1.2(b)) and the bidiagonal model of Dumitriu & Edelman (2002) to derive the chi-square representation, then proves the coupling and rate.

Corollary 4 (Accuracy versus exclusion). *If $\mathcal{R}_n(f_n) \sim bn^{-\alpha}$ with $b > 0$ at fixed d, k, then $0 < \alpha < k$ makes exclusion eventually disappear despite improving model accuracy; $\alpha > k$ yields an eventual crossing. At $\alpha = k$, the leading constants decide unless equal.*

The Gaussian law determines the rival's floor; the model's error and paired uncertainty determine the observed exclusion boundary. An independent evaluation reproduces the measured pattern across all nine contexts and 19 nonsaturated budgets, with sampling variation (Appendix G.5).

3 SEPARATING SYMMETRY-DEPENDENT ERROR

3.1 TARGET-PRESERVING TRANSFORMATIONS

For independent Haar rotations $Q \in O(n)$ and $U \in O(d)$, transform

$$(X, y, q, w) \longmapsto (QXU^\top, Qy, Uq, Uw). \tag{11}$$

This preserves the target and Gaussian task distribution. Row mixing preserves the full state $[X\ y]^\top[X\ y]$; conditional on that state, the support law is exactly the Haar orbit (Appendix B.3). Transporting recovered coefficients back gives

$$\widetilde{v}_{Q,U} = U^\top v(QXU^\top, Qy), \qquad \mathcal{K}_k(USU^\top, Ur) = U\mathcal{K}_k(S, r). \tag{12}$$

All transported coefficients are therefore compared with the original projector.

Let A average row mixing and B average transported feature rotations. These are commuting orthogonal projections in prediction L^2 (Elesedy & Zaidi, 2021). For $e = v - w$,

$$\begin{aligned} e_{00} &= ABe, & e_{10} &= (I - A)Be, \\ e_{01} &= A(I - B)e, & e_{11} &= (I - A)(I - B)e. \end{aligned} \tag{13}$$

The jointly invariant, row-only, coordinate-only, and interaction components are mutually orthogonal after population or full-orbit integration. Their energies sum to total coefficient error. The interaction is removed by either average and accounts for their overlapping contributions. The Krylov projector and query-linear projection commute with A, B; the decomposition also holds inside and outside each Krylov space and for offsets and query nonlinearity.

3.2 Class-preserving averaging

Proposition 5 (Symmetry projection). *For $T = AB$, $T\mathcal{C}_k \subseteq \mathcal{C}_k$ and*

$$\operatorname{dist}(Tf, \mathcal{C}_k) \le \operatorname{dist}(f, \mathcal{C}_k), \qquad U_{n,k}(Tv) = U_{n,k}(v) + \frac{1}{d} \sum_{(a,b)\neq(0,0)} \mathbb{E}\|(I - P_k)e_{ab}\|^2. \tag{14}$$

Proof. Every transformed class member transports into the same original Krylov space; averaging stays in that space. For $g \in \mathcal{C}_k$, $Tg \in \mathcal{C}_k$ and $\|Tf - Tg\|_{L^2} \le \|f - g\|_{L^2}$. Taking the infimum proves contraction. Orthogonality of (13) and the invariant teacher, applied to outside error, proves the usefulness identity. □

Thus averaging reveals departures while preserving $\mathcal{C}_k$. We apply this projection to the unpreconditioned class and evaluate the preconditioner menu on raw predictions.

3.3 Finite-grid estimation

We evaluate a full $m \times h$ cross of row and feature rotations. Two independent, pilot-corrected query banks estimate coefficient products without squared-estimation bias. Off-diagonal grid contractions recover the infinite-orbit component energies (Appendix B). The actual finite ensemble instead has expected error

$$R_{m,h} = E_{00} + E_{10}/m + E_{01}/h + E_{11}/(mh). \tag{15}$$

We report both infinite-orbit estimates and literal holdout errors of finite ensembles, retaining signed estimates before inference.

4 Experimental design

Models and independent units. We audit the released Garg regression checkpoint and canonical seed 12001: 12 causal blocks, width 256, eight heads, and unchanged native final heads (Garg et al., 2022a;b). The released checkpoint collection is also used by Fu et al. (2024b). Seed 12001 was trained for 500,001 Adam updates with the original curricula and all-prefix objective. One of the five registered independent runs has been fully evaluated. Each fixed-endpoint model is evaluated on independent prompts. Reported intervals quantify prompt uncertainty conditional on that model.

Paired context sweep. For each model, 1,024 independent episodes supply a 40×20 design and teacher. We evaluate nested prefixes $n \in \{20, 21, 22, 24, 25, 28, 30, 35, 40\}$ with 512 independent holdout queries shared across contexts. Primary risk is literal holdout MSE. Budgets are 1–19; $k = 20$ is a saturation control. The global preconditioner is frozen from 8,192 separate episodes per n. The context grid was fixed after exploration and evaluated on fresh episodes; evaluation streams are independent across models.

Crossed audit and readouts. Each model has a 256-prompt $n = 40$ audit and 128-prompt audits at each of $n = 22, 25$, all with 8×8 grids. A 128-query pilot, two independent 512-query coefficient banks, and 512 shared holdout queries separate coefficient, offset, and nonlinear error. At $n = 40$, intermediate probes use 32,768 calibration and 8,192 validation prompts; disjoint-half probes and an unchanged native final-head lens are controls. A matched 64-prompt subset extends the grid to 16×16. Short-context results use the native final head. Frozen affine references are fitted on separate calibration prompts for each model. Appendix D gives the calibration and training details.

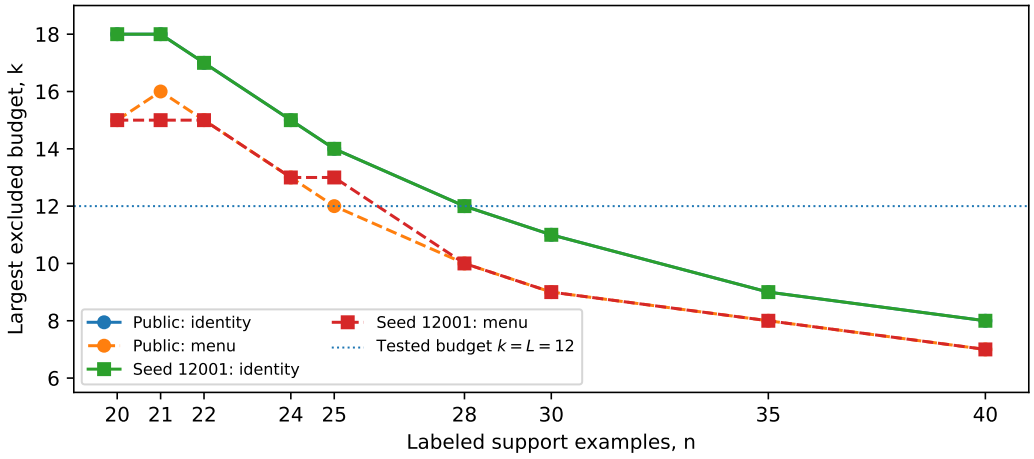

Figure 2: **Shorter contexts exclude larger families.** Curves give the largest budget with a positive simultaneous lower bound on oracle-minus-model error. The public identity curve is hidden beneath seed 12001's identical curve; menu boundaries differ by one update at two contexts. The horizontal line marks the tested one-update-per-layer hypothesis. The highest budgets depend on difficult prompts, unlike the robust 12-update comparison (Section 5.3).

Uncertainty and numerical controls. We bootstrap complete base episodes, keeping their contexts and query/orbit banks together. Native comparisons use 4,000 resamples with approximate max-$|t|$ simultaneous intervals; the crossed primary family includes all layers. Prompt-win fractions are descriptive. Exploratory oracle-ranked trimming uses 6,000 resamples, reselecting the trimming cutoff inside each resample. Production oracles use independently checked float64 constructions. Supplemental high-precision checks and the status of the complete boundary gate are distinguished in Appendix D.3. Synthetic null and alternative controls calibrate a reduced six-comparison family (Appendix F).

5 Results

The context sweep tests when prediction error can distinguish computational families; replication and symmetry audits assess the resulting exclusions.

5.1 Less accurate predictions can yield stronger exclusions

Figure 1 separates prediction accuracy from the strength of behavioral evidence. From forty to twenty-two examples, public-model RMS rises from 0.024 to 0.063, while the largest excluded identity budget rises from 8 to 17. The model worsens, but the rival deteriorates faster. The two audited models have matching identity boundaries at all nine contexts; menu boundaries agree at seven (Figure 2; Appendix E).

The budget-12 comparison makes this reversal concrete. With forty examples, the family is more accurate than the models. At $n = 22$, both 12-layer models outperform the budget-12 identity class by about 0.025 normalized MSE. They also beat the oracle-selected preconditioner menu; all four simultaneous intervals exclude zero (Table 1). These crossings reject the specified H_{12} description even with arbitrary support- and label-dependent scalar coefficients. The identity prompt-win fractions are 84% and 92% for the public model and seed 12001, respectively.

The exclusion also rules out close approximations. Corollary 2, applied to the existing simultaneous risk and advantage bounds, gives a prediction-distance lower bound of 0.088 signal-normalized RMS for the public identity class at $n = 22, k = 12$, versus model target RMS 0.063. The menu bound exceeds 0.051. Both models' bounds are reported in Table 11; Appendix B.9 derives their coverage-preserving transformation.

Table 1: **All six native exclusions reproduce in seed 12001.** Oracle-minus-model MSE, $\times 10^{-3}$, with approximate simultaneous 95% within-model prompt intervals shown for both models. The public checkpoint is the reference model.

(n, k)	Class	Public [interval]	Seed 12001 [interval]
$(22, 12)$	Identity	$25.294\,[20.022, 30.565]$	$24.555\,[20.126, 28.985]$
$(25, 12)$	Identity	$5.203\,[3.742, 6.664]$	$6.057\,[4.455, 7.658]$
$(22, 16)$	Identity	$6.728\,[4.364, 9.092]$	$7.068\,[4.604, 9.533]$
$(25, 14)$	Identity	$1.291\,[0.501, 2.080]$	$1.744\,[0.807, 2.680]$
$(22, 12)$	Menu	$13.456\,[9.781, 17.132]$	$12.769\,[9.342, 16.196]$
$(25, 12)$	Menu	$1.629\,[0.838, 2.421]$	$2.059\,[0.997, 3.121]$

5.2 Native class exclusions recur after independent training

Seed 12001 reproduces all six prespecified native contrasts, with positive simultaneous lower bounds (Table 1). It uses independent initialization, training, and evaluation streams. Appendix E.2 reports all 14 endpoints, including the intermediate and error-decomposition results.

5.3 Moderate-budget advantages survive removal of difficult prompts

At $n = 22, k = 12$, the identity and menu advantages remain positive in both models after removing the most oracle-difficult 10% of prompts. Retained mean advantages are about 0.014 for identity and 0.005 for the menu. The identity comparison at $(n, k) = (22, 16)$ also remains positive in both models. Table 7 gives the complete estimates and exploratory intervals, with the trimming cutoff reselected within each bootstrap resample.

The highest-budget frontier is more tail-dependent. After the same removal, public $n = 22, k = 17$ identity is unresolved, public $n = 21, k = 16$ menu favors the rival, and the $n = 25, k = 12$ menu comparison is unresolved in both models. These contrasts distinguish an advantage that persists away from difficult prompts from one sustained by them.

Hard prompts can amplify an advantage that persists without them. For a removed fraction α,

$$\mathbb{E}\Delta = (1 - \alpha)\mathbb{E}[\Delta \mid \text{retained}] + \alpha\mathbb{E}[\Delta \mid \text{removed}]. \tag{16}$$

At public $n = 22, k = 12$, roughly half the mean advantage comes from the hardest 10%, yet the retained mean stays large. At frontier budgets, that tail can sustain the entire positive mean. The trimmed and untrimmed means describe different parts of the task distribution.

5.4 Useful departures and symmetry-dependent error

The native crossings imply useful information outside $\mathcal{K}_{12}$ by (9). The final-head crossed audits measure this usefulness directly: at $n = 22$ and 25, the outside component helps both before and after joint averaging. At $n = 40$, individual predictions have a harmful outside component on average, while the joint-average component helps. Each sign is resolved by its audit's simultaneous interval (Table 12). The individual-prediction estimates average unaveraged Haar-grid cells, on samples separate from the native sweep.

This contrast links the two experiments: short contexts expose useful departures in native risk, whereas at forty examples symmetry-dependent error masks a useful component. There, 96–97% of coefficient error is symmetry-dependent, and the row/feature interaction contributes roughly 47% (Table 6). Joint averaging removes excess outside error while preserving useful information. Tables 8 and 10 retain the context-specific risks and component magnitudes.

5.5 Joint intermediate exclusions replicate more reliably than finite ensembles

At $n = 40$, joint-average layer-8/$\mathcal{K}_8$ and layer-9/$\mathcal{K}_9$ advantages have positive simultaneous lower bounds in both models (Figure 4). The public finite 8×8 ensembles also cross both floors. In

seed 12001, the layer-8 finite advantage is essentially zero and the layer-9 finite estimate is positive but unresolved. Finite-grid variation consumes almost all of the seed's layer-8 joint advantage, as predicted by (15).

The joint-limit result is therefore more stable across these models than its finite-ensemble realization. Appendix D.2 reports the probe and larger-grid controls, including the weaker seed-12001 comparisons. The low $n = 40$ floors constrain the budgets that these residual errors allow the risk test to exclude.

5.6 Fixed distortion and calibration controls

A frozen affine reference explains 31–35% of coefficient error on original, unrotated evaluation prompts. Its row invariance confines its contribution to the jointly invariant and coordinate-only components, leaving the large interaction component intact (Appendix B). The seed's larger surviving query-independent offset limits the corresponding improvement in total prediction risk (Appendix D.2).

Neither in-class control produces a false exclusion across 200 synthetic repetitions, and the designed alternative is detected in every repetition (at least one comparison). Appendix F reports the six-comparison calibration family, its complete-detection rate, and uncertainty intervals.

6 Relation to prior evidence

Complementary forms of evidence. Prior work combines algorithm constructions, restricted optimality and learning results, probes, and trajectory matching. Garg et al. (2022a); Akyürek et al. (2023); von Oswald et al. (2023) establish controlled ICL and algorithmic correspondences; Ahn et al. (2023); Mahankali et al. (2024); Zhang et al. (2024) analyze optimal or learned linear-attention predictors. Bai et al. (2023) give statistical guarantees and constructions for in-context algorithm selection. Our deliberately simple test adds a class-relative risk certificate to these complementary forms of evidence.

Implicit weights and richer updates. Akyürek et al. (2023) recover implicit weights by regressing query responses. Fu et al. (2024a) use induced weights and layerwise trajectories to compare GD, CG, BFGS, and Iterative Newton; their roughly three Newton iterations per middle layer describe a best-matching trajectory. Our budget is a lower bound from the best member of an entire span class, with its resolution set by the floor and residual prediction error (Appendix D.2). GD++ transforms features (von Oswald et al., 2023; Vladymyrov et al., 2024); Newton-type updates evolve a matrix state (Fu et al., 2024a); both exceed the premise of one original-Gram span expansion per layer. He et al. (2025) derive a centered-feature, debiased update for multi-head softmax attention. These provide concrete, richer alternatives to the tested linear-span hypothesis.

Spectra, context, and training variation. Average-case optimal *oblivious* polynomial methods are established (Pedregosa & Scieur, 2020); our finite-dimensional adaptive law applies MINRES results of Paquette & Trogdon (2023). Zhang et al. (2025) interpret training plateaus as principal-component regression in the large-context limit; Bordelon et al. (2026) analyze depth, width, context, and covariance scaling. Attention can generate degree exponential in depth (Li et al., 2025). These results distinguish architectural capacity and training dynamics from our observed output-class boundary. Poor targeted-attack transfer between *larger* same-recipe models suggests seed-dependent behavior (Anwar et al., 2025). Here the identity exclusion frontier matches across the two trained models despite this possibility. Distribution-shift comparisons supply complementary evidence (Hill et al., 2025).

Symmetry and identifiability. Elesedy & Zaidi (2021) establish the exact squared-risk gain from equivariant projection, used in Appendix B.4. Lutz et al. (2026) impose layerwise symmetries to expose classification recurrences; we average externally without changing the trained model. Our contribution is the combination with a class-preserving geometry and measured useful departures. Appendix H relates task-diversity, looped-computation, kernel-regression, language-model, and non-identifiability studies to this distinction.

7 DISCUSSION AND LIMITATIONS

What this means for algorithm-matching studies. Before interpreting an algorithm match, establish the family's best attainable risk and whether departures help solve the task. Risk gaps quantify separation from approximate explanations; class-preserving symmetries locate harmful variation. Shorter contexts expose useful departures even as prediction worsens. Appendix A illustrates a complementary design choice: the target and prior affect separation even with the observation model fixed.

Scope of the explanation. The rival grants arbitrary scalar adaptation within fixed directions. Aggregation across menu members, learned or changing preconditioners, multiple starting vectors, and query-dependent solvers enlarge it. The isotropic global calibration is scalar (Appendix B.7), while Jacobi and block-Jacobi use sample geometry. The Gaussian theorem orders marginal floors; model error and paired uncertainty determine the observed boundary. Above-floor risk leaves this test inconclusive, while direct functional comparisons can still distinguish predictors.

Interpreting the symmetry audit. Class preservation can fail for coordinate-tied preconditioners. The centered-feature update of He et al. (2025) is also row-sensitive. Its pure span extension loses one effective Gaussian sample, raising the floor, so the $n = 22$ crossing also excludes that family (Appendix B.10). Joint symmetry permits label-nonlinear maps such as $\|r\|^2 r$. Intermediate readouts establish decodability; the native lens fails at layers 8–9, and finite-ensemble confirmation varies across models. Averaging uses additional evaluations. These tests constrain compatible computations without identifying a unique optimizer.

Empirical scope. Results cover $d = 20$ noiseless isotropic regression, the released model, and seed 12001. One of the five registered independent runs has been fully evaluated. Bands quantify prompt uncertainty conditional on each model; synthetic calibration covers a reduced family. Short-context joint residuals are less precise and frontier exclusions more tail-dependent. The independent Gaussian-law comparison supports aggregate identity floors (Appendix G.5), complementing selected high-precision checks; the complete production boundary gate is not included in the reported validation (Appendix D.3).

8 CONCLUSION

Prediction accuracy and the strength of behavioral evidence are different objectives. Adaptive class oracles make that distinction measurable, and Gaussian MINRES theory explains why the rival improves as context grows. Native predictions in the released model and an independent replication cross depth-sized floors on shorter contexts, including after difficult-tail removal. Crossed symmetries distinguish useful information from residual error. Together, these results show how to choose comparisons that reveal computational differences rather than merely rank prediction accuracy.

GENERATIVE AI USAGE STATEMENT

Large language models (LLMs) were used extensively throughout this project. The tools used were ChatGPT Pro (GPT-5.5 Pro, GPT-5.6, and GPT-6) and Claude Fable 5.1. Assistance covered research development, theoretical reasoning, experimental design and implementation, analysis, and manuscript preparation, including substantive mathematical arguments and experimental suggestions.

Research questions, theory, and proofs. LLMs assisted in developing the conceptual framework, proposing and refining hypotheses, formulating mathematical claims, and assessing alternative computational explanations. They proposed proof strategies, intermediate arguments, and counterexamples; helped examine assumptions and potential gaps; and drafted and revised substantial portions of the proofs. This included assistance with substantive proof steps as well as exposition and notation.

Experimental methods, implementation, and synthetic data. LLMs assisted with the design of the native context sweep, crossed symmetry audit, comparison classes, controls, estimators, and numerical checks, including suggestions for experimental parameters. They were used extensively to write, revise, and debug code for model training and inference, synthetic-task sampling, Krylov projections, uncertainty estimation, numerical validation, and resumable experiment workflows. They also helped organize, extract, and reformat saved numerical results into analysis tables. Synthetic regression tasks and analytic controls were produced by executing the specified mathematical sampling and evaluation procedures with recorded seeds; LLM responses were not used as numerical observations or experimental labels. Reported model measurements came from executed computational runs rather than generated descriptions of hypothetical outcomes.

Literature review and source verification. The literature review combined papers the authors had already read and selected as relevant with additional papers recommended by LLMs. LLMs helped identify connections, summarize sources, examine novelty and attribution, and format references and citation locators. For sources incorporated following an LLM recommendation, the authors personally read the source papers and checked their relevance and the claims attributed to them. Literature choices therefore combined the authors' existing knowledge and reading with AI-assisted discovery and synthesis.

Analysis, interpretation, and presentation. LLMs assisted with analysis of saved outputs, recomputation of summaries, interpretation of results, examination of competing explanations, and critical feedback on the study's reasoning and presentation. They were also used extensively to draft and revise manuscript text, suggest structure and titles, prepare tables, write and edit scientific-figure code, and format the manuscript in LaTeX. The study concerns synthetic numerical regression tasks and trained models, rather than survey, interview, transcription, or qualitative-coding data. No translation was involved.

Authors' review and responsibility. The authors personally worked through the mathematical proofs and reviewed the code used for the reported experiments line by line. Implementation checks also included the analytic controls, software tests, cross-implementation comparisons, and precision checks documented in the paper. Some checking and analysis code was itself developed with LLM assistance. These checks are described with their actual scope; the manuscript identifies the scope of the completed validation. The authors assessed AI-generated suggestions and revisions and take responsibility for the research decisions and final content, including the mathematical claims and proofs, code, references, analyses, figures, and text produced with AI assistance.

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*, 2023. URL https://arxiv.org/abs/2306.00297.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.15661.

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. URL https://arxiv.org/abs/2411.05189.

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*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b2e63e36c57e153b9015fece2352a9f9-Abstract-Conference.html.

Blake Bordelon, Mary Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.01098.

Ioana Dumitriu and Alan Edelman. Matrix models for beta ensembles. *Journal of Mathematical Physics*, 43(11):5830–5847, 2002. doi: 10.1063/1.1507823. URL https://arxiv.org/abs/math-ph/0206043.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969, 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *Advances in Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=L8h6cozcbn. Earlier preprint title: Transformers Learn Higher-Order Optimization Methods for In-Context Learning: A Study with Linear Models.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Official implementation of the second-order in-context linear regression study, 2024b. URL https://github.com/DeqingFu/transformers-icl-second-order/tree/c99333c5fabfb41f2d51891b4f321030d4174bb9. GitHub repository; pinned source snapshot, September 2026.

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*, 2022a. URL https://arxiv.org/abs/2208.01066.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. In-context learning: Training code, configurations, and public checkpoints, 2022b. URL https://github.com/dtsip/in-context-learning/tree/338337192195827c58356f9b627e5f999ea703d2. GitHub repository; pinned source snapshot, September 2026.

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, 2024. URL https://proceedings.mlr.press/v235/gatmiry24b.html.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11398–11442, 2023. URL https://proceedings.mlr.press/v202/giannou23a.html.

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, 2025. URL https://proceedings.mlr.press/v267/he25q.html.

Joshua Hill, Benjamin Eyre, and Elliot Creager. Transformers don't in-context learn least squares regression, 2025. URL https://arxiv.org/abs/2507.09440. arXiv:2507.09440.

Yingcong Li, Xiangyu Chang, Muti Kara, Xiaofeng Liu, Amit Roy-Chowdhury, and Samet Oymak. When and how unlabeled data provably improve in-context learning, 2025. URL https://arxiv.org/abs/2506.15329v1. arXiv:2506.15329v1.

Patrick Lutz, Themistoklis Haris, Arjun Chandra, Aditya Gangrade, and Venkatesh Saligrama. Symmetry reveals layerwise dynamics: How transformers perform in-context classification. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2604.11613.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.03576.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.20914.

Elliot Paquette and Thomas Trogdon. Universality for the conjugate gradient and MINRES algorithms on sample covariance matrices. *Communications on Pure and Applied Mathematics*, 76 (5):1085–1136, 2023. doi: 10.1002/cpa.22081. URL https://arxiv.org/abs/2007.00640.

Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7553–7562, 2020. URL https://proceedings.mlr.press/v119/pedregosa20a.html.

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*, 2023. URL https://arxiv.org/abs/2306.15063.

Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. Position: Do pretrained transformers learn in-context by gradient descent? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44712–44740, 2024. URL https://proceedings.mlr.press/v235/shen24d.html.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2507.08802.

Max Vladymyrov, Johannes von Oswald, Mark Sandler, and Rong Ge. Linear transformers are versatile in-context learners. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2402.14180.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2212.07677.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Hassler Whitney. On functions with bounded nth differences. *Journal de Mathématiques Pures et Appliquées*, 36:67–95, 1957.

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

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

Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew M. Saxe. Training dynamics of in-context learning in linear attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76047–76087, 2025. URL https://proceedings.mlr.press/v267/zhang25br.html.

A AN EXPONENTIAL-FAMILY ILLUSTRATION: THE OBSERVABLE MATTERS

This example isolates the same design principle without a learned model. Fixing an observation family does not fix the separation between a prediction target and a rival function class. The loss, target coordinates, prior, and prompt length must be specified as well.

For a regular minimal exponential family with log-partition A and empirical sufficient-statistic mean $\widehat{\mu}$, the interior plug-in estimate of the natural parameter is $\widehat{\eta} = \nabla A^*(\widehat{\mu})$, where A^* is the Legendre dual (Wainwright & Jordan, 2008). On an open convex domain, for a symmetric matrix B,

$$\nabla A^*(\mu) = B\mu + a \quad \Longleftrightarrow \quad A^*(\mu) = \tfrac{1}{2}\mu^\top B\mu + a^\top \mu + c. \tag{17}$$

The forward implication follows by integration along line segments, and the reverse by differentiation. Thus a quadratic dual gives an exact affine null. This classical convex identity is a calibration example, not the main theorem of the paper.

The finite-prior target can differ from the plug-in map. Let $Y_i \mid \lambda \sim \mathrm{Poisson}(\lambda)$ and $\lambda \sim \mathrm{Gamma}(\alpha, \beta)$ in shape–rate parameterization. Multiplying likelihood and prior gives

$$\lambda \mid D \sim \mathrm{Gamma}(\alpha + n\widehat{\mu}, \beta + n). \tag{18}$$

Under squared error, two different estimands have posterior means

$$F_{\mathrm{rate}}(\widehat{\mu}) = \frac{\alpha + n\widehat{\mu}}{\beta + n}, \qquad F_{\log}(\widehat{\mu}) = \psi(\alpha + n\widehat{\mu}) - \log(\beta + n). \tag{19}$$

At fixed n, the first map is affine in $\widehat{\mu}$, whereas the second is not. Its curvature is $n^2\psi_2(\alpha + n\widehat{\mu})$; on a fixed bounded mean interval it tends to zero as $\alpha \to \infty$ with n fixed. A task-family label such as "Poisson" therefore does not imply a usable affine gap.

For a target F on a specified set K, let $E_K(F) = \inf_{L \text{ affine}} \sup_{\mu \in K} \|F(\mu) - L(\mu)\|$. Every admissible triple gives

$$E_K(F) \geq \frac{1}{4}\|F(x+h) - 2F(x) + F(x-h)\|. \tag{20}$$

Indeed, affine second differences vanish, and the three error terms have total coefficient magnitude four. The displayed lower bound follows directly from the triangle inequality. For historical context on finite-difference approximation, see Whitney (1957). A nonzero bound certifies target separation from the stated affine class, not separation between all direct and iterative implementations.

Two distinctions prevent an overinterpretation. First, $\mathbb{E}[\lambda \mid D]$ and $\mathbb{E}[\log \lambda \mid D]$ are different estimands, not one estimator expressed in different units. Second, an injective output transformation R preserves exact equality between fixed candidate maps: $R \circ f = R \circ g$ if and only if $f = g$. When changing coordinates for the same candidates, their class must also be transformed. Resetting the null to "affine in the new coordinates" changes the hypothesis. The role of this illustration is therefore to motivate the explicit observable and rival class in the regression audit, not to claim that non-Gaussian data automatically identify optimization.

B Additional derivations and estimator details

B.1 The linear-span restriction and its limits

Starting at zero, suppose an update can use scalar combinations of previous iterates and gradients of $\ell(z) = \frac{1}{2}z^\top Sz - r^\top z$. At the first update every direction is proportional to r. If all iterates through step k lie in $\mathcal{K}_k(S, r)$, applying S to them and subtracting r produces vectors in $\mathcal{K}_{k+1}(S, r)$. Induction proves inclusion of the update sequence, even when the scalar coefficients are arbitrary functions of all support data. This argument does not allow an arbitrary matrix to act on a vector or supply additional starting vectors for free.

For fixed inverse-preconditioner $C \succ 0$, replacing the gradient directions by $C(Sz - r)$ gives $\mathcal{K}_k^C$. Equivalently,

$$\mathcal{K}_k^C = C^{1/2}\mathcal{K}_k(C^{1/2}SC^{1/2}, C^{1/2}r), \tag{21}$$

since multiplication of its jth power by $C^{1/2}$ produces $(CS)^jCr$. An orthonormal basis in transformed coordinates is generally not orthonormal after this mapping; Euclidean reorthogonalization is required for the prediction oracle.

For generic full-rank Gaussian prompts, $\mathcal{K}_d$ saturates the coefficient space almost surely. The finite-dimensional polynomial representation of an inverse then gives no useful complexity identification by itself. An adaptive coefficient may hide expensive work. Likewise, the class does not encompass unrestricted prompt-dependent preconditioners, for which $C = S^{-1}$ solves the noiseless task in one nominal update.

B.2 Direct functional distance and useful outside error

For the coefficient $v = \mathbb{E}_q[qf]$, define $h = f - q^\top v$. Then $\mathbb{E}_q[qh] = 0$ and, for any coefficient z,

$$\frac{1}{d}\mathbb{E}_q(f - q^\top z)^2 = \frac{1}{d}\|v - z\|^2 + \frac{1}{d}\mathbb{E}_q h^2. \tag{22}$$

Minimizing over $z \in \mathcal{K}_k$ proves (5). Substituting $z = w$ and decomposing $v - w$ into its P_k and $I - P_k$ parts proves (9). The identity also clarifies why the risk comparison is one-sided: a predictor with a large unrelated error can fail to beat the class floor while remaining far from every class member.

B.3 Exact conditional row resampling

Write $Z = [X\ y]$. For any fixed $Q \in O(n)$, the law of (QX, Qy) equals that of (X, y): QX has the same Gaussian design distribution and $Qy = (QX)w$. More generally the same argument holds with independent isotropic Gaussian observation noise. If $Z_1^\top Z_1 = Z_2^\top Z_2$, the map taking corresponding columns of Z_1 to those of Z_2 defines an isometry on their column spans and extends to an element of $O(n)$. Each Gram fiber is therefore a single compact-group orbit. Its unique invariant probability measure is the Haar pushforward, which is the conditional law almost everywhere.

This orbit statement requires no inequality between n and d; invertibility is needed elsewhere for the noiseless target representation, not for the group action. For a prefix-processing model, however, mixing an entire long prompt changes its earlier prefix statistics. A prefix diagnostic must condition on that prefix's state. The native experiment directly evaluates each specified support prefix and does not mistake a changed prefix information set for irrelevant variation.

B.4 Orthogonal averaging and four components

For a target fixed by an orthogonal group average T, the exact identity of Elesedy & Zaidi (2021, Lemma 6) specializes to

$$\mathcal{R}(f) - \mathcal{R}(Tf) = \|f - Tf\|_{L^2}^2.$$

Indeed, $f - Tf$ is orthogonal both to Tf and to the target, so the cross term in the squared error vanishes. This is the risk-reduction identity used by the crossed decomposition.

Haar averaging is idempotent because the product of two independent Haar group elements is Haar. It is self-adjoint under the invariant distribution by replacing a group element with its inverse. It is

consequently an orthogonal projection. Row and feature actions commute, so their two projections A, B commute. Expanding $(A + (I - A))(B + (I - B))$ proves (13). Products of unlike projector factors contain either $A(I - A)$ or $B(I - B)$, proving orthogonality.

The teacher coefficient is fixed by both averaging operators, with feature transport included. Krylov projectors commute with them by (12). The same orthogonality therefore applies after projection into or outside the Krylov space. All three non-invariant outside-error components worsen expected target risk; only the jointly invariant component can carry target alignment after orbit integration. This is a property of population or orbit-integrated errors, not a pointwise claim about four perpendicular vectors at one support.

B.5 Unbiased coefficient and crossed estimates

For an independently fitted pilot $p(q) = a + b^\top q$, let $\rho(q) = f(q) - p(q)$. A bank of m_q independent Gaussian queries gives

$$\widehat{v} = b + \frac{1}{m_q} \sum_j q_j \rho(q_j), \qquad \widehat{a} = a + \frac{1}{m_q} \sum_j \rho(q_j). \tag{23}$$

Since $\mathbb{E}q = 0$ and $\mathbb{E}qq^\top = I$, these are unbiased conditional on the pilot. For independent banks A and B, $\mathbb{E}[\widehat{v}_A^\top H \widehat{v}_B] = v^\top H v$ for a support-determined projector H. Query coordinates may be shared across orbit cells within one bank, but the two complete banks must be independent. This prevents correlated within-bank estimation errors from being mistaken for orbit structure.

Write D_0, R_0, F_0, Z_0 for same-cell, same-row/different-feature, same-feature/different-row, and different-both cross-bank contractions. The component estimators are

$$\widehat{E}_{00} = Z_0, \quad \widehat{E}_{10} = R_0 - Z_0, \quad \widehat{E}_{01} = F_0 - Z_0, \quad \widehat{E}_{11} = D_0 - R_0 - F_0 + Z_0. \tag{24}$$

To see (24), condition on a fixed row level and average two independent feature levels. Their inner-product expectation is $\|\mathbb{E}_U z(Q, U)\|^2 = E_{00} + E_{10}$ after row integration. The analogous feature-conditioned product gives $E_{00} + E_{01}$. Independent levels of both give E_{00}; the same-cell product gives the sum of all four. Subtraction yields the component estimators. Expanding the squared mean of the finite grid yields (15).

For functional nonlinear remainders, common independent holdout queries estimate products between orbit cells. Subtracting cross-bank products of the pilot-corrected means and coefficients separates offsets and query-linear components. Independent holdout queries for the two factors would instead estimate a product of marginal means and would not recover a functional inner product. All outer uncertainty calculations keep the complete crossed grid attached to its base prompt.

B.6 Fixed affine distortion and the interaction

Let a diagnostic reference be $d(D) = c + (M - I)w$. Under feature rotation its transported value is $U^\top c + U^\top (M - I)Uw$. Put $\alpha = \mathrm{tr}(M)/d$ and $T = M - \alpha I$. Haar conjugation gives

$$\mathbb{E}_U U^\top M U = \alpha I, \qquad \mathbb{E}_U U^\top c = 0, \tag{25}$$

because the first average commutes with every orthogonal matrix and its trace is fixed. Scalar shrinkage survives both averages; the fixed bias and trace-free distortion belong to the coordinate-only component. Since d is row-invariant, $(I - A)Bd = (I - A)(I - B)d = 0$, so subtracting it cannot change row-only or interaction error.

A fixed bias can nonetheless look isotropic relative to the random Krylov directions. For $v = w + c$, rotation invariance of the random subspace implies $\mathbb{E}P_k = (k/d)I$ at generic rank k, and therefore

$$\mathbb{E}\|(I - P_k)c\|^2 = \frac{d - k}{d}\|c\|^2. \tag{26}$$

A flat residual per Krylov direction is thus consistent with a coordinate-fixed artifact; it does not establish independent isotropic noise or a particular solver.

B.7 THE ISOTROPIC GLOBAL MAP AND FINITE-SAMPLE PRECONDITIONING

Consider the population linear calibration $\min_C \mathbb{E}\|Cr - w\|^2$ with $r = Sw$, $w \sim N(0, I_d)$, and Gaussian X independent of w. Its normal equation is

$$C\mathbb{E}[rr^\top] = \mathbb{E}[wr^\top].$$

Since $\mathbb{E}[wr^\top] = \mathbb{E}S = I$ and rotation invariance makes $\mathbb{E}S^2$ scalar, it suffices to compute its trace. Diagonal entries of S have variance $2/n$ and off-diagonal entries have variance $1/n$, so

$$\mathbb{E}\,\mathrm{tr}(S - I)^2 = \frac{d(d+1)}{n}, \qquad \mathbb{E}S^2 = \left(1 + \frac{d+1}{n}\right) I.$$

Hence the unique population minimizer is $C_\star = nI/(n+d+1)$, which is SPD. This is the isotropic optimal preconditioner in Ahn et al. (2023, Theorem 1), in the present $S = X^\top X/n$ normalization. The same factor appears in the population-gradient-flow limit of Zhang et al. (2024, Theorem 4), under their initialization assumptions. Multiplication by this positive scalar does not change any Krylov space. This establishes the population redundancy of this particular global calibration under isotropy, not an equivalence of every finite-sample preconditioner. Jacobi and block-Jacobi use each realized S; individual diagonal entries fluctuate with standard deviation $\sqrt{2/n}$. The evaluated finite calibration and all promptwise menu choices remain distinct numerical comparisons.

B.8 NEWTON–SCHULZ DEGREE AND CLASS SATURATION

For the Iterative Newton initialization and recurrence used by Fu et al. (2024a), $M_0 = \alpha S$ and $M_{j+1} = 2M_j - M_j S M_j$, the following degree expansion follows by direct multiplication:

$$I - SM_{j+1} = I - 2SM_j + (SM_j)^2 = (I - SM_j)^2.$$

Induction gives $I - SM_j = (I - \alpha S^2)^{2^j}$. Since M_j is a polynomial in S,

$$M_j r = g_j(S)r, \qquad g_j(s) = \frac{1 - (1 - \alpha s^2)^{2^j}}{s}.$$

For nonzero α, the numerator has degree 2^{j+1} and is divisible by s, so g_j has degree $2^{j+1} - 1$ and its output is in $\mathcal{K}_{2^{j+1}}$. Normalizing the Gram matrix by n rescales α and the inverse representation; it does not change this degree statement. With the alternative initialization $M_0 = \rho I$, the residual is $(I - \rho S)^{2^j}$ instead, giving enclosing budget 2^j.

At generic $d = 20$, $\mathcal{K}_{20} = \mathbb{R}^{20}$. The first initialization reaches an enclosing budget of 32 after four iterations. This saturates the *comparison class*, not the convergence of M_j. Indeed, the residual at an eigenvalue is $(1 - \alpha s_i^2)^{2^j}$; it is not zero at finite j unless the initial factor is zero. Exact finite termination therefore does not follow from finite-dimensional polynomial interpolation. The observed exclusions permit these inverse-like recurrences but do not identify one of them.

B.9 SHARP RISK-ONLY DISTANCE BOUND

Let $b(Z) = \mathbb{E}[Y \mid Z]$ for any square-integrable regression problem. Conditional-mean orthogonality gives excess risk $r_f = \mathcal{R}(f) - \mathcal{R}(b) = \|f - b\|_{L^2}^2$. For any class $\mathcal{C}$ put $a_\mathcal{C} = \inf_{g \in \mathcal{C}} \|g - b\|_{L^2}^2$. For every $g \in \mathcal{C}$,

$$\sqrt{a_\mathcal{C}} \le \|b - g\|_{L^2} \le \sqrt{r_f} + \|f - g\|_{L^2}.$$

Taking the infimum proves Corollary 2, including nonconvex classes. Hence an ε-approximate class explanation is excluded if $\sqrt{r_f} + \varepsilon < \sqrt{a_\mathcal{C}}$.

Sharpness already holds for a nonzero closed linear subspace. Let $g_\star = P_\mathcal{C} b$ and $a = \|b - g_\star\|^2 > 0$. For $0 \le r_f \le a$, the predictor

$$f = g_\star + (1 - \sqrt{r_f/a})(b - g_\star)$$

has excess risk r_f and distance $\sqrt{a} - \sqrt{r_f}$ to the class. For $r_f \ge a$, take a unit vector $u \in \mathcal{C}$ and $f = g_\star + \sqrt{r_f - a}\, u$. It belongs to $\mathcal{C}$ and has that risk. Thus above-floor scalar risk alone cannot certify nonmembership, even though richer behavioral observations can. If $a = 0$, no positive universal lower bound follows from the two risks.

Propagating the existing simultaneous bounds. In the noiseless audit, write $R = R_n(f)$ and $\Delta = \inf_{g\in\mathcal{C}} R_n(g) - R$. Corollary 2 gives $\operatorname{dist}(f,\mathcal{C}) \geq \sqrt{R+\Delta} - \sqrt{R}$ when $\Delta > 0$, for both the identity class and the menu union. On any joint confidence event with $R \leq \overline{R}$ and $\Delta \geq \underline{\Delta} > 0$,

$$\boxed{\operatorname{dist}(f,\mathcal{C}) \geq \sqrt{\overline{R}+\underline{\Delta}} - \sqrt{\overline{R}} = \frac{\underline{\Delta}}{\sqrt{\overline{R}+\underline{\Delta}} + \sqrt{\overline{R}}}.}$$

For fixed positive Δ, the rationalized expression decreases with R; for fixed R it increases with Δ. This proves the displayed bound. Native model risk and oracle advantage are in the same recorded max-$|t|$ family. The transformation therefore inherits its approximate 95% simultaneous coverage within each model, with no additional resampling. Separate trained models retain separate confidence events. Table 11 gives the four principal comparisons; its displayed lower bounds are rounded downward. The numerical validation status of the source risks and floors is unchanged.

B.10 A CENTERED UPDATE CAN BE ROW-MIXING-SENSITIVE

The debiased-GD predictor of He et al. (2025, arXiv:2503.12734v2, Eq. (4.11)) uses $\bar{x}_i = x_i - n^{-1}\sum_j x_j$. Its coefficient is

$$v_{\mathrm{ctr}} = \frac{\eta}{n} X^\top H_n y, \qquad H_n = I - \frac{\mathbf{1}\mathbf{1}^\top}{n}.$$

After $(X,y) \mapsto (QX,Qy)$ it becomes $\eta X^\top Q^\top H_n Q y/n$. General Haar Q need not commute with H_n, despite preserving $X^\top X$ and $X^\top y$. Thus row sensitivity can be a concrete algorithmic feature, not random noise. Permutations preserve H_n; dense orthogonal mixing tests a stronger invariance.

The pure centered-span family has a larger Gaussian floor. Define $S_c = X^\top H_n X/n$, $r_c = X^\top H_n y/n = S_c w$, and

$$\mathcal{C}_k^{\mathrm{ctr}} = \{(D,q) \mapsto q^\top z(D) : z(D) \in \mathcal{K}_k(S_c, r_c)\}.$$

Coefficients may depend on the full original support and labels, but the output directions are confined to this centered space. The one-step surrogate above belongs to it at $k = 1$. For $n \geq d+1$ and $1 \leq k < d$, its target-risk floor satisfies

$$\boxed{F^2_{\mathrm{ctr},n,d,k} = F^2_{n-1,d,k} > F^2_{n,d,k}.}$$

To prove the equality, choose an orthogonal matrix O with $OH_nO^\top = \operatorname{diag}(I_{n-1}, 0)$. The first $n-1$ rows $\widetilde{X}$ of OX are iid Gaussian and independent of w, and

$$X^\top H_n X = \widetilde{X}^\top \widetilde{X}, \qquad S_c = \frac{n-1}{n}\widetilde{S}, \qquad r_c = \frac{n-1}{n}\widetilde{S}w, \quad \widetilde{S} = \frac{\widetilde{X}^\top \widetilde{X}}{n-1}.$$

Multiplying the matrix and starting vector by nonzero scalars leaves the Krylov span unchanged. Proposition 1 applies to the pointwise projection on that span, even when coefficients may use all the original data. Averaging gives the equality, and Proposition 3 gives the strict inequality.

Thus a population crossing below the uncentered floor, including the $n = 22, k = 12$ comparison, also excludes this pure centered-span family. This is a consequence of the Gaussian law, not a new fitted baseline. It does not transfer to a rule selecting between centered and uncentered spaces separately on each prompt: that rule's floor is the expectation of their promptwise minimum. The deduction concerns the exact centered surrogate and its span extension, rather than every softmax mechanism approximately resembling the surrogate in He et al. (2025).

C Polynomial calibration and saved-shard agreement

C.1 Fixed coefficients are a smaller class

Let $\widehat{w} = g(S)r$ with one degree-$(k-1)$ polynomial g shared by all prompts, and define the residual $q(s) = 1 - sg(s)$. Under noiseless isotropic teachers, the normalized error is

$$\frac{1}{d}\mathbb{E}\|q(S)w\|^2 = \int q(s)^2 \, d\mu_{n,d}(s), \tag{27}$$

where $\mu_{n,d}$ is the expected normalized spectral measure of S. The minimum over degree-k residuals with $q(0) = 1$ is the Christoffel value

$$F^2_{k,\mathrm{fixed}} = \left(\sum_{j=0}^{k} p_j(0)^2 \right)^{-1}, \tag{28}$$

with p_j orthonormal for $\mu_{n,d}$. To derive it, expand $q = \sum c_j p_j$ and minimize $\sum c_j^2$ subject to $\sum c_j p_j(0) = 1$. Cauchy–Schwarz gives the value and the equality-achieving coefficient vector. This is the average-case oblivious polynomial-optimization setting studied by Pedregosa & Scieur (2020).

In the Marchenko–Pastur limit with $0 < \gamma = d/n < 1$, take

$$t(s) = \frac{s - 1 - \gamma}{2\sqrt{\gamma}}, \qquad p_j(s) = U_j(t(s)) + \sqrt{\gamma}\, U_{j-1}(t(s)), \tag{29}$$

where $U_{-1} = 0$. Then $p_j(0)^2 = \gamma^{-j}$, so

$$F^2_{k,\mathrm{MP}} = \frac{(1-\gamma)\gamma^k}{1 - \gamma^{k+1}}, \qquad F^2_{k,\mathrm{MP}} \longrightarrow \frac{1}{k+1} \text{ as } \gamma \uparrow 1. \tag{30}$$

These are calibration formulas, not finite-dimensional adaptive oracle values. Label-dependent coefficients enlarge the class, so $F_{n,k} \le F_{k,\mathrm{fixed}}$. Neither the asymptotic expression nor an empirically optimized high-degree polynomial is substituted for the finite-prompt projection in the primary results.

C.2 Does a fixed filter reproduce the learned predictions?

A separate CPU-only analysis uses the 96 crossed-discovery prompt shards. It fits one fixed polynomial per layer and outer training fold to the learned joint coefficients, not to the regression teacher. Four outer held-out folds and inner degree/regularization selection keep each evaluated prompt out of its fitted predictor's training. A global spectral coordinate map is shared across prompts. Crossed and cross-bank products score agreement with the infinite joint predictor rather than charging finite-grid variation to the candidate.

Table 2: **Exploratory fixed-filter agreement on crossed-discovery data.** Full disagreement includes the learned query offset and nonlinear component. Different outer folds can select different degrees; each fitted candidate remains fixed across its held-out prompts. "Inverse" is the exact noiseless regression predictor, not a learned model.

Layer	Selected residual degrees	Agreement MSE	RMS (% signal)	Inverse agreement MSE
8	20,24,32	3.3602×10^{-5}	0.580	1.3821×10^{-4}
9	24,32	2.7331×10^{-5}	0.523	2.4814×10^{-5}
12	20,24,32	6.6928×10^{-5}	0.818	4.3548×10^{-5}

At layer 8, the selected polynomial captures part of a model-specific departure from exact inversion. At layer 9 and the final layer, the inverse reference is competitive or better, so a coarse approximation of the main mapping does not explain the small residual errors. Degrees 20–32 also do not support one degree per layer. Finally, a fitted polynomial gives an upper bound on best class agreement, not a lower bound: an imperfect fit cannot establish that prompt-adaptive coefficients are necessary. Cross-fitted score intervals are exploratory because outer training sets overlap and high-degree fits may extrapolate poorly to spectral tails.

D IMPLEMENTATION, CONTROLS, AND NUMERICAL STATUS

D.1 NATIVE PREDICTOR, CALIBRATION, AND INDEPENDENT TRAINING

Continuous inputs are interleaved feature and label tokens. A context with n labeled supports places its query at token index $2n$. Primary predictions use each checkpoint's original final normalization and head. Dense/cached agreement, transformed-batch agreement, upstream GPT-2 parity, and float32/float64 comparisons precede inference. Prefix checks ensure that changing the query label or future labels cannot change that query's prediction. Teacher prior $N(0, I_d)$ and error normalization by d do not imply rescaling the model's input labels.

The registered independent roster is 12001–12005; seed 12001 is the sole independently trained model reported here. It is trained to 500,001 updates with Adam, learning rate 10^{-4}, batch size 64, no dropout, and float32 with TF32 disabled. Dimensions start at five and increase by one every 2,000 updates to 20; total examples start at 11 and increase by two every 2,000 updates to 41. Loss is averaged over prediction positions. Validation is logged without selecting an endpoint. Each run has independent initialization and data streams. Partial checkpoints and repeated prompt evaluations do not increase the independent-model count.

For each intermediate block $\ell < 12$, let $h_\ell(D, q) \in \mathbb{R}^{256}$ be the query-position residual-stream vector immediately after the block and before the final layer normalization. The backbone is frozen. On $N_{\mathrm{tr}} = 32{,}768$ independent prompts with query labels $t_j = q_j^\top w_j$, standardize each feature coordinate using its training mean μ_ℓ and standard deviation s_ℓ, and write $\widetilde{h}_{\ell j} = (h_{\ell j} - \mu_\ell)/s_\ell$ coordinatewise. For each penalty λ, fit

$$(\widehat{a}_{\ell,\lambda}, \widehat{b}_{\ell,\lambda}) \in \arg\min_{a,b} \left\{ \frac{1}{N_{\mathrm{tr}}} \sum_j (a^\top \widetilde{h}_{\ell j} + b - t_j)^2 + \lambda \|a\|^2 \right\}.$$

The intercept is unpenalized. Independent validation MSE on 8,192 prompts selects λ separately per layer from $\{0, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$. The solver floors feature variances at 10^{-20}; at zero penalty it discards standardized-covariance eigenvalues at or below $10^{-10} \max\{\lambda_{\max}, 1\}$. Standardized coefficients are mapped back to an affine readout on the original residual stream. Primary final-layer outputs use the checkpoint's original final normalization and head, replacing any fitted final-layer diagnostic.

The half-A and half-B controls fit on disjoint halves of the training sample and are frozen before matched evaluation. The native lens applies the original final normalization and head at earlier blocks without fitting probe parameters. Short-context crossed results use only the native final prediction. The separate affine coefficient-error reference uses 512 calibration prompts per model; this count does not describe probe training or validation. Its parameters are frozen before evaluation. The first 96-prompt discovery and the exploratory fixed-filter fits in Appendix C are separate from the current 256-prompt primary decompositions.

Jacobi uses $\mathrm{diag}(S)^{-1}$; block maps invert consecutive diagonal blocks of sizes two and four. The global SPD map is calibrated on 8,192 independent episodes per model/context. This is a specified fitted map, not an optimum over all SPD preconditioners. The menu takes a minimum of exact target distances within the computed spaces, not a minimum of noisy model-distance estimates.

Immutable source and model identities. Both cited repositories are pinned to immutable commits in the bibliography. The public checkpoint's SHA-256 begins `7fe030f1d3dd631a`; seed 12001's begins `dae01d35b91c93f9`. Repository pins identify source snapshots consulted for this manuscript, rather than retroactively asserting which Git commit produced historical training.

D.2 RISK RESOLUTION AT $n = 40$

The native test can exclude a class only if its target floor exceeds model risk. The same condition applies to a joint-average probe: a remaining offset, query-nonlinear error, or within-space error can prevent a crossing even when its coefficient uses higher-budget directions. Table 9 reports the matched floors from the 256-prompt crossed audits, rather than an asymptotic or fixed-polynomial substitute. The floor falls by more than four orders of magnitude from $k = 8$ to 16. Consequently,

the layer-8 and layer-9 exclusions provide lower bounds, not a count of all computation available in those layers.

Public final joint risk	4.4993×10^{-5}
Seed-12001 final joint risk	1.2231×10^{-4}

The seed's latter value contains coefficient error 1.0456×10^{-5}, constant-in-query offset energy 1.1058×10^{-4}, and genuine query-nonlinearity energy 1.2705×10^{-6}. This is why a 97.1% removable *coefficient-error* fraction does not imply a 97.1% removable full prediction-risk fraction. The final native head and separately fitted intermediate readouts are different decoders; their joint-risk ordering is not an intervention demonstrating that a later layer destroys an earlier solution.

The primary intermediate measurements retain the distinction between joint-limit and finite-ensemble risks. For seed 12001 at layer 8, these are approximately 0.000524 and 0.000884, compared with the $\mathcal{K}_8$ oracle risk 0.000886. Its finite advantage is 2.16×10^{-6} with simultaneous interval $[-2.46, 2.51] \times 10^{-4}$; at layer 9 the finite advantage is 1.18×10^{-4} with interval $[-0.47, 2.83] \times 10^{-4}$. Both intervals span zero. Finite-grid variation therefore consumes nearly all the layer-8 joint-limit advantage. Full prediction-error reductions after joint averaging are 92.3% for the public model and 78.7% for seed 12001, distinct from their coefficient-error fractions.

Joint same-index crossings at layers 10 and 11 remain positive. The stronger layer-9/$k = 10$ comparison survives the public 16×16 check but is unresolved in seed 12001. Disjoint-half probes agree at approximately 10^{-6} normalized squared disagreement, whereas the unchanged native lens fails at layers 8–9. In the smaller seed-12001 readout controls, the layer-8 intervals also span zero. These control results use their own evaluation samples and uncertainty families.

D.3 Production oracle geometry

The production oracle is a Euclidean projection onto the intended solution space. Scaled, twice-reorthogonalized float64 constructions are checked against an independent construction using $C^{1/2}\mathcal{K}_k(C^{1/2}SC^{1/2}, C^{1/2}r)$ followed by Euclidean reorthogonalization. Rank, orthogonality, projector agreement, and saved-oracle consistency are checked. Squared residuals are evaluated directly rather than subtracting nearly equal projected and total energies.

A limited supplemental check regenerates spectrum-verified episodes and rebuilds every budget and all five preconditioned spaces with independent Chebyshev-column Householder QR at 80 and 120 decimal digits. Fourteen public-model prompt/context pairs and three seed-12001 stress cases were checked. Maximum absolute saved-oracle discrepancies were 5.77×10^{-13} and 7.45×10^{-13}, respectively; the largest positive overstatements were 5.77×10^{-13} and 3.61×10^{-13}. These are small on the checked cases, not bounds for the unselected sample.

The complete production boundary gate is not included in the reported validation. That gate combines fixed-random, large-condition-number, high-oracle-error, near-tie, and rank-loss selections at every context; it compares direct-CS Arnoldi and Chebyshev-column Householder QR at separate arbitrary precisions and escalates on disagreement. An orthogonal but incomplete basis can overstate an oracle floor, so span and rank checks are necessary. No failing prompt is silently discarded or replaced in the statistical sample. Synthetic validation in Appendix F is separate from production geometry verification.

E COMPLETE REPLICATION TABLES AND ADDITIONAL PLOTS

The following tables report the completed native and crossed evaluations. Native sweeps contain 1,024 independent episodes per model; primary $n = 40$ crossed audits contain 256 prompts per model; shorter-context crossed audits contain 128. Different model streams are not paired with each other. The exploratory fixed-polynomial analysis in Appendix C uses a separate 96-prompt sample.

E.1 NATIVE CONTEXTS AND PROMPT WINS

Table 3: **Native context sweeps.** Each slash gives identity/menu. Exclusion is the largest budget with positive simultaneous lower mean-advantage bound; majority is the largest budget with observed win fraction above $1/2$. The last column is the observed identity win rate at $k = 12$, not a confidence-certified majority.

Model	n	Model RMS	Exclusion	Majority	$k = 12$ wins (%)
Public	20	0.15103	18 / 15	17 / 14	84.0
Public	21	0.08742	18 / 16	16 / 14	87.5
Public	22	0.06321	17 / 15	16 / 13	84.3
Public	24	0.04242	15 / 13	14 / 12	77.0
Public	25	0.03798	14 / 12	13 / 11	68.3
Public	28	0.03100	12 / 10	11 / 9	38.7
Public	30	0.02890	11 / 9	10 / 8	21.9
Public	35	0.02568	9 / 8	8 / 7	1.8
Public	40	0.02402	8 / 7	7 / 6	0.0
12001	20	0.14485	18 / 15	18 / 15	87.1
12001	21	0.09128	18 / 15	17 / 15	89.1
12001	22	0.06605	17 / 15	16 / 14	91.5
12001	24	0.04061	15 / 13	14 / 12	80.9
12001	25	0.03767	14 / 13	13 / 11	73.2
12001	28	0.03154	12 / 10	11 / 9	38.4
12001	30	0.02909	11 / 9	10 / 8	19.2
12001	35	0.02557	9 / 8	8 / 7	2.0
12001	40	0.02373	8 / 7	7 / 6	0.0

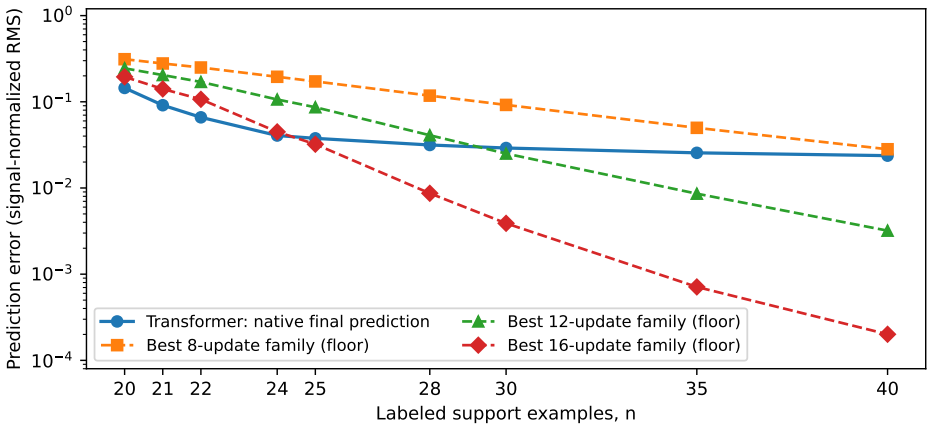

Figure 3: **Risk-versus-context replication in seed 12001.** Same construction as Figure 1, using the seed's 1,024 evaluation episodes. Each oracle has arbitrary prompt-dependent coefficients. Root-mean-square values are formed after averaging squared errors.

E.2 INDEPENDENT TRAINING: ALL 14 ENDPOINTS

Table 4: **Complete endpoint report for the independent replication.** The public checkpoint is separate from the registered cohort. Units apply to the displayed means and intervals. Each model's column gives its mean and approximate simultaneous 95% within-model prompt interval. All 14 endpoints are retained. Symmetry fractions use orbit-grid mean original coefficient error; affine fractions use separately measured unrotated-prompt coefficient error (Table 5).

Endpoint	Units	Public [interval]	12001 [interval]
Native $n = 25, k = 12$, identity	10^{-3}	$5.203\,[3.742, 6.664]$	$6.057\,[4.455, 7.658]$
Native $n = 22, k = 12$, identity	10^{-3}	$25.294\,[20.022, 30.565]$	$24.555\,[20.126, 28.985]$
Native $n = 25, k = 14$, identity	10^{-3}	$1.291\,[0.501, 2.080]$	$1.744\,[0.807, 2.680]$
Native $n = 22, k = 16$, identity	10^{-3}	$6.728\,[4.364, 9.092]$	$7.068\,[4.604, 9.533]$
Native $n = 25, k = 12$, menu	10^{-3}	$1.629\,[0.838, 2.421]$	$2.059\,[0.997, 3.121]$
Native $n = 22, k = 12$, menu	10^{-3}	$13.456\,[9.781, 17.132]$	$12.769\,[9.342, 16.196]$
Symmetry-removable coefficient error	10^{-4}	$4.390\,[4.038, 4.742]$	$3.531\,[3.229, 3.834]$
Layer 8 / $k = 8$, joint advantage	10^{-4}	$7.273\,[4.618, 9.928]$	$3.618\,[1.261, 5.974]$
Layer 8 / $k = 8$, finite advantage	10^{-4}	$4.848\,[2.319, 7.378]$	$0.022\,[-2.465, 2.508]$
Layer 9 / $k = 9$, joint advantage	10^{-4}	$2.969\,[1.750, 4.188]$	$2.647\,[0.978, 4.316]$
Layer 9 / $k = 9$, finite advantage	10^{-4}	$2.056\,[0.853, 3.259]$	$1.184\,[-0.467, 2.835]$
Affine-explained coefficient error	10^{-4}	$1.456\,[1.116, 1.796]$	$1.285\,[1.013, 1.558]$
Symmetry-dependent coefficient fraction	%	$96.012\,[95.125, 96.900]$	$97.124\,[96.190, 98.059]$
Affine-explained coefficient fraction	%	$31.397\,[26.126, 36.668]$	$34.762\,[29.690, 39.834]$

Seed 12001 completes the native and crossed evaluations, readout controls, precision subset, and short-context extensions. All six native lower bounds are positive. Joint-limit intermediate advantages replicate; layer-8 finite advantage is essentially zero and layer-9 finite advantage has an interval spanning zero. The stronger layer-9/$k = 10$ joint comparison is unresolved for this seed, including in its precision subset. This table reports completed measurements without imputing outcomes for seeds 12002–12005.

The two reported fractions use different baseline estimators. For the symmetry fraction, the denominator D_{orbit} is mean original coefficient error over the Haar grid. For the frozen-affine fraction, D_{original} is mean coefficient error on separately evaluated, unrotated prompts. Both estimate $d^{-1}\mathbb{E}\|v(D) - w\|^2$ under task-distribution invariance, but they differ in finite samples. If E_{joint} is joint-average coefficient error and G_{affine} is the frozen-reference gain, the reported fractions are

$$\rho_{\mathrm{symmetry}} = \frac{D_{\mathrm{orbit}} - E_{\mathrm{joint}}}{D_{\mathrm{orbit}}}, \qquad \rho_{\mathrm{affine}} = \frac{G_{\mathrm{affine}}}{D_{\mathrm{original}}}.$$

Table 5: **Baseline estimators for the final coefficient-error fractions at $n = 40$.** All entries are normalized squared-error quantities, $\times 10^{-4}$. The symmetry numerator is the original-minus-joint grid estimate; the affine numerator is measured on original prompts. The fractions in Table 4 retain these denominators rather than being renormalized to a common finite-sample estimate.

Model	D_{orbit}	D_{original}	$D_{\mathrm{orbit}} - E_{\mathrm{joint}}$	G_{affine}
Public	4.572306	4.637947	4.389983	1.456179
12001	3.635786	3.697093	3.531226	1.285175

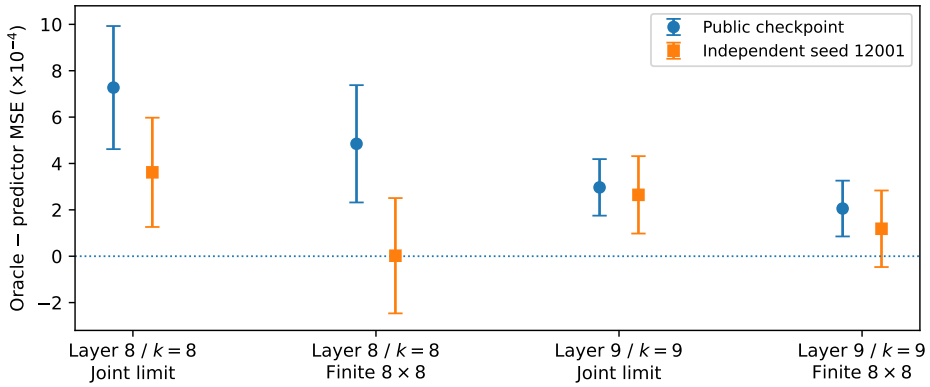

Figure 4: **Joint-limit and finite-ensemble results are distinct.** Oracle-minus-predictor MSE on 256 prompts per model; bars are approximate simultaneous 95% prompt intervals. Both joint-limit crossings replicate. Seed 12001's layer-8 finite advantage is essentially zero and its layer-9 finite interval also spans zero. This secondary result is weaker than the native replication in Table 1.

E.3 Component shares from the 256-prompt primary audits

Table 6: **Final coefficient-error decomposition, 256 prompts per model at $n = 40$.** Each pair of columns uses its own denominator: total coefficient error or coefficient error outside $\mathcal{K}_{12}$. Shares are ratios of mean energies, not means of promptwise ratios. Tiny jointly invariant outside errors are uncertain estimates, not a claim of exact equality with the target.

	Coefficient-error share (%)		Outside-$\mathcal{K}_{12}$ error share (%)	
Component	Public	12001	Public	12001
Jointly invariant	3.988	2.876	0.057	0.397
Row-only	12.222	8.362	13.322	7.983
Coordinate-only	37.344	41.822	36.751	42.667
Interaction	46.447	46.940	49.870	48.952

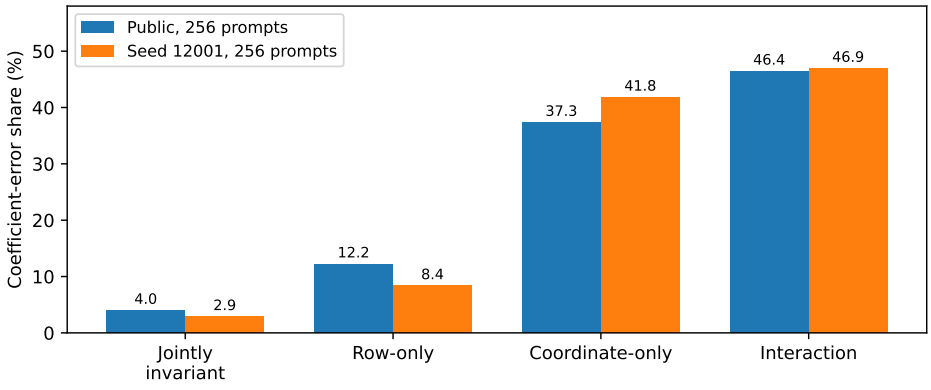

Figure 5: **The interaction remains the largest component.** Each model has its own total coefficient-error denominator and independent 256-prompt evaluation sample. Row and feature marginal sensitivities overlap; their sum would double-count the interaction.

The public symmetry-dependent fraction is 0.960125 $[0.951253, 0.968996]$ and the seed fraction 0.971242 $[0.961896, 0.980587]$. The affine explanatory fractions are 0.313971 $[0.261264, 0.366677]$ and 0.347618 $[0.296896, 0.398340]$, respectively. All four intervals belong to the primary within-model endpoint analysis.

E.4 SIGNED TAIL SENSITIVITY WITH THRESHOLD RESELECTION

For each comparison, remove the largest $\lceil \alpha N \rceil$ oracle errors with $N = 1024$. At each of 6,000 bootstrap resamples, resample paired episodes and recompute the ranking and cutoff. The intervals are pointwise percentile intervals for the trimmed statistic, not the native simultaneous family and not intervals conditional on a fixed retained subset. The analysis examined 0%, 1%, 5%, and 10% removal; Table 7 shows the 10% comparison.

Table 7: **Tail sensitivity in both models.** Normalized mean advantages and exploratory 95% intervals, $\times 10^{-3}$. The moderate $n = 22, k = 12$ identity and menu advantages persist. Frontier means are more tail-dependent, and several become unresolved or negative.

Model	(n, k)	Class	Full	Retained 90%	Retained interval
Public	$(22, 12)$	Identity	25.294	14.069	$[12.616, 15.577]$
Public	$(25, 12)$	Identity	5.203	2.217	$[1.862, 2.607]$
Public	$(22, 12)$	Menu	13.456	5.324	$[4.438, 6.354]$
Public	$(25, 12)$	Menu	1.629	-0.053	$[-0.224, 0.140]$
Public	$(22, 17)$	Identity	3.206	-0.429	$[-0.866, 0.089]$
Public	$(21, 18)$	Identity	3.255	-0.760	$[-1.466, -0.033]$
Public	$(21, 16)$	Menu	2.786	-1.385	$[-2.146, -0.556]$
Public	$(25, 14)$	Identity	1.291	-0.219	$[-0.367, -0.057]$
Public	$(25, 13)$	Menu	0.438	-0.675	$[-0.788, -0.543]$
Public	$(22, 16)$	Identity	6.728	1.462	$[0.808, 2.237]$
12001	$(22, 12)$	Identity	24.555	14.712	$[13.279, 16.310]$
12001	$(25, 12)$	Identity	6.057	2.726	$[2.352, 3.143]$
12001	$(22, 12)$	Menu	12.769	5.281	$[4.316, 6.323]$
12001	$(25, 12)$	Menu	2.059	0.077	$[-0.115, 0.283]$
12001	$(22, 17)$	Identity	3.738	-0.041	$[-0.541, 0.448]$
12001	$(21, 18)$	Identity	3.014	-0.684	$[-1.357, 0.036]$
12001	$(21, 16)$	Menu	1.909	-1.565	$[-2.385, -0.714]$
12001	$(25, 14)$	Identity	1.744	0.040	$[-0.119, 0.225]$
12001	$(25, 13)$	Menu	0.759	-0.546	$[-0.666, -0.428]$
12001	$(22, 16)$	Identity	7.068	1.841	$[1.217, 2.543]$

The population decomposition (16) concerns signed advantage, not only oracle-error mass. A substantial retained advantage can coexist with a large contribution from the removed tail. The hardest 10% amplifies the robust $k = 12$ effect; at the highest budgets it can sustain the full positive mean. No trimmed result changes the original evaluation sample or supplies a new maximum-budget primary claim.

E.5 NATIVE FINAL-HEAD CROSSED AUDITS AT SHORTER CONTEXTS

Table 8: **Short-context crossed results, 128 prompts per model/context.** Full prediction risks and joint-limit intervals are in $\times 10^{-3}$ units. Symmetry fractions are coefficient-error point estimates (%), not full-risk fractions. Negative lower bounds reflect signed-estimator uncertainty and are not clipped. Original risk averages unaveraged orbit-cell predictions.

Model	n	Original risk	Joint risk	Joint-risk interval	Coeff. symmetry (%)
Public	22	4.0665	0.5071	$[-0.5104, 1.5246]$	86.82
Public	25	1.3778	0.0379	$[0.0178, 0.0580]$	99.10
12001	22	3.7365	0.6530	$[-0.2930, 1.5990]$	82.09
12001	25	1.4287	0.1769	$[0.1409, 0.2130]$	97.48

These runs use the same final native head as the primary context sweep but different, smaller evaluation samples. At $n = 22$, full-error reduction is approximately 87.5% for the public model and 82.5% for seed 12001; the estimated joint residual is imprecise in both. At $n = 25$, the corresponding reductions are 97.2% and 87.6%. The short-context native oracle crossings are already established without averaging. These additional runs locate symmetry-dependent error where those crossings occur rather than exporting the $n = 40$ percentages to different contexts.

E.6 ORACLE FLOORS AND THE RESOLUTION OF INTERMEDIATE TESTS

Table 9: **Matched $n = 40$ floor calibration, 256 prompts per model.** All values are normalized MSE $\times 10^{-4}$. Probe-risk columns repeat each model's joint-limit risk to make the changing oracle floor visible. They are means, not additional significance tests. As the floor drops below residual error, this risk certificate stops excluding the class; it does not infer an upper bound on represented degree.

Model	k	Adaptive oracle	Layer-8 joint risk	Layer-9 joint risk
Public	8	8.72814	1.45504	0.19837
Public	9	3.16746	1.45504	0.19837
Public	10	1.15625	1.45504	0.19837
Public	12	0.11734	1.45504	0.19837
Public	16	0.00033	1.45504	0.19837
12001	8	8.85859	5.24087	0.93289
12001	9	3.58005	5.24087	0.93289
12001	10	1.29812	5.24087	0.93289
12001	12	0.12099	5.24087	0.93289
12001	16	0.00042	5.24087	0.93289

The original layer-8 and layer-9 tests are prespecified intermediate contrasts, not the largest degree those representations could express. Same-index joint-limit tests at layers 10 and 11 are also positive in seed 12001. The larger-grid and readout-control families have different uncertainty from the primary 256-prompt family; unresolved control results are retained rather than counted as confirmations from positive point estimates alone.

Table 10: **Outside-$\mathcal{K}_{12}$ prediction RMS, 256 prompts per model at $n = 40$.** RMS is the square root of mean signal-normalized prediction-component energy. Row, feature, and joint columns estimate the corresponding infinite averages. The target column is the teacher's own outside component. Values are rounded from the saved measurements.

Model	Original	Row average	Feature average	Joint average	Target
Public	0.01384	0.00882	0.00597	0.00341	0.00343
12001	0.01258	0.00868	0.00500	0.00366	0.00348

Quantitative approximate-explanation bounds. Table 11 propagates the native simultaneous bands using Appendix B.9. Its lower bounds apply to the complete prediction function, including any offset or query nonlinearity. The plug-in column is a descriptive evaluation of Corollary 2 at sample means. The last column uses an upper risk bound and lower advantage bound from the same confidence family, and is rounded downward to avoid overstating the lower bound.

Table 11: **Prediction-distance lower bounds at $n = 22, k = 12$.** $\overline{R}$ and $\underline{\Delta}$ are in normalized MSE units $\times 10^{-3}$; the last two columns are signal-normalized prediction RMS. Each model's last column inherits its native approximate simultaneous 95% event. No additional bootstrap or across-model pooling is used.

Model	Class	$\overline{R}$	$\underline{\Delta}$	Plug-in	Simultaneous lower
Public	Identity	4.7472	20.0220	0.1079	0.0884
Public	Menu	4.7472	9.7811	0.0689	0.0516
12001	Identity	5.8457	20.1257	0.1040	0.0846
12001	Menu	5.8457	9.3421	0.0648	0.0467

Usefulness connects native exclusion and symmetry averaging. Table 12 reports $U_{n,12}$ for the unchanged final head. Individual-prediction usefulness is positive at $n = 22, 25$ and negative at $n = 40$; joint-average usefulness is positive throughout. Every displayed sign is resolved within its own audit's simultaneous family. A positive value means that removing the outside linear component raises target MSE; negative values mean removing it helps. This tests target alignment directly, whereas matching the magnitude of the target's outside component alone would be insufficient.

Table 12: **Final-head outside-$\mathcal{K}_{12}$ usefulness, $\times 10^{-4}$.** Means and approximate simultaneous 95% intervals. "Individual" averages unaveraged predictions over the Haar grid, rather than a finite ensemble or the separate 1,024-episode native sample. Joint columns estimate infinite joint averages. There are 128 base prompts per model at each of $n = 22, 25$ and 256 at $n = 40$. Coverage is within each model/context audit; the table does not create a joint confidence family across runs.

Model	n	Individual $U_{n,12}(v)$ [interval]	Joint $U_{n,12}(ABv)$ [interval]
Public	22	$251.985\,[109.335, 394.634]$	$276.207\,[131.830, 420.583]$
Public	25	$47.175\,[19.169, 75.182]$	$53.429\,[24.887, 81.970]$
Public	40	$-1.681\,[-1.824, -1.539]$	$0.116\,[0.053, 0.179]$
12001	22	$225.976\,[133.543, 318.409]$	$246.728\,[151.909, 341.547]$
12001	25	$59.458\,[5.426, 113.491]$	$64.990\,[10.206, 119.773]$
12001	40	$-1.333\,[-1.469, -1.197]$	$0.115\,[0.040, 0.190]$

F FINITE-SAMPLE SYNTHETIC CALIBRATION

This experiment tests the risk-exclusion calculation on predictors whose class membership and conditional risk are known. It does not execute a neural model. We generate 200 independent repetitions; each contains 256 Gaussian regression episodes with $d = 20$, nested contexts $n \in \{22, 40\}$, and budgets $k \in \{8, 12, 16\}$. Every episode has 512 Gaussian holdout queries shared across its contexts and controls. The teacher prior, error normalization, and pairing match the model experiment.

F.1 PREDICTORS WITH KNOWN ANSWERS

Let $p = P_k w$, $e = (I - P_k)w$, and $u = r/\|r\|$. We use the following three query-linear coefficient maps:

$$\begin{aligned} z_{\text{projection}} &= p, \\ z_{\text{CG}} &= \text{CG}_k(S, r) + 0.02\sqrt{d}\,\eta u, \\ z_{\text{alternative}} &= p + \tfrac{1}{2}e, \end{aligned}$$

where $\eta \sim N(0, 1)$ is drawn once per support/budget, independently of the query bank. The perturbation is fixed across queries, so the second control stays in $\mathcal{K}_k$; adding fresh prediction noise would not be an in-class deterministic functional control. The first control is the least-favorable boundary null: its exact conditional risk equals the oracle. Orthogonal projection implies that the second control's conditional risk is at least that oracle, whatever η. The third has conditional risk $\|e\|^2/(4d)$ and advantage $3\|e\|^2/(4d)$, strictly positive almost surely at the tested unsaturated budgets.

The controls use known w to establish their exact answers; they are diagnostic functions, not proposed deployable predictors. The largest numerically measured outside energy of the perturbed CG coefficient over this experiment was 1.55×10^{-26} in signal-normalized units.

F.2 BANDS AND REPEATED-TEST RESULTS

For each control separately, the bootstrap family contains the six context/budget model risks and six oracle-minus-model advantages. We resample 256 complete base episodes 4,000 times and use the centered maximum absolute bootstrap statistic divided by the original estimated standard error, the same plug-in max-$|t|$ construction used in the native audit. An exclusion occurs when an advantage's simultaneous lower bound is positive. We record whether any of the six comparisons is excluded and whether all six are excluded. Each of the 200 repetitions regenerates supports, teachers, perturbations, queries, and bootstrap draws.

Table 13: Executed synthetic calibration. Counts are over 200 independent simulated experiments, each with six class comparisons. The null false-exclusion rate is the probability of at least one positive lower bound within that control's family. Intervals are exact two-sided 95% binomial intervals on the observed repetition counts.

Control	Any exclusion	All six	Any-exclusion rate [95% CI]
Target projection (null)	0/200	0/200	0 $[0, 0.0183]$
Perturbed CG (null)	0/200	0/200	0 $[0, 0.0183]$
Half-residual recovery	200/200	179/200	1 $[0.9817, 1]$

The alternative is excluded in all 200 repetitions at every $n = 22$ comparison and at $n = 40, k = 8$. Exclusion counts are 199/200 at $n = 40, k = 12$ and 180/200 at $n = 40, k = 16$. The mean advantage at that last cell is only 2.41×10^{-8}, so a known useful outside component can be difficult to detect even when the predictor removes three quarters of the oracle error. No in-class control produces a positive lower bound in this experiment.

This reduced-family check is not a coverage theorem for the full production audit. The synthetic calibration used random seed 738261. Independent spectral-QR and high-precision tests validate the control geometry on separate test matrices.

G EXACT GAUSSIAN ADAPTIVE FLOORS AND THE EXCLUSION WINDOW

This section proves Proposition 3 by applying the finite-dimensional MINRES residual law of Paquette & Trogdon (2023, Theorem 1.2(b)). Their theorem treats a Wishart matrix and a fixed right-hand side; the reduction below explains why it applies to our prompt-adaptive *target oracle*. It is not the fixed-polynomial optimization over an expected spectral measure in Appendix C.

G.1 THE ORACLE IS A MINRES RESIDUAL ON AN AUXILIARY SYSTEM

Let $u = w/\|w\|$. Since $r = Sw$ and multiplication by $\|w\|$ does not change the subspace,

$$Z_{n,d,k} := \frac{\|(I - P_k(S, Sw))w\|^2}{\|w\|^2} = \min_{z \in \mathcal{K}_k(S,Su)} \|u - z\|^2 \tag{31}$$

$$= \min_{t \in \mathcal{K}_k(S,u)} \|u - St\|^2 = \min_{\substack{\deg p \le k \\ p(0)=1}} \|p(S)u\|^2. \tag{32}$$

The middle equality uses $\mathcal{K}_k(S, Su) = S\mathcal{K}_k(S, u)$. The last minimization is the squared residual of MINRES for the auxiliary system $St = u$, started at zero. This is an identity for an oracle, not a solver supplied with the unknown teacher. In particular, the independent right-hand side in the theorem is u, *not* the correlated regression statistic $r = Sw$.

Gaussian w has independent radius and uniform direction, both independent of S. Rotational invariance of the Wishart matrix makes the conditional law of (32) the same for every fixed unit u. Thus Z is independent of $\|w\|$ and

$$F^2_{n,d,k} = \frac{1}{d}\mathbb{E}\|(I - P_k)w\|^2 = \mathbb{E}Z_{n,d,k}.$$

For the empirical normalization, the actual per-prompt oracle error is distributed as HZ with independent $H \sim \chi^2_d/d$. The radius is essential for the variance of the recorded error even though it cancels from its expectation.

G.2 THE EXACT CHI-SQUARE REPRESENTATION

In the notation of Paquette & Trogdon (2023, Theorem 1.2(b)), set the matrix dimension $N = d$, sample size $M = n$, and real-ensemble parameter $\beta = 1$. Their independent variables satisfy $\alpha_j \sim \chi_{n-j}$ and $\beta_j \sim \chi_{d-j-1}$. Squaring the displayed residual norm and putting $i = j + 1$ gives

$$Z_{n,d,k} \overset{\text{law}}{=} \left(1 + \sum_{j=1}^{k} \prod_{i=1}^{j} \frac{A_i}{B_i}\right)^{-1}, \quad A_i \sim \chi^2_{n-i+1}, \quad B_i \sim \chi^2_{d-i}, \quad 0 \le k < d, \tag{33}$$

with all A_i, B_i independent. The sum is empty at $k = 0$, giving $Z = 1$. At $k = d$, the residual is zero almost surely. The theorem is a finite-dimensional identity, not an asymptotic universality approximation.

For completeness, its orthogonal-polynomial derivation makes the orientation of the ratios explicit. The realized spectral measure from u is $\mu_{S,u} = \sum_i |\langle u, e_i\rangle|^2 \delta_{s_i}$. If $p_0, \dots, p_k$ are its orthonormal polynomials, expansion of p and Cauchy–Schwarz in (32) give $Z = (\sum_{j=0}^{k} p_j(0)^2)^{-1}$. Gaussian bidiagonalization preserving the starting vector produces $T = LL^\top/n$, with diagonal entries of L equal to $\sqrt{A_i}$ and subdiagonal entries $\sqrt{B_i}$; this is the real Gaussian specialization of the Laguerre bidiagonal model, with $m = d$, $a = n/2$, and $\beta = 1$ (Dumitriu & Edelman, 2002, §§ 3.1–3.2; Paquette & Trogdon, 2023). The leading $j \times j$ determinant of T at zero has magnitude $n^{-j}\prod_{i=1}^{j} A_i$, whereas the product of its first j off-diagonal entries is $n^{-j}\prod_{i=1}^{j}\sqrt{A_i B_i}$. Their squared ratio is

$$p_j(0)^2 = \prod_{i=1}^{j} \frac{A_i}{B_i},$$

proving (33). As an elementary check, $Z_{n,d,1} \sim \mathrm{Beta}((d-1)/2, n/2)$ and $F^2_{n,d,1} = (d-1)/(n + d - 1)$.

G.3 A COUPLING PROVES STOCHASTIC MONOTONICITY

Fix d, k, keep $B_1, \ldots, B_k$ fixed, and couple successive sample sizes by

$$A_i(n+1) = A_i(n) + G_{i,n}^2, \qquad G_{i,n} \overset{\text{iid}}{\sim} N(0,1),$$

independently of all initial variables. The chi-square degrees of freedom are correct at every n, and every A_i/B_i increases almost surely. The denominator in (33) therefore strictly increases for $k \geq 1$. Thus $Z_{n+1,d,k} < Z_{n,d,k}$ in this coupling, proving stochastic decrease and strict decrease of the mean. Coupling the same independent radius H proves the analogous distributional ordering for signal-normalized prompt errors HZ.

This is not a statement that extending a specific realized design always improves its fixed-budget oracle. For example, $S = I$ makes $Sw = w$, so the one-dimensional oracle is exact. Adding a row x generally changes that starting direction to a multiple of $nw + x(x^\top w)$, which need not be parallel to w. The Gaussian distribution, not a nested-space argument across prefixes, supplies the monotonicity.

G.4 AN EXPLICIT BOUND AND THE FIXED-DIMENSION RATE

Retaining only the last term in the denominator of (33) yields

$$Z_{n,d,k} \leq \prod_{i=1}^{k} \frac{B_i}{A_i}.$$

For $n > k+1$, independence, $\mathbb{E}B_i = d - i$, and $\mathbb{E}A_i^{-1} = (n-i-1)^{-1}$ give the upper bound in (10). The separate bound $Z \leq 1$ supplies the minimum with one.

Under the preceding coupling $A_i(n)/n \to 1$ almost surely. For fixed d, k, the highest-degree product dominates the denominator, giving

$$n^k Z_{n,d,k} \longrightarrow \prod_{i=1}^{k} B_i \quad \text{almost surely.}$$

Expectation convergence follows from uniform integrability. Indeed,

$$n^k Z_{n,d,k} \leq \prod_{i=1}^{k} \frac{nB_i}{A_i}, \qquad \mathbb{E}\left(\prod_{i=1}^{k} \frac{nB_i}{A_i}\right)^2 = \prod_{i=1}^{k} \frac{n^2(d-i)(d-i+2)}{(n-i-1)(n-i-3)},$$

which is bounded for all sufficiently large n. Hence $n^k F_{n,d,k}^2 \to \prod_{i=1}^{k}(d-i)$. The rate is n^{-k} for squared error and $n^{-k/2}$ for RMS; it is not a claim uniform in growing d or k.

Corollary 4 follows by dividing the model risk by this asymptotic floor. At equal exponents, $b < \prod_i (d-i)$ gives a crossing and $b > \prod_i (d-i)$ prevents it; equal leading constants require higher-order terms. Independently of such a model-rate assumption, $\Delta_{n,k} \leq F_{n,d,k}^2$ because risk is nonnegative. Thus longer contexts shrink the available absolute exclusion window, but they do not determine the trained model's curve, paired variance, or prompt-win frequencies. Identity is in the preconditioner menu, so its floor upper-bounds the menu floor; the chi-square law itself is not asserted for the prompt-dependent preconditioners.

At fixed k and proportional dimensions $d/n \to \gamma \in (0, 1]$, the ratios instead converge to $1/\gamma$, recovering $(\sum_{j=0}^{k} \gamma^{-j})^{-1}$ by bounded convergence. This recovers the leading-order MINRES limit of Paquette & Trogdon (2023, Theorem 1.3(b)). This fixed-k limit should not replace the exact formula at budgets close to $d = 20$.

G.5 INDEPENDENT NUMERICAL CHECK AGAINST THE COMPLETE CONTEXT GRID

We evaluated (33) with 10^6 scalar draws per context, independently of any matrix projection or model inference. We compared all 9×19 nonsaturated identity floors for each of the public and seed-12001 samples: 342 mean comparisons. Table 14 displays three representative budgets. The chi-square calculation reproduces the context/budget dependence of the measured floors with sampling variation, rather than exact equality of independent Monte Carlo means.

Table 14: Independent scalar-law check of the identity-class floors. Entries give law / saved public-sample RMS; $z_{\max}$ is the largest absolute mean discrepancy in combined standard-error units across both models and all $k = 1, \ldots, 19$ at that context. Law estimates use 10^6 draws; saved samples have 1,024 episodes each. $z_{\max}$ uses the saved-sample variance; the law-variance comparison is reported below.

n	$k = 8$	$k = 12$	$k = 16$	$z_{\max}$
20	0.31067 / 0.31648	0.24526 / 0.25101	0.19461 / 0.19708	1.53
21	0.27630 / 0.28060	0.20109 / 0.20775	0.14071 / 0.14243	1.71
22	0.24497 / 0.25133	0.16288 / 0.17114	0.09831 / 0.10356	2.15
24	0.19094 / 0.19597	0.10361 / 0.10714	0.04466 / 0.04736	1.86
25	0.16818 / 0.17079	0.08172 / 0.08152	0.02935 / 0.03078	1.96
28	0.11487 / 0.11674	0.04004 / 0.04087	0.00860 / 0.00939	1.71
30	0.08918 / 0.09030	0.02510 / 0.02681	0.00398 / 0.00395	4.25
35	0.04938 / 0.04895	0.00864 / 0.00856	0.00073 / 0.00074	1.78
40	0.02879 / 0.02907	0.00339 / 0.00335	0.00018 / 0.00017	1.89

The largest standardized discrepancy is 3.01 when using the law-derived per-prompt variance $\mathrm{Var}(HZ) = (1 + 2/d)\mathbb{E}Z^2 - (\mathbb{E}Z)^2$ plus scalar-integration variance. All 342 are below the descriptive Bonferroni-normal reference threshold 3.80. This is a distributional consistency check, not a coverage theorem or per-prompt numerical certificate. Using only the 1,024-prompt sample variance instead gives a maximum discrepancy of 4.25 at seed-12001 $n = 30, k = 19$, whose sample variance is much smaller than the law-derived variance. A disclosed follow-up simulating 4,096 independent size-1,024 samples from the exact law places that observed mean at the 8.6th percentile. No floor or confidence interval in the primary experiment is replaced.

H ADDITIONAL RELATED WORK AND INTERPRETATION SCOPE

Learning an estimator versus realizing a computation. Bai et al. (2023) construct transformers with provable statistical performance and in-context algorithm selection and aggregation. Their post-ICL validation can return a convex combination of near-best candidate predictors (Proposition 10). In contrast, our menu selects one preconditioner per prompt and does not cover aggregation across members. Zhang et al. (2024) establish population-gradient-flow convergence for one-layer linear attention under a specified small, balanced initialization, and characterize the learned predictor under distribution shifts. Raventós et al. (2023) show empirically that increasing pretraining task diversity can move the learned predictor away from the Bayes estimator for the finite training-task prior and toward ridge regression associated with a broader Gaussian prior. These studies motivate specifying both the training distribution and the tested output class.

Looping, depth, and numerical procedures. Giannou et al. (2023) construct fixed-depth transformers whose repeated application executes programs, separating parameter depth from total sequential computation. Gatmiry et al. (2024) analyze multistep preconditioned gradient descent in a structured looped linear-transformer parameterization, including conditions under which training learns such solutions. For a fixed preconditioner, those zero-initialized iterates lie in the corresponding preconditioned Krylov space; our finite menu need not contain their learned preconditioner. Kernel-regression constructions and layerwise solver matching provide a related numerical setting (Yan et al., 2026). These are positive results for specified computational models; our oracle instead bounds attainable risk within a declared output class. Its budget does not charge arbitrary coefficient computation or impose a general arithmetic-circuit limit on attention.

Behavioral disagreement and causal interpretation. Shen et al. (2024) use discrepancies in order sensitivity to challenge specified gradient-descent equivalences for pretrained language models, providing a precedent for testing necessary behavioral consequences rather than relying only on optimizer matching. This concerns specified equivalences: sequential SGD and the prompt-dependent coefficients in our class can be order-sensitive. Our averaging theorem requires class preservation, not order invariance of each predictor. Méloux et al. (2025) find non-unique circuits, interpretations, and causal alignments in small Boolean-task networks, using enumeration within specified candidate-search spaces. The ambiguity persists under certain causal-alignment criteria. Sutter et al. (2025, Theorem 1, under Assumptions 1–5) establish a vacuity result for input-restricted distributed abstraction with unrestricted nonlinear alignment maps, under assumptions including countable inputs, input-injective hidden representations, and a network that solves the task. Its scope does not directly provide a no-go theorem for our continuous Gaussian task or fixed linear readouts. These results motivate restricting the explanatory class and the allowed alignment before interpreting a match. Our one-sided class exclusions leave viable richer explanations and do not identify a causal state variable.