TASK GEOMETRY AND REVERSIBLE SPECIALIZATION IN IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

What determines whether in-context generalization persists during pretraining? We study nonlinear Transformers on linear-regression tasks whose training priors match in task count, probabilities, mean, covariance, and task-vector norms, while differing in geometry. Across 54 independent pairs, the geometry intervention changed training-budget-averaged normalized excess risk by 0.0902 (95% CI 0.0846 to 0.0957). Both conditions improved early; their trajectories subsequently diverged. Continuation interventions from shared source weights reversed specialization, including after resetting the optimizer state. A separately prespecified four-dimensional experiment with 40 independent pairs yielded a final clean-MSE geometry contrast of 0.2529 (95% CI 0.1696 to 0.3362). A paired supervision experiment with 16 independent four-arm blocks showed that the objective moderates the geometry contrast, with an interaction of 0.02676 in normalized-risk AUC (95% CI 0.01518 to 0.03833). We also test the limits of candidate explanations: a frozen harmonic descriptor that performed well during development predicted new geometries worse than a finite-Bayes-risk baseline across 106 independent geometries. Separately, an existence result shows that equal finite Bayes-risk-reduction profiles need not identify the same Bayesian prediction function in sufficiently rich moment-matched prior families. The combined results characterize the magnitude, reversibility, and tested scope of specialization beyond covariance summaries, while distinguishing prediction of scalar risks from identification of an internal learning algorithm.

1 INTRODUCTION

What determines whether in-context generalization persists as pretraining continues? A Transformer can initially solve new tasks and subsequently specialize to its training distribution. Understanding this trajectory requires more than identifying a checkpoint at which few-shot learning is successful. It requires separating properties of the task distribution, changes induced by continued optimization, and the algorithms that might explain the resulting predictions.

Task diversity and geometry already influence in-context learning (ICL). Finite-task regression studies connect task count to departures from the finite-prior Bayesian predictor (Raventós et al., 2023); geometric coverage produces specialization–generalization transitions (Goddard et al., 2025); covariance alignment admits precise analysis in restricted attention models (Lu et al., 2025; Letey et al., 2025). Strategy-selection accounts explain transient generalization through competition between predictors (Wurgaft et al., 2025). These findings motivate a narrower question: *when task count, probabilities, mean, covariance, and norms are held fixed, how does task geometry change the retention and reversibility of generalization in a nonlinear Transformer?*

We construct two eight-task linear-regression priors that match those summaries but differ in angular arrangement. A paired study with 54 independent training pairs measures a sustained separation after early improvement. Continuation interventions from common trained weights reverse part of that separation, including after resetting optimizer state. A prespecified 16-block experiment shows that supervision objective changes the geometric contrast. A fresh 40-pair experiment transfers the direction of the final-risk contrast to full-dimensional four-dimensional task supports. A targeted replication of the closest strategy-selection study reproduces its transient-generalization trajectory; in all three paired seeds, final-query supervision attenuates the return toward memorization.

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.

Our contribution is the measured magnitude and dynamics of a controlled distributional intervention, together with tests of explanatory limits. We do not claim the first observation that higher moments or task geometry matter. In particular, a harmonic descriptor that looked promising during development transfers worse than a finite-Bayes-risk baseline on 106 new geometries. A separate response-level test finds an approximation residual for a common mixture of two Bayesian predictors. An existence proposition distinguishes finite scalar risk profiles from identification of a full Bayesian prediction function. These results constrain the explanation; they do not establish a universal task-diversity threshold or identify a unique internal learning algorithm.

2 Relation to prior work

Diversity, coverage, and alignment. Raventós et al. (2023) vary finite pretraining-task diversity and compare learned regression with discrete-prior and Gaussian-prior estimators. Goddard et al. (2025) study angular coverage of task space, its interaction with task count, and specialization. Our intervention holds task count and covariance fixed and measures paired training trajectories and continuation effects. Letey et al. (2025) derive covariance-alignment predictions in a solvable model and evaluate nonlinear Transformers; Lu et al. (2025) obtain high-dimensional asymptotics for linear attention. Our finite-dimensional experiment tests the reach of summary-based descriptions in a different nonlinear learner, not the validity of those theorems under their assumptions.

Learning dynamics and strategies. Singh et al. (2023) demonstrated that emergent ICL can disappear during continued training. Singh et al. (2025) studied cooperative and competitive interactions between learning strategies as an explanation of emergence and transience. Our study examines a different intervention: task geometry under matched task count, probabilities, mean, covariance, and norms, using paired training trajectories and continuation experiments. Nguyen & Reddy (2024) analyze competing learning kinetics in a classification setting. Lin & Lee (2024) distinguish retrieval and task learning in a probabilistic account of ICL. Wurgaft et al. (2025) model training-dependent preference for memorizing and generalizing Bayesian predictors. Their regression task family is Gaussian, whereas our primary continuous test family is uniform on a circle. We therefore separate two tests: a finite/circle dictionary audit in our geometry experiment, and a targeted replication using their public generator, architecture, all-output objective, and reported $D = 64$ condition. The latter changes only the objective in paired final-query arms. It reproduces transient generalization under the source objective and reveals substantial objective dependence; it does not reproduce the full task-count-by-diversity map or hierarchical fit.

Scope of mechanistic claims. Causal function-vector studies in language models (Todd et al., 2024) motivate intervention-based tests, but risk similarity alone does not establish a shared mechanism. We distinguish a causal effect of the *training distribution* from causal identification of an internal algorithm. Supplementary activation and language-model tests retain this distinction, including both failed interfaces and a positive fresh prompting control.

3 A moment-matched task intervention

Let $x_i, x_q \sim \mathcal{N}(0, I_2)$ independently, $y_i = x_i^\top w + \epsilon_i$, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.25$. A training window contains $m = 16$ noisy demonstrations and one query from the same w. For $0 < \theta \le \pi/4$, the training prior is uniform over

$$\mathcal{W}_\theta = \{\pm e_1, \pm e_2, \pm R_\theta e_1, \pm R_\theta e_2\}. \tag{1}$$

Both priors have eight distinct tasks, equal probabilities, zero mean, covariance $I_2/2$, and unit task norms. The primary contrast is 5° versus 45°. Each independent replicate pair receives a common random rotation, fixed throughout its runs. Evaluation tasks are uniform on the unit circle. They are almost surely outside the finite training support, but belong to the same linear function family.

For Gaussian inputs, squared functional distance between tasks equals squared Euclidean distance between their vectors. Their fourth angular harmonic has magnitude $|\mathbb{E}e^{4i\phi}| = |\cos(2\theta)|$, unaffected by the common rotation. Rotation averaging makes the direction of a single task marginally uniform in both conditions; the intervention concerns the *fixed within-run prior* reused across win-

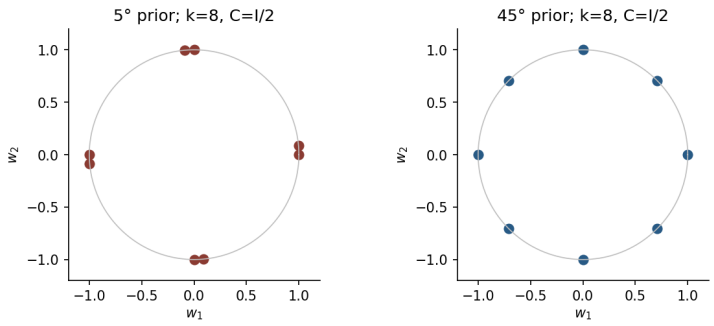

Figure 1: Training-task geometry before the common rotation. The two priors match task count, probabilities, mean, covariance, and norms. The continuous test family is the entire unit circle.

dows. Higher moments, conditional inference difficulty, and possibly gradient noise change together. We do not isolate one harmonic as the cause.

An exact restricted control. Conditional on all inputs X, both priors give $\mathbb{E}[y \mid X] = 0$ and $\mathrm{Cov}(y \mid X) = XCX^\top + \sigma^2 I$, with $C = I_2/2$. Thus every predictor linear in context labels, conditional on inputs, has the same squared population risk under the two priors when its coefficients are fixed. In particular, for $u = m^{-1}\sum_i x_i y_i$ and $\widehat{y} = x_q^\top A u$,

$$S = \mathbb{E}[uu^\top] = \frac{m+1}{m} C + \frac{\mathrm{tr}\, C + \sigma^2}{m} I, \quad A_* = CS^{-1}, \quad R_*^{\mathrm{linear}} = \frac{3 + 2\sigma^2}{m + 3 + 2\sigma^2}. \tag{2}$$

The population gradient at fixed A is invariant as well. This control does not imply identical finite-mini-batch trajectories, nor does it cover arbitrary deep linear-attention parameterizations. The nonlinear Transformer is not constrained to this class.

3.1 Experimental and statistical design

The model has four pre-LayerNorm causal Transformer blocks, hidden width 128, four heads, MLP expansion four, and 798,465 parameters. Numeric input and response tokens alternate across 33 positions. We train the last-query prediction with AdamW for 30,000 updates, batch size 256, learning rate 3×10^{-4} after 500 warm-up updates, gradient clipping at one, and no dropout. FP32 is used with TF32 disabled. Each batch is balanced over training tasks. Full settings and validation are in the appendix.

Within each pair, initialization, task indices, inputs, noise, and rotation are shared. Pairs are independent. Each pair has a trajectory bank of 512 test functions with eight contexts each, shared across conditions and checkpoints. A separate final bank has 2048 functions with eight contexts each. Evaluation targets are clean responses; demonstration noise remains. The continuous-circle posterior is the test-family oracle, distinct from Gaussian-prior ridge regression.

The primary estimand integrates normalized excess risk over the prespecified linear update grid, including initialization:

$$e_s(t) = \frac{R_s(t) - R_s^*}{1 - R_s^*}, \qquad U_s = \frac{1}{T}\int_0^T e_s(t)\, dt, \qquad D_s = U_{5^\circ,s} - U_{45^\circ,s}. \tag{3}$$

We use trapezoidal integration at 1000-update intervals. This area summarizes performance over the training budget and combines acquisition, sustained accuracy, and later deterioration; it is not a standalone measure of forgetting. The denominator uses population zero-predictor risk one. A 95% paired t interval entirely beyond ± 0.02 supports a meaningful difference; a 90% interval entirely inside this margin supports equivalence of the area estimand. The sample of 54 pairs was fixed using conservative pilot-variance-based power simulations before confirmation. Pilot runs are excluded. The independent unit is a training pair, not a context, checkpoint, or token. The protocol was committed locally before confirmation; it was not externally preregistered. All other intervals are explicitly descriptive unless stated otherwise.

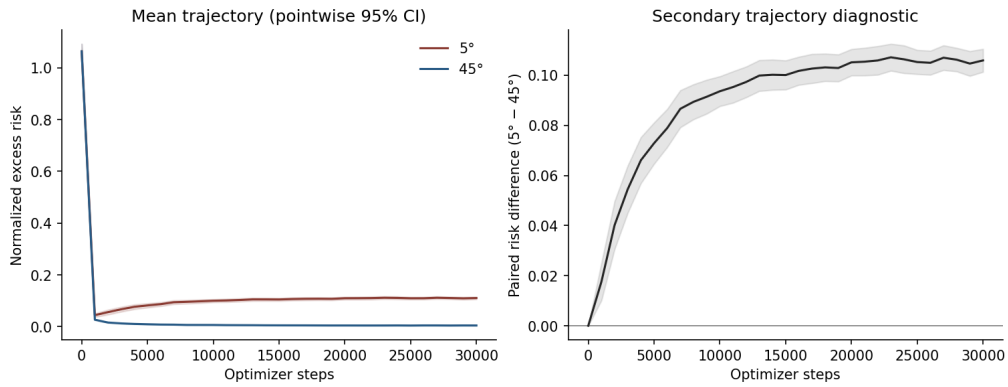

Figure 2: Primary trajectories across 54 independent pairs. Both conditions improve early, then differ in retention. Bands are pointwise 95% intervals, not simultaneous bands. The primary statistic integrates the full checkpoint grid.

4 Geometry changes retention; continuation reverses specialization

4.1 A confirmed trajectory contrast

All 108 primary runs completed. The mean paired normalized-area difference is 0.09016 (SD 0.02045; 95% CI $[0.08458, 0.09574]$), exceeding the practical margin. A paired percentile bootstrap gives $[0.08470, 0.09552]$. Mean areas are 0.11495 and 0.02479 at $5°$ and $45°$, respectively.

The difference is principally in retention after early improvement (Figure 2). The prespecified criterion of three consecutive checkpoints below 0.1 normalized excess is reached by 40/54 runs at $5°$ and 54/54 at $45°$. All recorded onsets among reaching runs are at 1000 updates; finer onset times are unresolved. Of the 40 reaching $5°$ runs, 31 subsequently exceed the threshold, versus 0/54 at $45°$. Non-reaching runs remain censored.

Independent final evaluation gives neural MSE 0.11351 versus 0.00938, a difference of 0.10414 (descriptive 95% CI $[0.09988, 0.10839]$). The common circle oracle has risk 0.00447; ridge has 0.00949; finite-prior posterior risks are 0.17336 and 0.05104. Neither network is identified with its finite-prior posterior by these risk comparisons. The closeness of the $45°$ risk to ridge likewise does not identify its algorithm. Prespecified width-256 and exposure-matched batch-1024 series retain the contrast; the latter changes update count as well as gradient noise (appendix).

4.2 An intervention from common trained weights

Six source pairs at update 30,000 are continued for 5000 updates under each of three priors: $5°$, $45°$, and continuous circle. These were H1 pair identifiers 0–5, selected as the first six independent pairs without reference to their losses, trajectories, or continuation outcomes. Each continuation is run with retained and reset AdamW states, giving 72 branches from six independent pairs. Starting weights and future input/noise streams are matched within source and optimizer treatment; evaluation uses a new common bank.

Switching $5° \to 45°$ reduces final MSE relative to staying at $5°$ by 0.09544 with retained optimizer state (descriptive 95% CI $[0.08075, 0.11013]$), and 0.09614 after reset ($[0.08413, 0.10815]$). Conversely, $45° \to 5°$ worsens risk by 0.03490 and 0.03469. Circle continuation improves the $5°$ source by 0.11687 and 0.11665.

This is a causal effect of the continuation distribution under the controlled training procedure. It supports reversible behavioral specialization: poor test performance is not an irreversible architectural incapacity. Persistence after optimizer reset makes accumulated AdamW moments alone insufficient as an explanation. Similar point estimates do not establish formal equivalence between optimizer treatments, and the intervention does not identify the internal algorithm mediating the change.

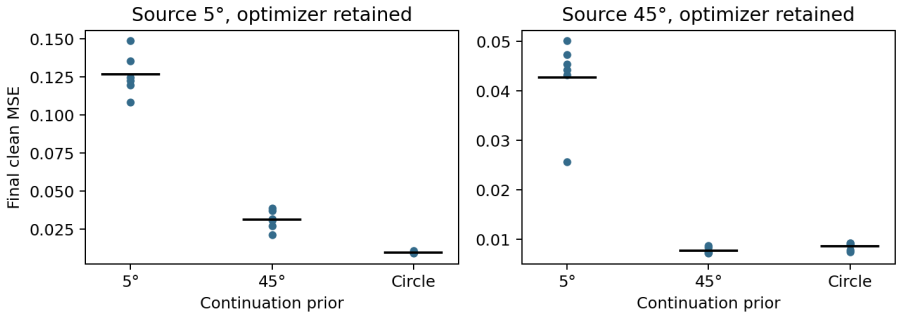

Figure 3: Continuation from common source weights with retained optimizer state. Dots are six independent source pairs; bars are means. Panels have different vertical scales. Reset-state effects are reported in the text. The 72 continuation branches are not 72 independent source replications.

4.3 Fresh scope checks

We first conducted two separately specified exploratory series at the original 30,000-update horizon: a six-pair four-dimensional extension with two additional two-dimensional control pairs, and a four-pair two-dimensional all-prefix experiment. These series comprise 24 training runs. All completed, underwent prediction reconstruction and final-weight replay, and were excluded from the 54-pair confirmation. The subsequent 40-pair four-dimensional and 16-block paired-objective experiments were separate studies.

Four-dimensional extension. An initial six-pair extension used training supports confined to a union of two orthogonal planes and was inconclusive: final-MSE contrast 0.11723 (descriptive 95% CI $[-0.10293, 0.33740]$). We then froze a separate full-dimensional design before seeing its outcomes. Each prior is uniform on the signed rows of two orthonormal bases in $\mathbb{R}^4$; the second is obtained by the fixed dense rotation $R_\theta = \exp((\theta\pi/180)A)$ specified in the appendix, with $\theta \in \{5, 45\}$ measured in degrees. Both component bases span all four dimensions. After identifying each antipodal pair $\{v, -v\}$, every triple of the resulting eight unoriented directions has rank three; the signed 16-task support itself necessarily contains dependent triples such as $(v, -v, u)$. The conditions match 16 tasks, probabilities, mean, covariance $I_4/4$, and norms.

All 40 fresh pairs completed. The prespecified final clean-MSE contrast, $\mathrm{MSE}_{5^\circ} - \mathrm{MSE}_{45^\circ}$, is 0.25289 (SD 0.26057; 95% paired-t CI $[0.16955, 0.33622]$). Its lower endpoint exceeds the frozen practical threshold 0.02. Thirty-four of 40 contrasts are positive; a deterministic pair bootstrap gives $[0.17387, 0.33361]$, and leave-one-pair-out means range from 0.23910 to 0.26960. Thus the original direction transfers beyond the union-of-two-planes support restriction. The prior remains finite rather than continuous on the sphere, and changing dimension also changes task count and the demonstration-to-dimension ratio; this is not an isolated dimension-scaling law. A post hoc cross-fitted checkpoint-selection audit finds positive mean deterioration in both four-dimensional conditions, with 33 of 40 arms showing positive deterioration in each condition and greater deterioration on average at 5°. Mean selected risks in both conditions remain substantially above the ridge reference. The result supports heterogeneous deterioration after checkpoint selection, rather than acquisition and subsequent loss of ridge-level generalization; no common transition point is identified (appendix Figure 8).

Supervision across context lengths. Four fresh two-dimensional pairs average prediction loss over contexts with 1–16 demonstrations, excluding the zero-context position. Evaluation still uses 16 demonstrations. Mean final risks are 0.12964 and 0.00606, giving contrast 0.12359 (descriptive 95% CI $[0.11403, 0.13314]$); all four contrasts are positive. The effect therefore persists under this alternative objective. There is no fresh last-query-only arm with the same seeds, so this is not a paired causal estimate of changing the objective. Both series and individual contrasts are plotted in the appendix.

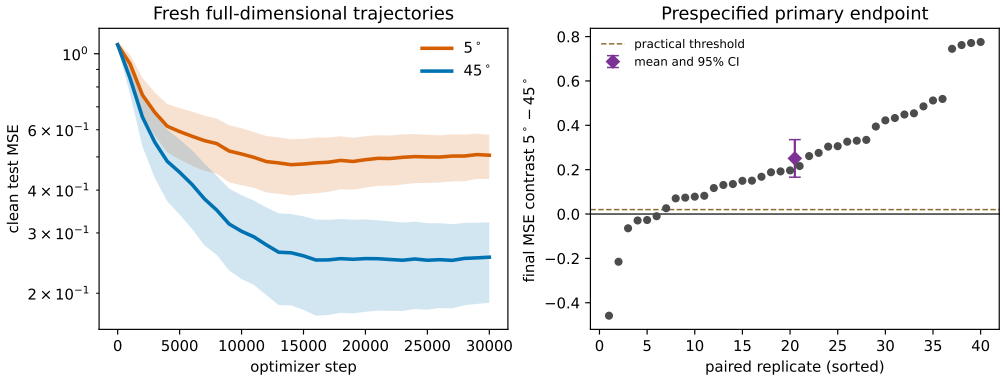

Figure 4: Fresh four-dimensional transfer. Left: mean clean-risk trajectories across 40 paired replicates; bands are pointwise 95% intervals. Right: sorted paired final contrasts and the primary paired-t interval. The dashed line is the prespecified practical threshold, not zero.

We then ran a separately frozen paired test with 16 new quartet blocks, each crossing $5°/45°$ geometry with last-query/all-prefix supervision. The pilot blocks used only to select sample size are excluded. The primary interaction, $(\mathrm{AUC}_5 - \mathrm{AUC}_{45})_{\mathrm{all}} - (\mathrm{AUC}_5 - \mathrm{AUC}_{45})_{\mathrm{last}}$, is 0.02676 (95% paired t CI $[0.01518, 0.03833]$, $p = 1.83 \times 10^{-4}$). Thus objective causally moderates the geometric contrast in this setting. The geometry contrast is positive in all 16 blocks under both objectives (0.0933 last-query, 0.1201 all-prefix). Descriptively, all-prefix minus last-query AUC is $+0.01750$ at $5°$ and -0.009257 at $45°$. The result resolves the matched-objective control but shows that the effect is not invariant to supervision.

Targeted closest-source replication. We also reproduced the $D = 64$ linear-regression condition of Wurgaft et al. (2025) with three fresh paired seeds and 100,000 updates. All three source-style all-output arms first favored the generalizing predictor and later moved toward the memorizing predictor: their prespecified transience indices were 0.190, 0.209, and 0.302. Changing only the loss to the final query reduced these to 0.004, 0.104, and 0.006. The mean paired attenuation was 0.196 (descriptive 95% t CI $[-0.042, 0.434]$). This reproduces the qualitative source trajectory in three of three seeds and supports objective dependence, but three pairs are a mechanistic control rather than a powered population claim (Appendix C.4). These indices pool all 32 positions, which is not position-matched for the final-query arm. In a post hoc last-position-only reanalysis, the paired attenuation is 0.212, 0.245, and 0.351 (mean 0.269; descriptive 95% paired-t interval $[0.089, 0.449]$). The 10%-above-observed-minimum residual criterion automatically admits a checkpoint and is not a fit-quality test. Two-fold held-out-task mixture residuals and complete curves are reported in the appendix; substantial last-position prediction-space residual remains for the all-output arms.

Pretrained-model prompting check. A separately frozen test evaluated one pretrained base model on 64 fresh numerical tasks under four paired sentence-prompt conditions. Correct demonstrations reduced capped MSE from 1.30669 without context to 0.47455; shuffled and irrelevant demonstrations gave 1.11773 and 1.67922. All three task-bootstrap control-minus-clean contrasts had positive multiplicity-adjusted one-sided lower bounds (Appendix J.7). This establishes use of correct demonstrations for one model and interface, not transfer of the synthetic training-geometry effect or task diversity in natural text.

5 Tests of Explanatory Scope

We distinguish predicting a scalar risk, reproducing a function of answers, and identifying an internal computation. Success at one level does not imply success at the others.

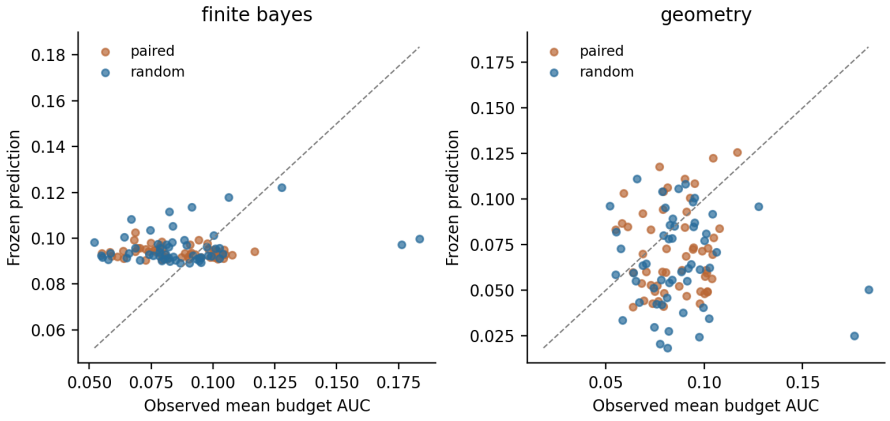

Figure 5: Forecasts frozen before 106 new task geometries. Each point averages two training seeds. The geometric predictor transfers worse than the finite-Bayes-risk baseline. Dashed lines indicate perfect prediction.

5.1 A FROZEN GEOMETRIC FORECAST FAILS TO TRANSFER

An exploratory fourth-harmonic descriptor interpolated two held-out geometries well. To test broader transfer, we developed a predictor on eight geometries, selected its specification without the new outcomes, and froze it before evaluating 106 independent geometries from two new generators, with two training seeds per geometry. Each prior contains four quarter-turn orbits: 16 equally probable unit vectors, zero mean, and covariance $I_2/2$. The geometric predictor adds fourth/eighth angular-moment magnitudes to a finite-Bayes-risk regression baseline. Training lasts 5000 updates. The target is the same form of normalized training-budget area, now over this shorter horizon.

The geometric predictor has mean absolute forecast error 0.02852, versus 0.01624 for the baseline. Its paired improvement is -0.01228 (95% CI $[-0.01650, -0.00806]$; Holm-adjusted $p = 9.14 \times 10^{-8}$ across two primary tests). We average seeds within geometry before comparing absolute errors; inference is stratified over the two fixed generators. Bootstrap and leave-one-geometry sensitivity preserve the negative sign. A prespecified width-64 check is also negative. These results reject the proposed improvement of this frozen descriptor on these generators, not the usefulness of all geometric descriptions. No target-geometry refitting was performed. The 106-geometry evaluation assesses the transfer performance of this fixed predictor; it does not quantify variability across alternative eight-geometry development samples.

A second primary test evaluates a gradient-based prediction of one SGD continuation step from the current network. Its absolute-error improvement over a zero-change forecast is 6.61×10^{-6} MSE (95% CI $[5.39, 7.83] \times 10^{-6}$; Holm-adjusted $p = 9.35 \times 10^{-18}$). Even the conservative 97.5% upper bound 8.01×10^{-6} is below the prespecified practical reference 10^{-5}. This local forecast uses current weights and gradients; it is neither a prior-only theory nor a prediction of long AdamW trajectories. Full design and the conditional Taylor remainder bound are in the appendix. Descriptively, this is a 66.3% MAE reduction; the reference slightly exceeds the entire zero-change MAE.

5.2 A TWO-PREDICTOR DICTIONARY LEAVES A RESPONSE RESIDUAL

Let M be the finite-prior posterior predictor, G the continuous-circle posterior, and h the network. We test the family $G + a(M - G)$ with one common $a \in [0, 1]$. Under a fixed context distribution, writing $v = M - G$, the best unconstrained coefficient is $a_* = \langle h - G, v \rangle / \|v\|^2$ when $\|v\| > 0$; clipping gives the best common convex mixture. This is a projection in prediction space, not a decomposition of the network's task risk.

For the four-pair two-dimensional all-prefix experiment of Appendix C.2, we fit a on 4096 independent final-bank contexts and evaluate it on another 4096. Mean held-out prediction-distance MSE is 0.04390 at 5° and 0.001591 at 45°; fitted finite-predictor weights average 0.6918 and 0.0033.

Network task risks on these held-out halves are 0.126745 and 0.005932, respectively. The response residual persists despite changing the supervision objective (0.05121 in last-query 5° H1 states; appendix). The test does not exclude context-dependent coefficients, richer dictionaries, or other Bayesian models; the targeted source replication above tests the original setting separately.

5.3 Equal scalar profiles need not identify Bayesian answers

For a unit-norm task prior P, define the finite Bayes risk-reduction profile

$$J_j(P) = \mathbb{E}_{w\sim P, X, \epsilon}\|\mathbb{E}_P[w \mid X_{1:j}, y_{1:j}]\|^2, \qquad j = 1, \ldots, m. \tag{4}$$

This quantity is not mutual information. It measures improvement over the zero predictor under the training prior. A basic existence result clarifies what such a profile can identify.

Proposition 1. *Fix $m \geq 2$ and $\sigma > 0$ in two-dimensional Gaussian-design regression. There exist two uniform priors, each on $4(m + 2)$ distinct unit vectors with mean zero and covariance $I_2/2$, not related by an orthogonal transform, with identical $J_1, \ldots, J_m$ but different posterior-mean prediction functions on a positive-probability set under the common continuous-circle test law.*

The proof constructs an open $(m + 1)$-dimensional family of noncongruent quarter-turn-symmetric priors. The continuous map to m profile values cannot be injective by invariance of domain (Brouwer, 1911). Distinct priors have distinct posterior-mean functions by the Gaussian score identity and injectivity of Gaussian convolution (Efron, 2011). Continuity extends the distinction to a positive-probability set of designs. The full argument is in the appendix.

At $m = 16$, the construction uses 72 tasks, not the sixteen-task experimental family. It establishes neither unequal scalar test risks nor unequal Transformer trajectories, and gives no practical separation bound. It therefore does not disprove approximate prediction of learning curves from information profiles. Its role is to delimit exact identification claims; the classical mathematical ingredients are not new.

6 Discussion and limitations

The main result is a controlled effect of geometry on the retention of ICL generalization, and a continuation intervention showing that the resulting specialization can be reversed in the tested learner. The area test sizes the contrast; its reading as retention rests on the prespecified threshold summary (31/40 reaching 5° runs versus 0/54 at 45° later lost it). Matching count, frequencies, mean, covariance, and norms does not match the entire distribution, and nonlinear sensitivity to the remaining structure is expected. Our evidence quantifies that sensitivity and its dynamics rather than proposing a universal covariance-sufficiency theorem to refute.

The explanatory tests impose substantive limits. The frozen harmonic descriptor fails to improve prediction on the tested new geometries. At the primary setting, the local gradient forecast reduces the zero-change forecast's MAE by 66.3%, but does not meet the prespecified absolute improvement target, which slightly exceeds the entire zero-change MAE on this sample. This result concerns one SGD step rather than long-horizon AdamW dynamics. A common two-predictor mixture leaves a prediction-space residual. These are separate findings, not a derivation of the observed training curves from first principles. The theorem concerns exact Bayesian prediction functions in a richer family, not the measured neural effect.

External validity remains restricted. The fresh four-dimensional experiment transfers the direction beyond the earlier union-of-planes support, but uses a finite task prior and changes task count and the demonstration-to-dimension ratio. The paired objective experiment finds a positive geometry contrast under both objectives, with significantly different magnitude. The closest-source replication strengthens the connection to strategy-selection dynamics, but its three paired seeds only establish a descriptive intervention in that synthetic setting. Activation tests in the appendix do not identify a reusable query-independent task vector. A fresh Qwen3-4B-Base prompting test shows causal use of correct sentence demonstrations after an earlier joint two-format gate failed, but does not reproduce the training-geometry intervention or establish natural-language task diversity. The practical value is a controlled evaluation design for distinguishing early acquisition from later retention and for testing candidate explanations without substituting scalar risk similarity for algorithmic identity.

REPRODUCIBILITY STATEMENT

The supplementary code contains frozen configurations, scientific source files, aggregate outcomes, audit records, and commands for re-running training and analysis. The appendix specifies evaluation banks, pairing, statistical units, pilot separation, and numerical controls. Returned checkpoints were verified by checksums, and scientific audits reconstructed saved risks and replayed selected or complete final states as described for each series. The lightweight supplementary package does not include all large weight and per-example arrays; its inventory explicitly distinguishes included artifacts from the complete locally retained data. Hardware and library changes may prevent bitwise reproduction.

AI USE DISCLOSURE

Generative AI tools were used throughout the research workflow, including literature search and synthesis, hypothesis refinement, experimental design, implementation of synthetic-data generators and training code, cleaning and reformatting of the public text dataset, development of mathematical claims and proofs, experiment orchestration, statistical analysis, interpretation, figure preparation, translation, and manuscript drafting and editing. The synthetic datasets were sampled by explicit numerical generators, not produced as free-form language-model outputs. Computational checks do not replace author review of scientific claims or independent assessment of proofs. Human authors are responsible for the submitted content.

REFERENCES

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009. doi: 10.1214/09-AOS689. URL https://arxiv.org/abs/0809.5032.

L. E. J. Brouwer. Beweis der invarianz des n-dimensionalen gebiets. *Mathematische Annalen*, 71: 305–313, 1911. doi: 10.1007/BF01456846.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372/.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Elisabeth Gassiat, Alice Cleynen, and Stéphane Robin. Finite state space non parametric Hidden Markov Models are in general identifiable. *arXiv preprint arXiv:1306.4657*, 2013. URL https://arxiv.org/abs/1306.4657.

Chase Goddard, Lindsay M. Smith, Vudtiwat Ngampruetikorn, and David J. Schwab. When can in-context learning generalize out of task distribution? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 19585–19599, 2025. URL https://proceedings.mlr.press/v267/goddard25a.html.

Mary I. Letey, Jacob A. Zavatone-Veth, Yue M. Lu, and Cengiz Pehlevan. Pretrain-Test Task Alignment Governs Generalization in In-Context Learning. *arXiv preprint arXiv:2509.26551*, 2025. URL https://arxiv.org/abs/2509.26551v1.

Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30135–30188, 2024. URL https://proceedings.mlr.press/v235/lin24l.html.

Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy*

of Sciences, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122. URL https://arxiv.org/abs/2405.11751v3.

Alex Nguyen and Gautam Reddy. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. *arXiv preprint arXiv:2412.00104*, 2024. URL https://arxiv.org/abs/2412.00104v2.

Qwen Team. Qwen2.5-0.5B model card, 2024a. URL https://huggingface.co/Qwen/Qwen2.5-0.5B. Revision 060db6499f32faf8b98477b0a26969ef7d8b9987.

Qwen Team. Qwen2.5-1.5B model card, 2024b. URL https://huggingface.co/Qwen/Qwen2.5-1.5B. Revision 8faed761d45a263340a0528343f099c05c9a4323.

Qwen Team. Qwen3-4B-Base model card, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Base. Revision 906bfd4b4dc7f14ee4320094d8b41684abff8539.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2e10b2c2e1aa4f8083c37dfe269873f8-Abstract-Conference.html.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58692a1701314e09cbd7a5f5f3871cc9-Abstract-Conference.html.

Aaditya K. Singh, Ted Moskovitz, Sara Dragutinović, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. Strategy coopetition explains the emergence and transience of in-context learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55720–55739, 2025. URL https://proceedings.mlr.press/v267/singh25c.html.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function Vectors in Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah D. Goodman. In-Context Learning Strategies Emerge Rationally. *arXiv preprint arXiv:2506.17859*, 2025. URL https://arxiv.org/abs/2506.17859v2.

Haoren Xu and Guanhua Fang. Self-Attention as a Covariance Readout: A Unified View of In-Context Learning and Repetition. *arXiv preprint arXiv:2605.10466*, 2026. URL https://arxiv.org/abs/2605.10466v1.

A Full experimental methods

A.1 Data and paired randomization

Each training window contains 16 noisy demonstrations and one query from a shared linear function in two dimensions. Inputs are independent standard normals and demonstration/query noise has standard deviation 0.25. The batch contains 256 windows, balanced exactly across the eight training functions. We optimize the MSE of the noisy final query response.

Within a replicate pair, initial network weights, task identities, input draws, noise draws and task rotations are shared. Across replicate pairs they are independent. Evaluation draws task directions uniformly from the unit circle, almost surely excluding the finite training support. The query target for evaluation is the clean linear response; demonstration noise remains present.

Each pair has an independent trajectory-evaluation bank of 512 task functions with eight contexts per function. Both conditions and all checkpoints within that pair use this common bank. An independent final bank contains 2048 functions with eight contexts each. Development and confirmatory seed namespaces are disjoint.

A.2 Architecture and optimization

The primary model is a four-block pre-LayerNorm causal Transformer with hidden size 128, four attention heads per block and an MLP expansion factor of four. Numeric input and label tokens alternate, followed by the query input, for a total length of 33. Learned type and positional embeddings are added to the input projection. A scalar head reads the final position. The specified architecture has 798,465 trainable parameters.

AdamW uses learning rate 3e-4, betas (0.9,0.95), epsilon 1e-8, matrix-weight decay 0.01 and no decay on biases or normalization parameters. The learning rate warms up for 500 steps and then remains constant. Gradient norm is clipped to 1; dropout is zero. Training and evaluation use FP32 with TF32 disabled. The frozen confirmation budget is 30,000 optimizer steps, with evaluation every 1000 steps including initialization.

A.3 Oracles and controls

The test-family oracle integrates the posterior over the unit-circle angle. Starting at 512 equally spaced angles, quadrature resolution is doubled through at most 16,384 points. Convergence is declared at the first doubling for which the maximum episode-wise Euclidean change in the posterior-mean task vector is below 10^{-4} and the absolute change in estimated clean risk is below 10^{-5}; otherwise evaluation fails rather than silently accepting the last grid. It differs from ridge regression with a Gaussian prior of covariance I/2; both are reported separately. A finite Bayesian predictor and finite MAP predictor use the eight actual pretraining functions. Their extrapolation behavior is diagnostic rather than the definition of successful neural generalization.

Positive controls train on a fresh task direction from the continuous family for every window. Negative controls generate the query task independently of its demonstrations. The negative control has zero available contextual Bayes gain, so its normalized excess score is undefined and only raw risk is reported.

A.4 Primary estimand and decision rules

For a replicate with model risk $R(t)$ and common oracle risk R^*, define

$$e(t) = \frac{R(t) - R^*}{1 - R^*}, \qquad U = \frac{1}{T}\int_0^T e(t)\,dt.$$

The integral uses the trapezoidal rule on the prespecified linear-budget checkpoint grid. The primary difference is $D_s = U_{5^\circ,s} - U_{45^\circ,s}$. The practical margin is 0.02 of the available contextual improvement. A 95% paired confidence interval entirely outside [-0.02,0.02] supports a meaningful difference. A 90% interval entirely inside those bounds supports equivalence of the AUC estimand;

other outcomes are inconclusive. Equal areas do not imply pointwise equality of trajectories. The denominator uses the known population zero-predictor risk of 1, rather than the empirical zero-predictor risk in each bank. Replicate-level confidence intervals include variation from both training randomness and independently sampled evaluation banks; the 4096 episodes within a bank are not treated as independent model-training replicates.

The confirmatory sample size was fixed at 54 independent pairs before confirmation began. Four pilot pairs produced a paired SD of 0.02181; its one-sided 90% upper bound was 0.04942. Normal-data simulations at that conservative SD selected the first $n \geq 20$ achieving at least 80% power both to place a 95% CI beyond the 0.02 margin under a true difference of 0.05 and to establish 90%-CI equivalence under a true difference of zero. At n=54 the simulated powers were approximately 0.990 and 0.806, respectively. Selection used the variance, not the pilot mean. The pilot uses a shorter training horizon, so these are planning approximations rather than guaranteed powers for the confirmatory experiment.

Pilot outcomes are not pooled with confirmation. The frozen local protocol and manifests were committed before confirmatory execution; this was not an external preregistration. Secondary outcomes include final risk, first occurrence of three consecutive checkpoints below 0.1 normalized excess, and subsequent loss of the capability. Non-reaching runs are censored, not assigned a fictitious threshold time.

A.5 Robustness and computational provenance

Two prespecified secondary series use six independent pairs each. The width series increases hidden size to 256 (3,169,793 parameters), retaining 30,000 updates and all other primary settings. The batch series uses 1024 windows per update and 7500 updates, with a 125-update warm-up and evaluation every 250 updates. This preserves the number of training windows and the checkpoint grid measured in data exposure. It changes both the gradient-noise regime and the number of parameter updates, so it does not isolate gradient noise. These six-pair series have descriptive intervals and are not powered as additional confirmatory tests.

The execution uses three NVIDIA A100-SXM4-40GB Colab runtimes with Python 3.13.15, PyTorch 2.11.0+cu128 and CUDA 12.8. Each A/B pair remains on one runtime. Work was initially assigned to two runtimes and then redistributed across three by complete pair; interrupted runs resumed from serialized optimizer, data-generator and PyTorch RNG states. An independent serialized CUDA restart test reproduced uninterrupted weights and predictions exactly. No seed, sample-size, training-budget or numerical-precision adjustment followed this operational redistribution.

Recorded numerical and reporting libraries include NumPy 2.1.3, SciPy 1.16.3, pandas 2.2.3 and Matplotlib 3.10.0.

All saved checkpoints include predictions; the latest checkpoint additionally stores optimizer and RNG state. Checkpoints are copied to local persistent storage with archive SHA256 verification, while the original scientific computations remain in Colab. Historical pilot source hashes included macOS metadata sidecars. After removing these non-executable files, all actual Python sources matched the frozen per-file checksums. The canonical confirmatory source hash is `d6c6c17d3b6322a9f570ed16c9de7c7ebaf75d993c34316949c744d844c23c06`. The supplied scientific CLI scripts are the reproducibility entry points; the reported primary execution used these scripts.

B Additional primary-study results

On the independent final banks, mean clean-target MSE was 0.11351 at 5° and 0.00938 at 45°. The paired difference was 0.10414 (descriptive 95% CI 0.09988–0.10839). Table 1 reports the reference predictors on the same banks. The common circle oracle had risk 0.00447. The learned predictors outperform their respective finite-prior Bayesian predictors on this continuous test family, so prediction-risk comparisons do not support equating either learned predictor with exact finite-prior retrieval. Similarity of the 45° model's risk to ridge regression also does not establish algorithmic identity.

Table 1: Mean final clean-target MSE across 54 independent evaluation banks. Confidence intervals for neural risk are descriptive and unadjusted for secondary comparisons; other entries are reference-risk means.

Predictor	5° condition	45° condition
Transformer	0.11351 [0.10919, 0.11783]	0.00938 [0.00838, 0.01037]
Continuous circle posterior	0.00447	0.00447
Gaussian-prior ridge	0.00949	0.00949
Finite-prior posterior	0.17336	0.05104
Finite-prior MAP	0.17695	0.05515
Restricted linear control	0.16353	0.16353
Zero predictor	1.00145	1.00145

B.1 Prespecified robustness

All 24 prespecified robustness runs completed and passed their stage audits. At hidden size 256, the paired AUC difference was 0.11658 (descriptive 95% CI 0.10886–0.12430). With batch size 1024 and matched total data exposure, it was 0.08868 (0.08097–0.09638). Both series retained the direction observed in confirmation. Their six-pair intervals are unadjusted descriptive checks, not separately powered confirmatory tests, and we do not interpret differences between series as a formal width-by-geometry interaction.

Table 2: Prespecified robustness results. Brackets denote descriptive paired 95% t intervals. MSE is measured on independent final banks.

Series	Pairs	AUC difference	Final MSE 5°	Final MSE 45°
Width 256	6	0.11658 [0.10886, 0.12430]	0.13890	0.00850
Batch 1024	6	0.08868 [0.08097, 0.09638]	0.11661	0.00915

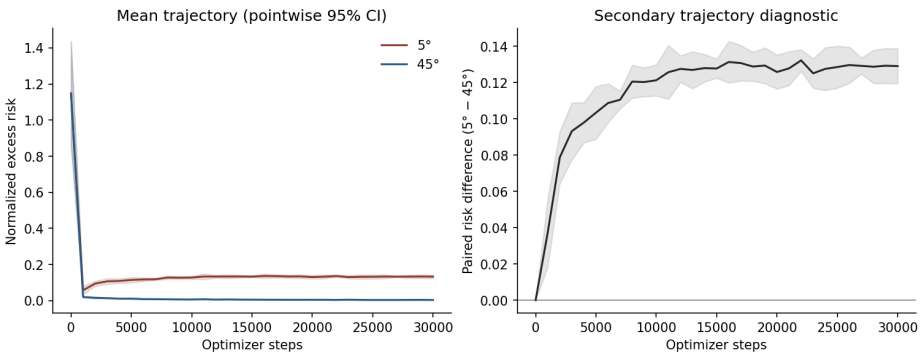

Figure 6: Width-256 robustness trajectories across six independent pairs. Shaded intervals are pointwise and descriptive.

The wide 5° model reached the sustained threshold in 2/6 runs, both of which later exceeded it; the wide 45° model reached it in 6/6 with no later exceedance. In the batch series, both conditions reached the threshold in 6/6 runs, followed by later exceedance in 5/6 at 5° and 0/6 at 45°. Thus, persistence of generalization remained a relevant behavioral distinction in both secondary series.

Aggregate outcomes for both robustness series accompany the code supplement; complete per-example arrays and checkpoints are retained separately.

B.2 Completion and resource record

The final return passed a local inventory of 144 unique complete runs and 10,228 files, including all required weights, predictions, evaluation banks and latest optimizer/RNG states. The independently

repeated complete primary analysis matched the earlier complete primary return exactly. All three task-owned A100 sessions and the final CPU analysis session were stopped after their artifacts were returned and verified.

Table 3: Recorded resources for the non-pilot series. Timing sums completed training segments only; it excludes prior interrupted segments, setup, transfer, audits and runtime idle time, and is not billed GPU time or compute units. Memory is peak allocated tensor memory reported by PyTorch, not total device or reserved memory.

Series	Runs	Updates	Training windows	Segment seconds	Peak allocated bytes
Confirmation	108	3,240,000	829,440,000	28,720	340,029,440
Width	12	360,000	92,160,000	5,545	678,913,536
Batch	12	90,000	92,160,000	1,926	1,268,625,920

The data contain numeric tokens of length 33 per training window; these counts should not be compared directly with natural-language token counts. The frozen manifests, per-run outcomes, source tables and checksum inventory accompany the manuscript. Scientific calculations for the original studies, including final aggregation, ran in Colab. The initial scope checks ran on RunPod H200; the subsequent full-dimensional transfer ran on Colab A100. Local computation handled file integrity, independent aggregation, and document production.

C Dimensional and supervision scope checks

Earlier exploratory four-dimensional runs at 5000 updates were heterogeneous. We therefore fixed a fresh six-pair extension at the original 30000-update horizon, with two fresh two-dimensional control pairs. In dimension d, the prior contains a signed coordinate basis and its rotation by either 5° or 45° within each two-dimensional coordinate plane, followed by a shared random orthogonal transformation. Each prior has 4d equally probable unit vectors, mean zero, and covariance I/d. Thus task count and covariance match within each dimension; the change from eight to sixteen tasks prevents interpreting the cross-dimensional comparison as an isolated dimension intervention. Evaluation tasks are uniform on the unit sphere. Both d=4 training supports lie in a union of two orthogonal two-dimensional subspaces before the common rotation; the test family fills the sphere. This shared support restriction limits extension to generic four-dimensional priors. Keeping 16 demonstrations also changes the demonstration-to-dimension ratio from 8 to 4, so this is a transfer test at fixed context length, not an isolated scaling law.

The architecture and AdamW settings match the primary study, including 500 warmup steps; only the input projection changes with dimension. Trajectory evaluation uses 2048 independent task/context draws per pair, with an independent 8192-draw final bank. Seeds are fresh and paired across the two priors. Replicate counts were fixed from compute timing before these outcomes. The primary descriptive endpoint for this extension is the final clean-risk contrast; its paired t interval has five degrees of freedom. It is not a new powered confirmation or pooled enlargement of the 54-pair study.

At d=4, mean final MSE was 0.33422 at 5° and 0.21699 at 45°. The paired difference was 0.1172 (descriptive 95% CI -0.1029 to 0.3374). This extension does not establish a stable positive transfer of the original contrast to d=4; the original two-dimensional confirmation remains valid within its own scope. The raw-MSE training-budget average contrast was 0.0666 (descriptive 95% CI -0.1517 to 0.2849). Unlike the original primary AUC, this secondary area is not normalized by an oracle. The two fresh d=2 controls had mean final risks 0.11936 and 0.00829; no inferential claim is made from these two controls.

C.1 Prespecified full-dimensional four-dimensional transfer

The six-pair result above was not pooled with the follow-up. Before new outcomes, we fixed 40 independent pair IDs 400–439, a final-risk primary endpoint, and a practical decision margin. Each

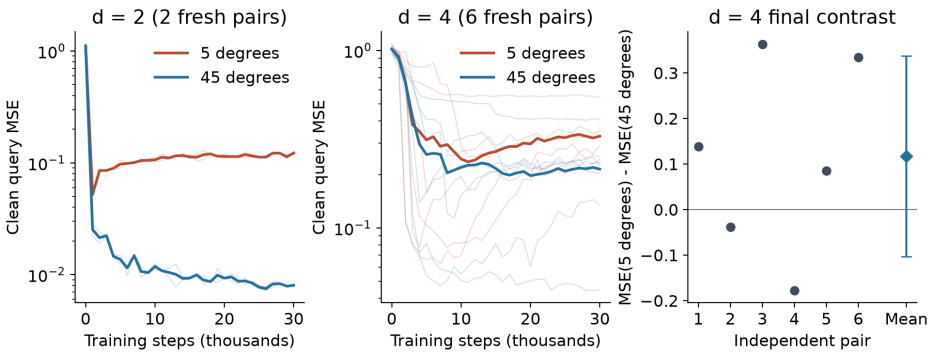

Figure 7: Fresh dimensional-transfer trajectories and individual final contrasts. Thin curves are independent training replicates, thick curves are means, and the vertical axis of each trajectory panel is logarithmic, with different vertical ranges. The final-contrast interval is a descriptive paired t interval across six d=4 pairs.

training prior is uniform on the positive and negative rows of an orthonormal basis and a second orthonormal basis obtained through a dense four-dimensional rotation. Specifically,

$$A = \frac{1}{\|A_0\|_2} A_0, \qquad A_0 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ -1 & 0 & 4 & 5 \\ -2 & -4 & 0 & 6 \\ -3 & -5 & -6 & 0 \end{pmatrix}, \qquad R_\theta = \exp((\theta\pi/180)A)\,.$$

The first four unoriented directions are the rows of I_4 and the next four are the rows of R_θ; their negatives produce the 16 signed tasks. A common random global orthogonal transform is applied within each pair. Both conditions therefore contain 16 equally probable unit tasks, have zero mean and covariance $I_4/4$, and have two component bases that each span all four dimensions. The technical gate removed antipodal duplicates before testing general position: every triple among the eight unoriented directions had rank three, with minimum third singular values 0.01383 at 5° and 0.07062 at 45°. No such claim is made for arbitrary triples of the signed support, which includes both v and $-v$. The task law is nevertheless finite and does not have continuous sphere support.

The architecture, final-query objective, optimizer, context length, noise level, and 30,000-update budget match the earlier four-dimensional extension. Task identities are sampled independently within each minibatch, whereas the earlier extension used task-balanced minibatches. Both geometry arms use the same sampling scheme and paired task-index stream. Within each pair, the two angles share initialization, task-index stream, inputs, noise, evaluation banks, and global rotation. The independent final bank contains 8,192 episodes per pair. The primary endpoint is final clean $\mathrm{MSE}_{5^\circ} - \mathrm{MSE}_{45^\circ}$. A 95% paired-$t$ interval with lower endpoint above 0.02 defines a positive result; a 90% interval entirely within $[-0.02, 0.02]$ defines practical equivalence. No interim outcome analysis or seed extension was allowed.

The primary mean is 0.252886 (SD 0.260570; 95% CI $[0.169552, 0.336220]$), producing the prespecified positive decision. A one-sided test against the 0.02 margin gives $p = 7.86 \times 10^{-7}$. Thirty-four of 40 pairs are positive (exact sign $p = 8.36 \times 10^{-6}$); the two-sided Wilcoxon signed-rank $p = 1.46 \times 10^{-7}$. A deterministic 200,000-draw pair bootstrap gives $[0.173870, 0.333609]$, and every leave-one-pair-out mean is positive, ranging from 0.239102 to 0.269600. The prespecified secondary contrasts were a raw training-budget AUC difference of 0.204265 (95% CI $[0.114775, 0.293754]$) and an early-to-final risk-change difference of 0.164300 (95% CI $[0.060783, 0.267817]$). The latter endpoint, termed *retention* in the frozen protocol, is $C_\theta = R_\theta(30{,}000) - R_\theta(1{,}000)$ on the shared trajectory bank; the contrast is $C_{5^\circ} - C_{45^\circ}$. It does not by itself establish acquisition followed by loss and is distinct from H1's three-consecutive-checkpoint criterion and the post hoc checkpoint-selection audit below.

All 80 arm archives passed local SHA256 verification and complete Google Drive readback before entering the ledger. A separately implemented author-run analysis recomputed 2,600 checkpoint and final metrics from saved predictions, including equality of paired final banks. Runtime interruptions discarded only unacknowledged work; frozen seeds and endpoints were unchanged. The source

archive, attempt ledger, per-arm archive checksums, analysis code, and machine-readable summaries accompany the reproducibility record.

In a separate post hoc checkpoint-selection and deterioration audit, we restored all 80 arm archives, verified their SHA256 digests, and reconstructed all 2,480 checkpoint risks (maximum discrepancy 4.44×10^{-16}). Ridge predictions reconstructed from the stored tokens differed from the saved independent-final-bank ridge predictions by at most 2.57×10^{-7}. For each arm, the lowest-risk checkpoint among the 31 recorded steps was selected using 1,024 episodes from the fixed trajectory bank and evaluated on the other 1,024; the halves were then swapped and the held-out estimates averaged. This prevents the selected minimum from being evaluated on the same episodes that selected it.

The mean selected held-out MSE was 0.43571 at 5° and 0.21190 at 45°, versus mean trajectory-bank ridge MSE 0.02184. Median arm-level selected-risk-to-ridge ratios were 20.56 and 4.62, respectively; thus early relative improvement in 4D should not be called ridge-level generalization. Final-minus-selected held-out deterioration averaged 0.06992 at 5° (33/40 positive; descriptive pair-bootstrap 95% interval $[0.04947, 0.09169]$) and 0.04293 at 45° (33/40 positive; $[0.03164, 0.05482]$). The paired difference in deterioration averaged 0.02699 (bootstrap interval $[0.00480, 0.04953]$), but was positive in only 24/40 pairs. The independent 8,192-episode final bank reproduced the final-risk scale (means 0.50622 and 0.25333, versus 0.50563 and 0.25484 on the trajectory bank); saved intermediate states were not reevaluated on that second bank. Selected steps span 2,000–30,000 at 5° and 1,000–30,000 at 45°. Hence these exploratory measurements support heterogeneous post-best deterioration, not a single identified transition point or the claim that either condition consistently reached ridge-level generalization before deteriorating. The frozen final-risk decision is unchanged.

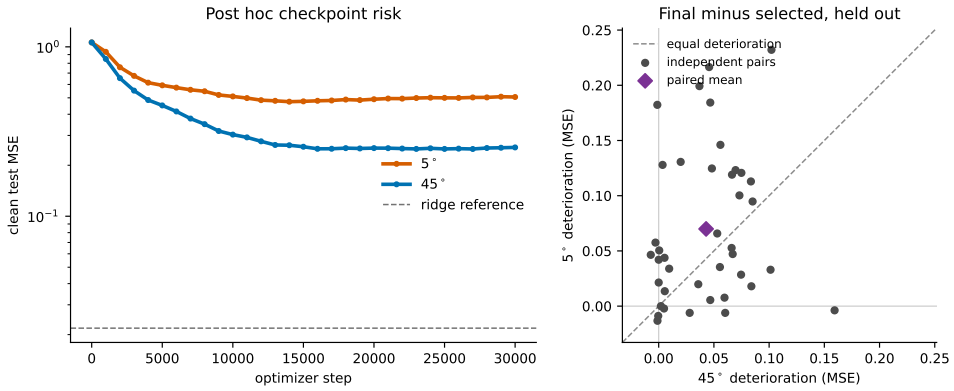

Figure 8: Post hoc D4 checkpoint-selection and deterioration audit. Left: mean clean-query risk across the 40 independent pairs at each geometry, with the shared-bank ridge reference. Right: each pair's out-of-fold final-minus-selected risk deterioration; the diagonal marks equal deterioration in the two geometries. Checkpoint selection and evaluation use disjoint episodes from the trajectory bank. The separate final bank evaluates only the final checkpoint.

C.2 Supervision across context lengths

A separate four-pair exploratory check uses fresh seeds and the original d=2 priors, but trains the same Transformer to predict every response with 1 through 16 available demonstrations. The loss averages the 16 noisy-response squared errors within each window; these losses are correlated and are not counted as independent replicates. The zero-demonstration position is excluded. Evaluation remains clean risk at 16 demonstrations on the continuous-circle family. Causal-mask checks verified that changing current and future labels cannot affect the corresponding predictions. All other training settings, step counts, and evaluation-bank sizes match the scope checks in the main text.

Mean final MSE was 0.12964 at 5° and 0.00606 at 45°. The paired geometry contrast was 0.1236 (descriptive 95% CI 0.1140 to 0.1331), with three degrees of freedom. The positive contrast persists under all-prefix supervision in this small exploratory sample. Since there is no freshly trained last-query-only arm with the same seeds, it does not estimate a paired causal effect of changing the

objective. It also leaves the continuous task-family prior different from the Gaussian-prior regression setting of Wurgaft et al. (2025); this control is not a full replication or a refutation of that work.

C.3 Confirmatory paired-objective experiment

The sample-size pilot used four quartet blocks and was excluded from inference. Its frozen variance rule selected the maximum planned main sample: 16 new blocks (IDs 200–215), each crossing 5°/45° geometry and last-query/all-prefix supervision. All 64 models trained for 30,000 updates from common within-block initializations and data random-number streams. The quartet is the statistical unit; checkpoints and arms are not additional replicates. Unlike H1, which uses population zero-predictor risk equal to one, this experiment normalizes excess risk using the zero predictor's empirical risk on the shared evaluation bank: $(R_{\text{network}} - R_{\text{circle}})/(R_{\text{zero}} - R_{\text{circle}})$. The primary interaction is computed from these empirically normalized AUCs.

The 16 primary interactions have mean 0.026756 and standard deviation 0.021728. The paired t interval is $[0.015177, 0.038334]$, with $t_{15} = 4.9255$ and two-sided $p = 0.000183$. It excludes zero and is not contained in the prespecified equivalence interval $[-0.02, 0.02]$. The simple geometry contrasts, $\text{AUC}_{5^\circ} - \text{AUC}_{45^\circ}$, are 0.093304 under last-query supervision (descriptive 95% CI $[0.080012, 0.106596]$) and 0.120060 under all-prefix supervision ($[0.115328, 0.124791]$); both are positive in all 16 blocks, and their difference is the primary interaction. Descriptive all-prefix-minus-last-query AUC contrasts are 0.01750 at 5° and -0.009257 at 45°; the descriptive final-risk interaction is 0.01685. All 384 checkpoint archives passed full Drive readback, and a fresh CPU runtime reproduced the primary summary.

A post-confirmatory sensitivity audit retained all blocks. The exact two-sided sign-flip test gives $p = 0.000549$, the Wilcoxon signed-rank test gives $p = 0.000580$, and a deterministic 200,000-draw block bootstrap gives a percentile 95% interval $[0.01636, 0.03692]$. Fourteen of sixteen interactions are positive; leave-one-block-out means range from 0.02472 to 0.02908. These checks were specified after the primary result was known and are robustness analyses, not an additional confirmation.

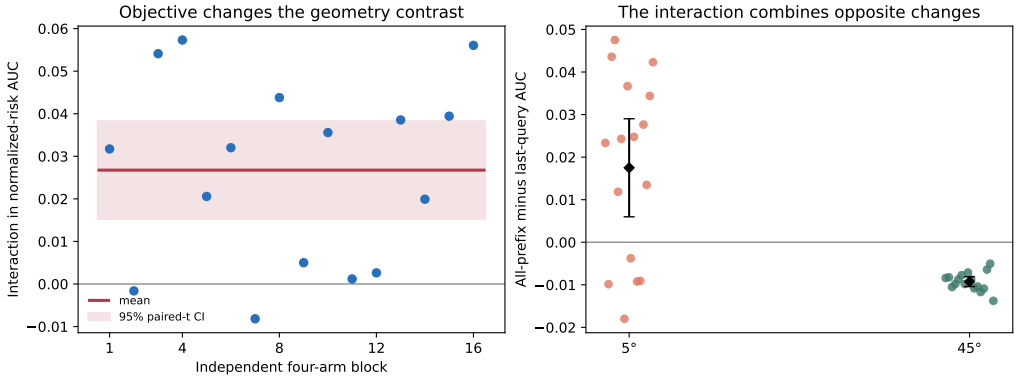

Figure 9: Paired-objective main experiment. Left: every independent four-arm block, the interaction mean, and its paired-t interval. Right: the two simple all-prefix-minus-last-query effects with descriptive paired intervals. The positive interaction combines increased normalized-risk AUC at 5° with decreased AUC at 45°.

C.4 Targeted replication of source-style transient generalization

We used the public generator and GPT-NeoX implementation of Wurgaft et al. (2025) at source commit `7fcd0a67`: dimension eight, 64 fixed Gaussian training tasks, Gaussian out-of-distribution tasks, context length 32, eight Transformer blocks of width 64, one attention head, and 100,000 updates. The positive-control objective averages squared error over all output positions, as in the source implementation. For each of three fresh seeds, a paired arm shares initialization, task bank, data stream, and evaluation bank but computes loss only at the final position. The technical gate

reproduced the source loss exactly and gave bitwise-identical uninterrupted and resumed 200-step runs under both objectives.

At each checkpoint we project model predictions onto the best common convex mixture of the finite-prior memorizing predictor and Gaussian-prior generalizing predictor. The frozen transience index is the final fitted memorizing weight minus its minimum after the two-predictor residual enters a threshold 10% above its trajectory minimum. This threshold was defined relative to the observed trajectory minimum, so every arm necessarily enters it; entry alone is not evidence that the mixture fits well. The original coefficient and index pooled all 32 prediction positions, including positions not directly supervised by the final-query arm. For each position scope, entry is the first recorded checkpoint whose same-scope residual is at most 1.1 times its trajectory minimum. Transience is the final fitted memorizing weight minus its minimum over all recorded checkpoints from entry onward; the minimum is not restricted to checkpoints meeting the residual threshold. These coefficients are fitted on the full evaluation bank. The two-fold analysis below separately tests held-out approximation error. All-output transience was 0.19046, 0.20924, and 0.30222; paired final-query transience was 0.00426, 0.10421, and 0.00634. The three paired differences were 0.18619, 0.10502, and 0.29588. Their mean is 0.19570 and their descriptive 95% paired-t interval is $[-0.04224, 0.43363]$. The interval is wide because the protocol fixes only three mechanistic pairs; checkpoints and evaluation examples are not additional replicates.

A post hoc position-matched reanalysis fitted the same common mixture separately on the last response and on the preceding 31 responses. On the last response, all-output transience was 0.21579, 0.24469, and 0.35067; matched final-query transience was 0.00351, 0, and 0. The three paired differences were 0.21229, 0.24469, and 0.35067 (mean 0.26921; descriptive 95% paired-t interval $[0.08942, 0.44901]$). The direction therefore survives, and becomes larger under the position-matched measurement, but this reanalysis is post hoc and three pairs do not establish a universal objective effect. Earlier positions are shown separately in Figure 11.

To audit fit quality independently of the threshold, we used the 500 fixed OOD tasks in a two-fold even/odd task split at every checkpoint and position scope. One common mixture coefficient was fitted on 250 tasks and tested on the other 250, then the roles were reversed and test errors pooled. At the final checkpoint on the last response, mean held-out prediction-space residual was 1.55227 MSE for all-output training and 0.01559 for final-query training. Relative to the same-scope distance from the network to its nearer single Bayesian endpoint, these held-out residuals were 0.761 and 0.985, respectively; one denotes no improvement over that endpoint, not perfect prediction. The normalizer is the smaller of the two full-bank, same-scope prediction distances from the network to the individual reference predictors; it is descriptive, not the performance of a separately trained endpoint-selection rule. Across all 198 checkpoint-by-arm-by-scope records, the held-out residual exceeded the within-bank optimum by at most 2.90% of that optimum. Full absolute and relative residual trajectories, including the preceding positions, appear in Figure 11 and the accompanying CSV. A small fit/test gap does not make the absolute residual small, nor is prediction-space distance an additive component of clean test risk.

Every one of the 60 nonzero checkpoint packages passed local SHA256 verification and complete Google Drive readback. A fresh CPU runtime recomputed all 66 checkpoint metrics from saved individual predictions with maximum absolute discrepancy zero. The experiment targets one reported condition and does not reproduce the complete task-count-by-diversity map or fit the source hierarchical model.

For the four-pair two-dimensional all-prefix experiment in Appendix C.2, we separately audit a common convex mixture of the finite-prior and continuous-circle posterior predictors. A scalar weight is fit to the network's answers on the first 4096 examples of each independent final bank; prediction-distance MSE is evaluated on the other 4096. Circle quadrature at 2048 and 4096 points passed the fixed convergence tolerances. Table 4 separates held-out distance from the network's actual task risk. The audit measures the approximation error of this common-weight dictionary, not an additive risk component or the identity of an internal algorithm.

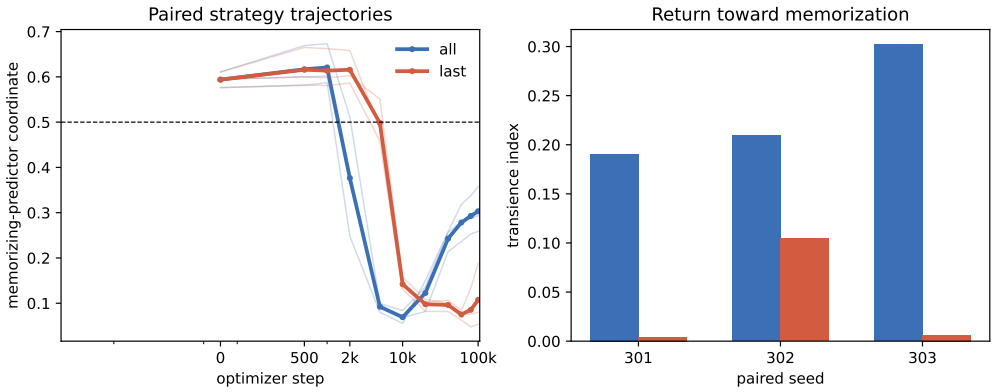

Figure 10: Targeted closest-source replication. The author-style all-output objective shows an early move toward the generalizing predictor followed by a return toward memorization in every seed. Final-query supervision attenuates this return in each matched pair. Lines are independent training seeds; the fitted coordinate is a diagnostic projection and its nonzero residual is reported separately. Fitted coordinates and transience indices here pool all 32 response positions; Figure 11 separates the last position from the preceding positions.

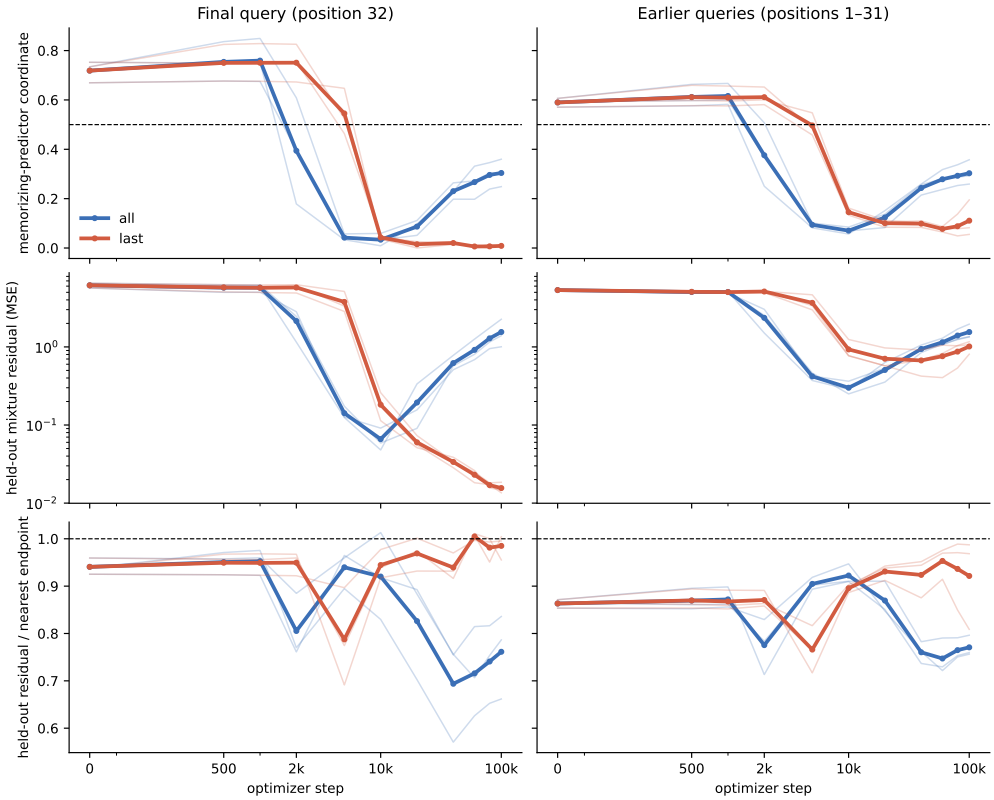

Figure 11: Post hoc W position-matched audit. Left: the last response, which both objectives directly supervise; right: the preceding 31 responses, shown separately. Thin lines are the three paired seeds and thick lines their means. The middle panels give the complete out-of-fold prediction-space residual curves; the bottom panels divide those residuals by the distance to the nearer single reference predictor at each checkpoint. The coefficient was fitted on disjoint tasks from those used to score each held-out point.

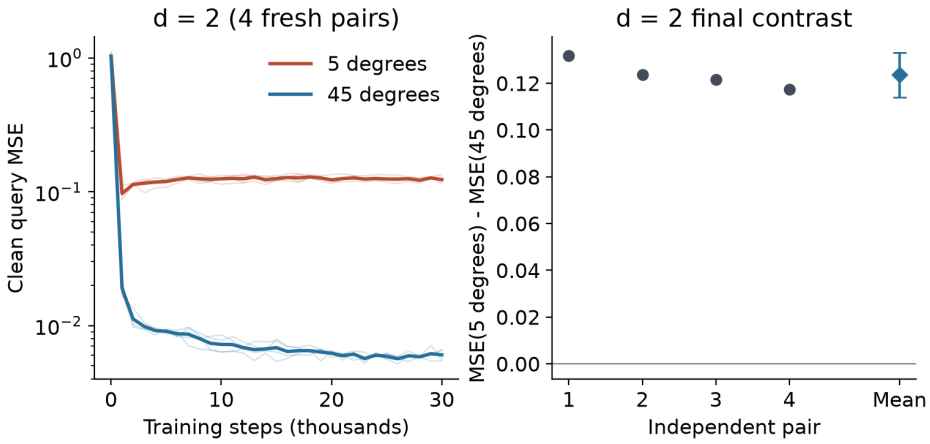

Figure 12: Four-pair two-dimensional geometry contrast with supervision across context lengths, separate from the W replication above. Thin curves show four independent pairs; thick curves show means. Final risk uses an independent evaluation bank at 16 demonstrations. The final-contrast interval is descriptive and unadjusted.

Table 4: Response-dictionary audit for the four-pair two-dimensional all-prefix experiment in Appendix C.2. Entries are means across four trained models per geometry. Mixture weights are fitted on separate contexts from those used to measure prediction distance and task risk.

Prior	Fitted finite weight	Held-out prediction-distance MSE	Held-out model task MSE
5 degrees	0.6918	0.043895	0.126745
45 degrees	0.0033	0.001591	0.005932

The 24 runs from the two initial exploratory series—12 four-dimensional runs, four two-dimensional control runs, and eight all-prefix runs—completed on one NVIDIA H200 using PyTorch 2.8.0+cu128, FP32 and disabled TF32. Checks reconstructed every saved trajectory/final risk from per-example arrays and replayed final saved weights on their complete evaluation banks. The scientific protocols and training source snapshots were committed before their respective runs. A timing-only scheduler clarification during execution retained the same planned sample, seeds, endpoints, and shared wall budget; any use of timing recovery is recorded separately. Primary H1 sources and outcomes remain unchanged.

D EXPLANATORY TESTS: DERIVATIONS, DESIGN, AND DETAILED RESULTS

The earlier continuation experiment provides a causal starting point: changing the training prior from common weights altered subsequent risk in both directions, including after optimizer-state reset (the continuation intervention in the main text). We next ask which descriptions can predict network behavior beyond the original pair of geometries.

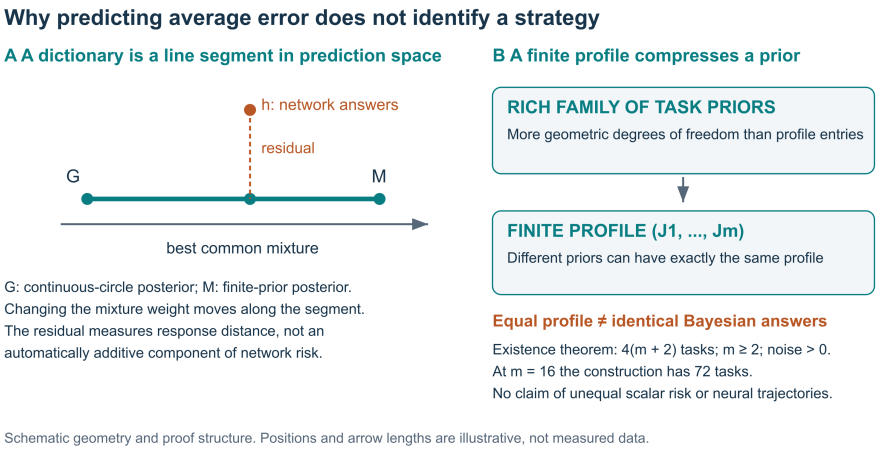

Figure 13: Two distinct limits on identifying a strategy. Left: a common convex mixture spans a line segment in prediction space; the illustrated optimum lies in its interior, while endpoint clipping is needed in general. Right: the existence theorem concerns equal finite profiles with different Bayesian answer functions in a richer prior family. Coordinates are schematic.

D.1 STRATEGY DICTIONARIES AND LOCAL RESPONSE

Our follow-up separates three explanatory levels. A distributional descriptor predicts which training conditions are difficult. A strategy dictionary approximates the network's answers. A dynamical account predicts how these answers change under further training. Success at one level does not establish the others. In particular, matching a Bayesian predictor's task risk does not imply matching its predictions.

The closest strategy-selection account (Wurgaft et al., 2025) motivates a mixture of finite-prior and generalizing predictors. Its regression generalizer uses a Gaussian task prior; our continuous task law is uniform on the circle. We therefore evaluate both the covariance-matched ridge comparator and a numerically converged circle-posterior comparator. We retain an explicit distinction between a functional-form adaptation to our last-query objective and replication of the original study. The classification dynamics in (Nguyen & Reddy, 2024) involve a match-to-sample mechanism and cannot be inserted unchanged as a quantitative regression model. The relevance of geometric coverage (Goddard et al., 2025) and covariance alignment (Letey et al., 2025) further limits novelty claims to the controlled dynamics and their quantitative explanatory tests.

Let M denote the finite-prior posterior predictor and G the continuous-circle posterior predictor. Let h denote a trained predictor and $v = M - G$, with inner products taken over a fixed distribution of test episodes. If $\|v\|^2 > 0$, define

$$a_* = \frac{\langle h - G, v \rangle}{\|v\|^2}, \qquad r = h - G - a_* v.$$

Orthogonality gives the elementary identity

$$\|h - (G + av)\|^2 = \|r\|^2 + (a - a_*)^2 \|v\|^2.$$

Thus clipping a_* to the unit interval yields the smallest approximation error attainable by any common convex weight. This lower bound distinguishes a failure of the two-predictor dictionary from an inaccurate temporal law for selecting its weight. It does not constrain episode-dependent weights or richer dictionaries. An empirically optimized lower bound is optimistic; we additionally estimate the weight on separate contexts and evaluate its transferred approximation error.

For a complementary local account, write $L_P(\theta)$ for training risk under prior P and $R_Q(\theta)$ for a fixed test risk. Population gradient descent from common weights gives $\theta'_P = \theta - \eta g_P$, where $g_P = \nabla L_P(\theta)$. If the test-risk Hessian has operator norm at most H on both update segments, Taylor's theorem gives

$$R_Q(\theta'_B) - R_Q(\theta'_A) = -\eta\langle \nabla R_Q, g_B - g_A\rangle + \epsilon,$$

$$|\epsilon| \le \frac{H\eta^2}{2}(\|g_A\|^2 + \|g_B\|^2).$$

For stochastic gradients the linear term uses their expectations, and the remainder bound uses their second moments. For AdamW, the corresponding expansion must use actual parameter increments, including its optimizer state. We therefore test one-step SGD interventions on AdamW-trained states as a distinct local experiment. Gradients evaluated on a separate fit bank predict a risk contrast evaluated on an independent test bank. These measurements have access to the current network and are not prior-only forecasts.

In this empirical design, prediction error includes Monte Carlo error in the estimated test gradient. Consequently, an observed discrepancy need not scale quadratically with the step size even when the underlying Taylor remainder does. Failure of the empirical predictor would not falsify Taylor's theorem; conversely, predictive success would not certify an unmeasured uniform Hessian bound.

The link to matched moments can be made explicit without assuming convergence of an infinite Fourier expansion. Define $s_\theta(\phi) = \langle \nabla R_Q, \nabla L(\theta \mid \phi)\rangle$. If a trigonometric polynomial $s_K = \sum_{|k|\le K} c_k e^{ik\phi}$ has uniform approximation error at most δ_K, then, for $\mu_k(P) = \mathbb{E}_P[e^{ik\phi}]$,

$$\left| \int s_\theta \, d(P_B - P_A) - \sum_{|k|\le K} c_k[\mu_k(P_B) - \mu_k(P_A)] \right| \le 2\delta_K.$$

The result follows by bounding each probability-weighted remainder. Quarter-turn symmetry eliminates modes not divisible by four; covariance matching alone does not eliminate the remaining modes. A sign guarantee requires the predicted linear contrast to exceed both approximation and Taylor remainders. Finite-grid measurements do not certify the required uniform bound. These are conditional elementary results, not a new universal convergence theorem or a guarantee about long training trajectories.

D.2 Independent prediction and intervention design

The next stage was fixed after a development pilot and before observing its new training outcomes. The pilot contained eight task geometries from two generators, with two training seeds per geometry. Every prior consists of four quarter-turn orbits with equal weights, giving sixteen distinct unit vectors, zero mean and covariance I/2. The development generators placed orbit origins in a random narrow sector or in reflected pairs. The two evaluation generators used four independent angular origins or two separated pairs with independently varying within-pair distances. Generator definitions were written before the pilot was analyzed.

We froze a comparison between a finite-Bayes-risk regression baseline and a geometric predictor adding the magnitudes of the fourth and eighth angular moments. The model choices and regularization were selected using leave-one-geometry-out validation on the pilot. Constant, information-profile and explicitly adapted strategy-law comparators were retained as secondary results. The latter was fitted in prediction space; its early acquisition regime was represented by a shared development curve rather than assuming an untrained network already implemented either Bayesian

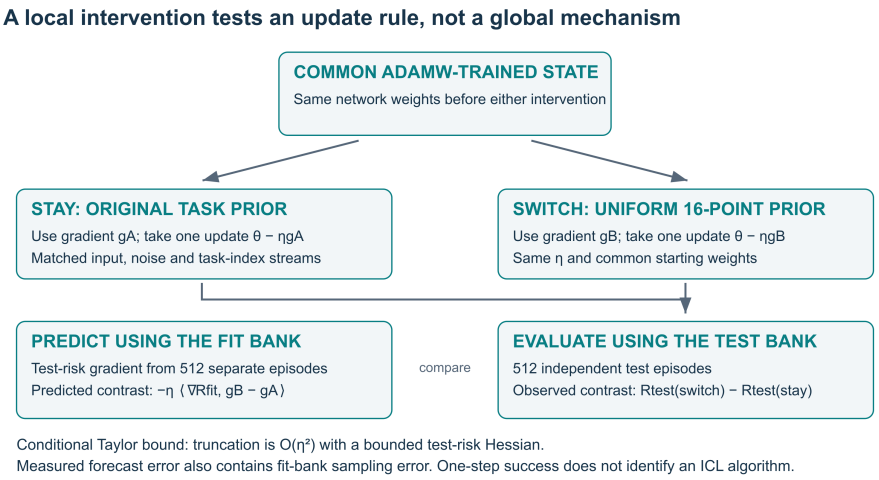

Figure 14: Local intervention and measurement separation. Both updated networks are evaluated on the same independent test bank. The prediction combines both update gradients with a test-risk gradient estimated on a separate fit bank; compare denotes an error comparison, not data flow between the banks. The tested step is SGD on AdamW-trained weights.

endpoint. Generator descriptors used separate 8192-window banks. No predictor was refitted on the new evaluation generators.

For exact reconstruction, every retained linear forecast has the form

$$\widehat{U}(x) = \bar{U} + \beta^\top z, \qquad z_k = (x_k - \mu_k)/s_k,$$

where $\bar{U} = 0.1101233985$, the feature means and population standard deviations are computed only on the eight pilot geometries, and the intercept is unpenalized. The finite-Bayes baseline uses $x = (r_{\text{finite}})$, $\lambda = 0.1$, $\mu = (0.0915117202)$, $s = (0.0692898254)$, and $\beta = (0.0186404661)$. The primary geometry forecast uses $x = (r_{\text{finite}}, |\mu_4|, |\mu_8|)$ and $\lambda = 0.01$, with

$$\begin{aligned}\mu &= (0.0915117202,\ 0.5517936661,\ 0.7304272471),\\ s &= (0.0692898254,\ 0.4224896827,\ 0.2063796325),\\ \beta &= (0.1286938920,\ -0.1032079310,\ -0.0093374581).\end{aligned}$$

Coefficients solve $(Z^\top Z + \lambda I)\beta = Z^\top(U - \bar{U})$ after this fixed standardization. The secondary information forecast uses $(r_{\text{finite}}, \Delta L, J_1, J_2, J_4, J_8, J_{16})$ with $\lambda = 10$ and

$$\begin{aligned}\beta_{\text{info}} = (&0.00333438,\ 0.00354484,\ 0.00284913,\ 0.00294290,\\ &0.00279458,\ 0.00049517,\ 0.00372076).\end{aligned}$$

Its complete frozen centering and scaling vectors are included in the supplementary `frozen_predictors.json`. The adapted strategy law uses frozen parameters $(\alpha, \beta, \log\gamma) = (0.38405263, 0.34375568, 1.06607051)$ and $K_G = 1344$. These numerical specifications were frozen before the 106 evaluation geometries and reproduce the published forecasts without refitting.

Here $J_j(P) = \mathbb{E}\|\mathbb{E}_P[w \mid D_j]\|^2$ denotes the finite-prior Bayes risk reduction from observing j demonstrations, with unit-norm tasks and isotropic Gaussian queries. It is not mutual information. The information comparator uses J at context lengths 1, 2, 4, 8 and 16 together with finite-prior OOD risk and the training-loss gap. It is not an isolated test of adding the full sixteen-entry J profile, nor does its performance identify a unique contribution from J.

The pilot planning standard deviation was inflated to its upper 80% bound. A normal approximation targeting 80% power for a 0.005 improvement in normalized AUC prediction error, with a conservative two-sided alpha of 0.025, gave 106 geometries. We therefore fixed 53 independent geometries

from each evaluation generator, with two seeds each, yielding 212 trainings. This planning calculation does not guarantee the same power under a different generator distribution. The original 54-pair H1 confirmation remains a separate experiment.

All new models used the original four-layer Transformer architecture with hidden width 128, batch size 256, sixteen demonstrations, noise standard deviation 0.25, AdamW and a learning rate of 0.0003 after 250 warmup updates. Training lasted 5000 updates. TF32 was disabled. Curves used 1024 evaluation episodes at updates 0, 250, 500, 1000, 2000, 3000, 4000 and 5000, with an independent 2048-episode final bank. We retained the optimizer and random-stream state for deterministic resumption. Numerical posterior integration and matched-moment controls preceded the scientific runs.

The primary forecast target was the trapezoidal training-budget average of normalized excess risk. Within each geometry, predicted and observed AUC were first averaged over the two training seeds. Absolute forecast errors were then calculated, and their paired difference compared the frozen baseline and geometric predictor. This ordering matches the pilot estimand; averaging absolute seed-level errors would define a different quantity.

The second primary outcome assessed local prediction at the final trained state. From common weights, one SGD update under the original prior was compared with one under a uniform sixteen-point circle prior. Both branches used the same sampled inputs, noise and task-index streams. A test-risk gradient estimated on 512 fit episodes predicted the risk contrast evaluated on an independent 512-episode test bank. The fixed primary step size was 0.0001. Additional step sizes 0.00001, 0.00003 and 0.0003, and the 1000-update source state, were secondary diagnostics. This experiment changes the learning distribution and deliberately uses a one-step SGD update on AdamW-trained weights; it does not claim to reproduce the earlier long AdamW continuations.

For the local outcome, the improvement was absolute error of the zero-effect forecast minus absolute error of the gradient forecast. Seed-level improvements were averaged within geometry. The practical reference magnitude of 0.00001 MSE was selected from the pilot scale before confirmation. The local forecast has access to current weights and gradients; its comparison with zero should not be interpreted as superiority over a distribution-only theory with the same information budget.

Both primary effects used equal weighting of the two fixed evaluation generators, separate within-generator variance estimates and Welch–Satterthwaite degrees of freedom. Two-sided p-values were adjusted by Holm. We report pointwise 95% intervals and conservative 97.5% intervals for the two primary quantities. A 5000-repetition heteroskedastic null simulation checked the implementation, yielding rejection frequency 0.0524 and 95% interval coverage 0.9476. These intervals condition on the frozen development-stage predictors and characterize the specified generator population, not arbitrary task families or all possible predictor-fitting datasets.

A prespecified width-transfer check used the first six geometries of each evaluation generator, two seeds each, at hidden width 64. Its 24 trainings used the same frozen predictors without target-width adaptation and were designated exploratory. Matching entire information profiles and intervening on within-window task persistence were not part of this test; strong information-profile sufficiency remains untested.

D.3 Results of the frozen explanatory tests

All 212 confirmatory trainings and all 24 width-transfer trainings completed. Independent replay of each final checkpoint reproduced the saved predictions exactly. Configuration checks found complete coverage with no substituted seeds. The primary analysis was executed only after this coverage check; no predictor was refitted or replaced after seeing the results.

The geometric predictor failed its proposed improvement test. Its mean absolute AUC forecast error was 0.02852, versus 0.01624 for the finite-Bayes-risk baseline. The paired improvement was negative, -0.01228 (95% CI -0.01650 to -0.00806; Holm-adjusted p=9.14e-8). The negative direction occurred in both evaluation generators: -0.01206 for paired geometries and -0.01249 for random geometries. The result is evidence of worse transfer for this particular frozen predictor, not merely an inconclusive improvement test. It does not establish that geometric information is intrinsically unhelpful or that another fitted model could not use it.

Table 5: Prespecified primary effects across 106 geometries. Positive values denote better prediction. Intervals are stratified by the two fixed generators; p-values use Holm adjustment across these two tests.

Quantity	Mean improvement	Conservative 97.5% interval	Holm p
AUC forecast A	-0.01228	[-0.01712, -0.00743]	9.14e-8
Local contrast B, MSE	6.61e-6	[5.21e-6, 8.01e-6]	9.35e-18

The local gradient forecast reduced the mean absolute error of the zero-effect prediction from 9.97e-6 to 3.36e-6 MSE. Its improvement was 6.61e-6 (95% CI 5.39e-6 to 7.83e-6). This supports predictive usefulness at the tested state and step size. However, even the upper conservative interval endpoint is below the prespecified practical reference of 1e-5. Statistical evidence for a positive improvement therefore does not imply attainment of the practical target. Nor does this result identify a specific ICL algorithm: a local Taylor prediction can apply to many trained differentiable models.

We additionally reanalyzed the saved run-level contrasts at both retained stages and all four step sizes (Table 6). Let Δ_i and $\widehat{\Delta}_i$ denote the observed and predicted one-step contrasts. Define $\mathrm{MAE}_{\mathrm{forecast}} = N^{-1}\sum_{i=1}^{N}|\widehat{\Delta}_i - \Delta_i|$ and $\mathrm{MAE}_0 = N^{-1}\sum_{i=1}^{N}|\Delta_i|$. Relative error reduction is $1 - \mathrm{MAE}_{\mathrm{forecast}}/\mathrm{MAE}_0$. Each table row contains $N = 212$ runs from 106 geometries; these runs are not treated as 212 independent geometries. At the prespecified final-state step size, this reduction is 66.28% (9.974e-6 to 3.363e-6 MSE). Because an ideal forecast cannot improve absolute error by more than the zero-change MAE, the sample mean maximum possible improvement is 9.974e-6, just below the fixed 1e-5 practical reference. This scale observation does not change the prespecified decision. The predicted effect has the correct sign in 192/212 runs (90.57%); the majority-sign comparator attains 111/212 (52.36%). An ordinary descriptive regression of observed on predicted contrasts has slope 0.947 and intercept 0.312e-6 MSE, versus ideal values one and zero. These diagnostics are post hoc, use correlated step-size measurements of the same 106 geometries, and are not additional independent confirmations.

Table 6: Post hoc diagnostics of the saved one-step SGD contrasts. Each row contains two runs for each of 106 geometries. MAEs are in units of 10^{-6} MSE. Calibration is ordinary least squares of observed contrast on predicted contrast; sign accuracy excludes zero contrasts (none occur here). The four step sizes within a stage reuse the same trained models and are not independent tests.

Source step	SGD η	Zero MAE	Forecast MAE	Relative gain	Sign correct	Calibration slope
1000	10^{-5}	3.683	1.389	62.28%	194/212	0.875
1000	$3 \cdot 10^{-5}$	11.039	4.184	62.10%	194/212	0.870
1000	10^{-4}	36.664	14.143	61.43%	194/212	0.850
1000	$3 \cdot 10^{-4}$	108.863	44.231	59.37%	195/212	0.796
5000	10^{-5}	1.004	0.336	66.53%	192/212	0.956
5000	$3 \cdot 10^{-5}$	3.011	1.009	66.49%	192/212	0.955
5000	10^{-4}	9.974	3.363	66.28%	192/212	0.947
5000	$3 \cdot 10^{-4}$	29.370	10.217	65.21%	192/212	0.927

A 10,000-resample stratified bootstrap gave 95% intervals of [-0.01647, -0.00823] for A and [5.45e-6, 7.83e-6] for B. Excluding any one geometry left A between -0.01272 and -0.01170 and B between 6.39e-6 and 6.72e-6. These secondary checks preserve the two primary directions; they do not replace the frozen tests.

Among secondary comparators, the adapted strategy law had MAE 0.01535, the information comparator 0.02018 and the constant 0.02705. The strategy law's lower descriptive MAE is not a prespecified superiority result. Its reasonably accurate scalar forecast also does not imply that its two underlying predictors reproduce network responses. The mean best common circle/finite mixture approximation error at update 5000 was 0.01085 across the new geometries. This is a prediction-distance residual, not an additional test-risk component that can automatically be added to the network's MSE.

The right panel of Figure 16 reports the full checkpoint trajectory of this residual rather than only a thresholded pass/fail summary; the machine-readable supplement also gives every geometry, seed, and checkpoint value.

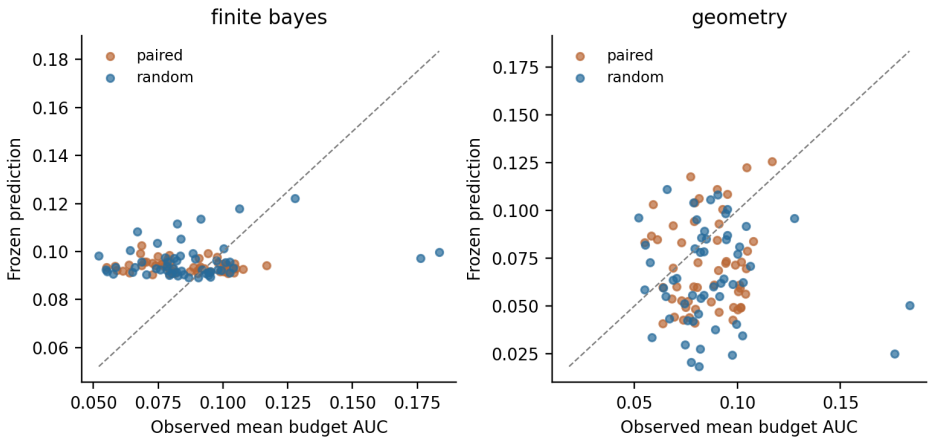

Figure 15: Frozen forecasts on 106 new geometries, averaging two training seeds within each geometry. The geometric predictor transfers worse than the finite-Bayes-risk baseline. Dashed lines show perfect prediction.

The earlier development audit makes the distinction especially clear. In the saved 5° states at update 30,000, the best circle/finite mixture had approximation MSE 0.04861. Fitting its weight on separate contexts and testing the transferred weight gave 0.05121. The covariance-matched ridge version had a similar optimal residual, 0.04876. Thus the common-weight dictionary misses substantial aspects of the answers even when its mixture coefficient is optimized. This limits the specified dictionary, not episode-dependent mixtures or Bayesian inference in general.

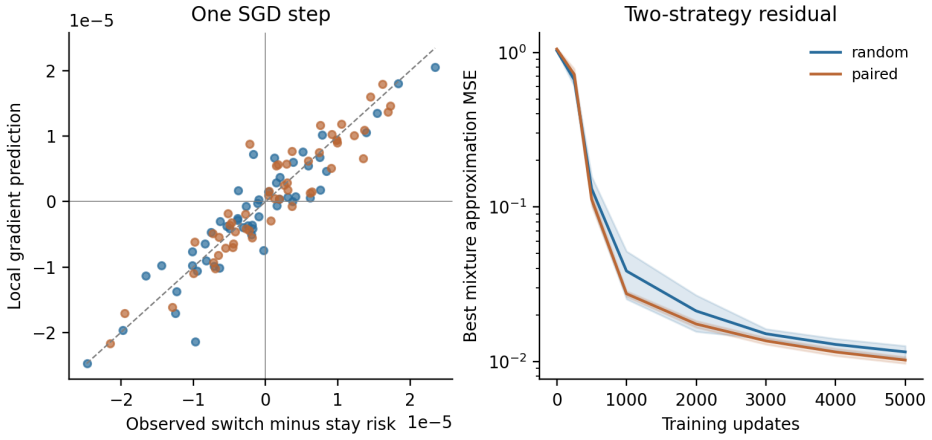

Figure 16: Local intervention predictions and the residual of the best two-strategy approximation. Left: geometry-averaged observed and predicted switch-minus-stay risk contrasts after one SGD update at step size 0.0001. Right: approximation error over training; bands are pointwise descriptive normal intervals across geometries.

At width 64, the same frozen geometric predictor also transferred worse than the finite-Bayes-risk baseline: improvement -0.02704, with descriptive 95% CI [-0.04223, -0.01184] over twelve geometries. The constant comparator had lower descriptive MAE than all other retained predictors at this width, indicating that even the stronger width-128 forecasts did not transfer reliably without adaptation. The local gradient improvement remained positive, 8.56e-6, with descriptive 95% CI [4.85e-6, 1.23e-5]. This small exploratory sample does not establish width-invariant performance or attainment of the practical reference. No target-width refitting was performed.

A complementary theoretical limitation is given in Appendix I. For any finite profile length m at least two, a sufficiently rich family contains noncongruent priors with equal information profiles but

different Bayesian prediction functions. The construction uses 4(m+2) tasks and does not establish unequal scalar risks or neural trajectories. It therefore restricts exact identification of the full strategy, while leaving approximate curve prediction as an empirical question.

E Population invariance derivation

Condition on a task vector w. Gaussian fourth moments give

$$\mathbb{E}_x[xx^\top (x^\top w)^2] = \|w\|^2 I + 2ww^\top.$$

Independent observation noise contributes $\sigma^2 I$ to the same-demonstration term. Cross-demonstration terms contribute $ww^\top$. Averaging over tasks with covariance C yields

$$S = \mathbb{E}[uu^\top] = \frac{m+1}{m} C + \frac{\operatorname{tr} C + \sigma^2}{m} I, \qquad \mathbb{E}[uw^\top] = C.$$

The clean-query population risk is $\operatorname{tr}(ASA^\top) - 2\operatorname{tr}(AC) + \operatorname{tr} C$. Its gradient is $2AS - 2C$, hence $A_* = CS^{-1}$. The identities are independent of higher task moments.

More generally, conditional on all inputs, any predictor linear in context labels has a squared-error loss quadratic in the joint response vector. Its expected risk therefore depends on the first two conditional moments. This fixed-predictor invariance does not constrain a nonlinear Transformer to the same class and does not imply identical finite-mini-batch optimization paths.

F Covariance direction is not covariance sufficiency

A recent softmax analysis expresses the fixed-query population attention readout as an exponentially tilted mean, or the gradient of a log moment-generating function (Xu & Fang, 2026, Theorem 3.2). Its elliptical-distribution result includes a query-dependent positive scalar multiplying the covariance direction (Xu & Fang, 2026, Theorem 3.4). Directional alignment should therefore not be read as a universal statement about response magnitude.

For clarity, our own elementary counterexample uses scalar keys and values with tilt q. Standard Gaussian X and equiprobable X∈{-1,+1} both have mean zero and variance one, but E[X exp(qX)]/E[exp(qX)] equals q in the first case and tanh(q) in the second. The covariance alone does not determine the exact softmax readout. This observation neither tests our trained model nor explains its trajectory. It clarifies why the empirical contribution of H1 must concern effect magnitude and learning dynamics under the specified controls, rather than the abstract possibility that nonlinear attention depends on higher moments.

G Angular harmonics of the population training objective

The following elementary identity sharpens the geometric description; it is our derivation for this experiment, not a measured neural mechanism. Write the task angles as $\alpha + j\pi/2 + b\theta$, with $j \in \{0,1,2,3\}$, $b \in \{0,1\}$, and common rotation α. Summing the four roots of unity gives

$$\mathbb{E}[e^{ik\phi}] = 0 \quad \text{if } 4 \nmid k, \qquad \mathbb{E}[e^{i4r\phi}] = e^{i4r\alpha} e^{i2r\theta} \cos(2r\theta).$$

For $\theta = \pi/4$, the nonzero modes are exactly the multiples of eight. Consequently, this prior matches uniform-circle averages of every polynomial in the task coordinates of total degree at most seven. The 5° prior matches degrees up to three, but has a nonzero fourth harmonic. These identities hold after any common rotation.

For a fixed set of network parameters, let $L(\phi)$ denote the expected training loss for task direction ϕ, averaging over inputs and observation noise. If its Fourier series is absolutely convergent, write $L(\phi) = \sum_k \widehat{L}_k e^{ik\phi}$. Its finite-prior population objective is then exactly

$$\mathbb{E}_{\mathcal{W}_{\theta,\alpha}}[L(\phi)] = \sum_{r \in \mathbb{Z}} \widehat{L}_{4r}\, e^{i4r\alpha} e^{i2r\theta} \cos(2r\theta),$$

whereas the continuous-prior objective is $\widehat{L}_0$. Thus the two finite priors retain different angular components of the task-dependent loss, despite equal covariance. For label-linear squared-error prediction, the loss is at most quadratic in the task vector and has no relevant fourth or higher harmonic, recovering the invariance in Appendix E.

Appendix J reports finite-grid measurements of angular loss and gradient components at selected H1 states. These diagnostics extend H1 but do not constitute interventions that isolate one Fourier mode. The identity alone neither predicts the sign of the trained-model contrast nor establishes which component caused an observed trajectory difference; optimization changes the loss function's Fourier coefficients over time.

H INDIVIDUAL CONFIRMATORY CONTRASTS

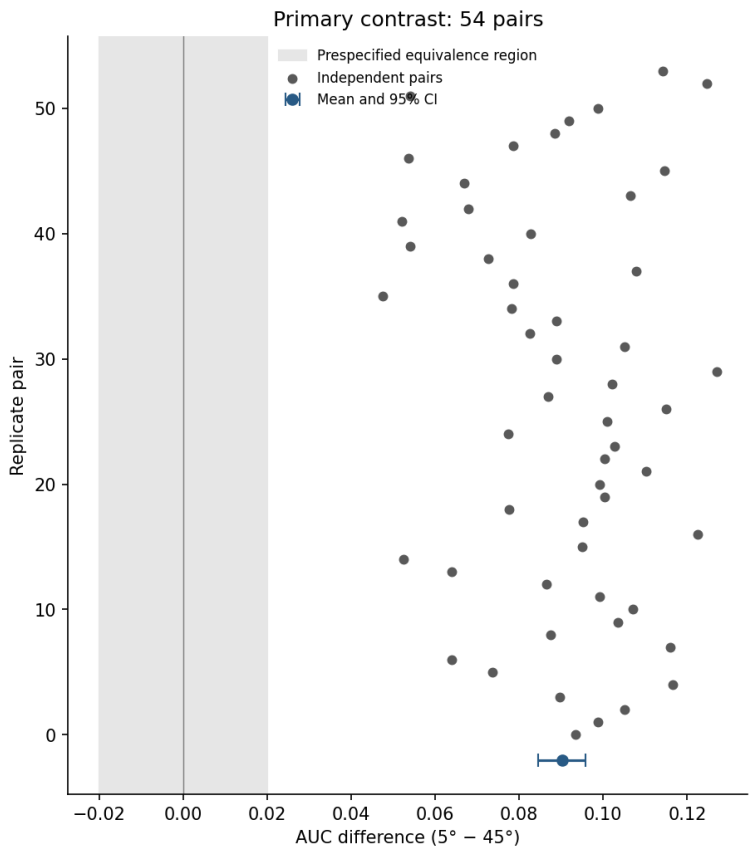

Figure 17: Every confirmatory paired AUC contrast and the mean with its 95% confidence interval. The shaded region marks the prespecified ± 0.02 equivalence margin.

I A FINITE INFORMATION PROFILE DOES NOT IDENTIFY THE FULL BAYESIAN PREDICTOR

Proposition. Fix an integer m at least two and positive observation noise. In two-dimensional Gaussian-design regression, there exist two uniform priors, each supported on 4(m+2) distinct unit vectors with mean zero and covariance I/2, that are not related by an orthogonal transformation, have identical values of $J_1, \ldots, J_m$, and yield different posterior-mean prediction functions on a positive-probability set under a common continuous-circle test law.

Proof. Set $L = m + 2$. Choose ordered orbit origins $0 = a_0 < a_1 < \cdots < a_{L-1} < \pi/2$, and let each origin generate its four quarter-turn rotations. Require the first angular gap to be the unique smallest gap and the closing gap to be the unique largest. These strict inequalities define a nonempty open set $U \subset \mathbb{R}^{L-1}$. For example, gaps proportional to $1, 2, \ldots, L-1, L+1$ satisfy them. Every resulting uniform prior has the stated moments and cardinality.

Different parameters in U are not related by rotation: the unique largest gap fixes its successor as the origin modulo a quarter turn. Reflection followed by this canonical alignment reverses the internal gap order, moving the unique smallest gap away from the first position. Thus no reflection maps one member of U to another.

Couple task selection by a common discrete index and use common Gaussian input and noise draws as the orbit origins vary. The finite-prior posterior mean is continuous in the origins and observations, and its norm is at most one. Dominated convergence makes each J_j continuous. Consequently,

$$F : U \subset \mathbb{R}^{m+1} \longrightarrow \mathbb{R}^m, \qquad F(P) = (J_1(P), \ldots, J_m(P))$$

is continuous. If it were injective, embedding its image in a coordinate hyperplane of $\mathbb{R}^{m+1}$ would give a continuous injection between equal-dimensional Euclidean spaces with a nonopen image. This contradicts invariance of domain (Brouwer, 1911). Hence two distinct, noncongruent priors share the profile exactly.

It remains to show that these priors define different prediction functions. Condition first on a design whose first two rows form the identity and whose remaining rows vanish. The informative observation is $z = w + \epsilon$. Writing f_P for its positive Gaussian-convolution density, differentiation under the finite mixture gives

$$\mathbb{E}_P[w \mid z] = z + \sigma^2 \nabla \log f_P(z).$$

This is the multivariate Gaussian posterior-mean identity associated with Tweedie's formula (Efron, 2011). Equality of the two posterior means everywhere would imply equality of the log-density gradients, hence equality of the normalized densities. Gaussian convolution is injective because its Fourier multiplier never vanishes. Therefore the priors would be identical, a contradiction.

Although the exact conditioned design has probability zero, posterior means are continuous in the design and labels. A nonzero difference persists on an open neighborhood with positive probability under Gaussian inputs and positive observation noise. An independent Gaussian query turns a nonzero difference in posterior means into a nonzero squared prediction distance. The distance is integrable because the posterior means have bounded norms. This proves the proposition.

The argument applies classical topological and Gaussian identities; no priority claim is made for those ingredients. At m=16, this construction uses 72 tasks, so it does not establish nonidentifiability within our sixteen-task experimental family. It gives neither a numerical effect-size bound nor a difference in scalar OOD risk: distinct predictors can have equal risks. It consequently does not refute approximate prediction of learning curves from an information profile. It excludes interpreting a finite profile as a universally identifying description of the entire Bayesian prediction function without restrictions on the prior family.

J EARLIER EXPLORATORY FOLLOW UP STUDIES

These studies were specified after the H1 result and before their own outcomes. They are exploratory. Reused checkpoints, branches and contexts do not increase the independent confirmatory sample. Protocols and code were committed before execution; this is not a claim of registration in an external registry.

J.1 CHECKPOINT DIAGNOSTICS

Twelve pairs, selected by the first twelve identifiers rather than outcome, were evaluated at steps 0, 1000, 3000, 10000, 20000 and 30000 on a new angular evaluation bank. Each bank contains 128 directions and 32 common Gaussian contexts per direction. Shared inputs and observation noise reduce directional comparison variance; these contexts are not treated as independent training replicates. We record clean risk, squared prediction differences from ridge and finite/circle posterior means, and Fourier amplitudes of the angular risk. A separate midpoint-label intervention tests affine dependence on context responses. Population-gradient Fourier norms were estimated on four pairs, at steps 1000 and 30000, using 32 directions and 64 common contexts. These finite-grid, parameterization-dependent norms do not establish a causal mechanism.

Under the 5° prior, mean clean risk increased from 0.05548 to 0.11712 between steps 1000 and 30000. Squared prediction distance to ridge increased from 0.05258 to 0.11636, whereas distance to the eight-function posterior decreased from 0.16075 to 0.07315. The fourth risk harmonic increased from 0.01634 to 0.06275. The paired between-condition difference in risk changes was 0.08197 (descriptive 95% CI 0.06260–0.10133). These observations support behavioral specialization while neither equating the trained predictor with the finite posterior nor identifying an internal Bayesian algorithm.

At 45°, final model risk 0.00904 coexists with squared prediction distance 0.00796 from ridge. Similar mean risks therefore do not imply similar pointwise predictions. The midpoint intervention also departs from exact label-affine behavior. Because the intervention changes the latent task/noise law, its defect is an algorithmic response diagnostic rather than another in-distribution generalization risk.

J.2 SWITCHING THE CONTINUATION PRIOR

The complete intervention, source-pair count, effect estimates, and interpretation are reported in the continuation intervention in the main text.

J.3 TASK INFORMATION AND RESTRICTED ALGORITHM TESTS

Four independent H1 pairs were tested at steps 1000 and 30000 under both priors. A recipient context has its demonstration responses permuted. A donor has independent demonstrations of the same latent function and the same query input. A wrong-task donor provides a matched intervention control. Each of 16 attention-head outputs, immediately before its output projection, is patched at the final query position. For each trained model separately, 128 selection contexts choose one head; another 128 contexts evaluate that choice. Head indices are not aligned across independently initialized networks.

Held-out same-task patches reduce recipient MSE in all four prior/checkpoint groups on average. Mean reductions are 0.76218 and 0.49576 at 5° for steps 1000 and 30000, and 0.79471 and 0.33089 at 45°. Wrong-task patches are worse on average. These are raw risks in heavily corrupted contexts and must not be compared numerically with the H1 normalized AUC effect.

The intervention establishes causally used task-specific information at the selected location. Its presence after 5° generalization deterioration argues against the simple claim that all useful task information disappeared. Since donor and recipient queries are identical, the transferred activation may encode query-specific computation; the experiment does not establish a query-independent function vector or a unique internal optimizer. These interventions address the distinction raised by function-vector studies (Todd et al., 2024); they do not replicate their language-task protocols.

Numeric regression does not instantiate the canonical repeated-token induction benchmark, so these heads are not identified as induction heads.

For fixed inputs X, fixed initialization and learning rates independent of the response vector y, finite-step gradient descent on squared regression loss remains affine in y. The update

$$u_{t+1} = (I - \eta_t X^\top X)u_t + \eta_t X^\top y$$

preserves this property by induction. Such predictors obey the midpoint identity, as does a fixed Gaussian-prior posterior mean. A nonzero midpoint defect excludes exact global identity with that restricted class. It does not exclude adaptive learning rates, label-dependent stopping or initialization, nonlinear feature learning, constrained inference, or other learned optimizers. With nonzero Gaussian observation noise, describing an intervention as outside the task family concerns its latent generating law, not zero support under the original observation distribution.

To distinguish query-specific task information from a reusable function representation, we extracted residual activations at the last demonstration-label position from a 32-token prefix containing no query. Each donor activation was reused across four independent new queries. For each of 16 model states, 128 functions fitted a centered ridge probe from hidden activations to the two-dimensional task vector; the probe row space defined a rank-two projection. A separate 64-function set selected the layer by rescue, and another 64 functions tested that selection. Controls included full residual replacement, half-dose transfer, wrong-task and random rank-two displacements, complement displacements, and deletion from clean contexts. Control displacements were norm matched. Prefix invariance and the expected null effect of final-layer prefix replacement were validated.

The selected rank-two transfer changed corrupted-context MSE only by approximately 0.00003 or less in the four prior/checkpoint group means. Task-subspace deletion also had negligible effects. Full residual replacement provided partial rescue. Thus, this particular linearly decoded, query-free bottleneck was not supported, despite successful query-position head transfer. Weak linear decodability, distributed computation, nonlinear representations and other token positions remain alternatives; this is not evidence against all function vectors or all ICL mechanisms.

J.4 Geometric descriptors and held out prediction

To test the sufficiency of the fourth harmonic itself, the subsequent exploratory suite uses sixteen unit vectors formed by four crosses with offsets a,-a,b,-b and common rotation. We use a = 5° or 10° and choose b so that cos(4 a)+cos(4 b)=1. Before the common rotation,

$$\mathbb{E}[e^{4i\phi}] = \frac{\cos(4a) + \cos(4b)}{2} = \frac{1}{2}.$$

Writing A=cos(4 a), the eighth harmonic is

$$\mathbb{E}[e^{8i\phi}] = A^2 + (1 - A)^2 - 1 = -2A(1 - A).$$

Changing a therefore varies higher structure while preserving cardinality, probabilities, norms, mean, covariance and the complex fourth harmonic under a common orientation. This elementary construction does not isolate the eighth harmonic as the only altered property. The following tests report its completed exploratory evaluation.

The exploratory suite comprised 28 configurations across four independent replicate blocks, totaling 112 new trainings. Unless otherwise specified, each used 5000 updates, batch size 256, hidden width 128, four blocks, 16 demonstrations and noise standard deviation 0.25. Other optimizer settings match H1, with a 250-update warmup. Each checkpoint uses 2048 evaluation episodes, and the independent final bank uses 4096. Intermediate geometries have separate training and evaluation random streams; common initial weights and orientation within each replicate block are retained where dimensions match. This shorter experiment does not replace the 30000-update confirmation. All 112 runs passed checks of regenerated targets and references, risk reconstruction, checkpoint coverage and prior moments. Fifty-six two-update suite smoke runs were excluded from scientific counts.

For the intermediate-angle family, four geometries (10°, 20°, 30°, 40°) fitted fixed linear predictors and two (15°, 35°) were held out. A scalar fourth-harmonic predictor achieved absolute errors of 0.000715 and 0.000294 on final independent-bank mean MSE; a predictor using finite-posterior risk on the common continuous test distribution achieved 0.001482 and 0.000841. This is interpolation within one controlled family, not evidence that either descriptor is sufficient across arbitrary priors.

For the two sixteen-task priors with matching fourth harmonic, the mean final-risk difference was 0.000946 (descriptive 95% CI -0.002915 to 0.004807). No difference was resolved at this budget and sample size. A non-significant difference does not establish equivalence or sufficiency of the matched harmonic.

We additionally estimated the sixteen-value profile $J_m = \mathbb{E}\|\mathbb{E}[w \mid D_m]\|^2$ using 32768 windows per prior. With unit-norm tasks and isotropic Gaussian queries, J_m is the reduction in finite-prior Bayes prediction risk relative to the zero predictor. It is not mutual information. For 5° and 45°, J_1 was 0.61066 and 0.48370, respectively; easier early inference under the training prior therefore need not imply better generalization to new circle tasks.

A fixed ridge predictor using this profile, fitted only on the four training geometries with regularization 0.001, predicted the 20 noninitial checkpoint-risk means at the two held-out angles. Mean absolute curve errors were 0.007717 and 0.009601, versus 0.007970 and 0.012026 for a constant covariance-only predictor. The scalar finite-posterior-risk predictor gave 0.007257 and 0.009732. The profile was not uniformly better than that scalar baseline. Two held-out geometries cannot establish a general ranking or sufficiency theorem. The curve analysis uses checkpoint evaluation banks; the separate final-risk analysis uses independent final banks. Their endpoint errors are different estimands.

J.5 Optimization dimension and scale

Hidden widths 32, 64, 128 and 256 correspond to 52161, 202625, 798465 and 3169793 parameters. Mean final risks at 45° were 0.14146, 0.01515, 0.01303 and 0.01459; at 5° they were 0.19678, 0.07205, 0.07285 and 0.11132. The improvement from width 32 to 64 does not identify a sharp capacity boundary: the grid is sparse, training duration is fixed, and geometry continues to matter at larger widths. The width-256 contrast was 0.09673 (descriptive 95% CI 0.05094–0.14252); intervals for the other three width contrasts crossed zero with only four pairs.

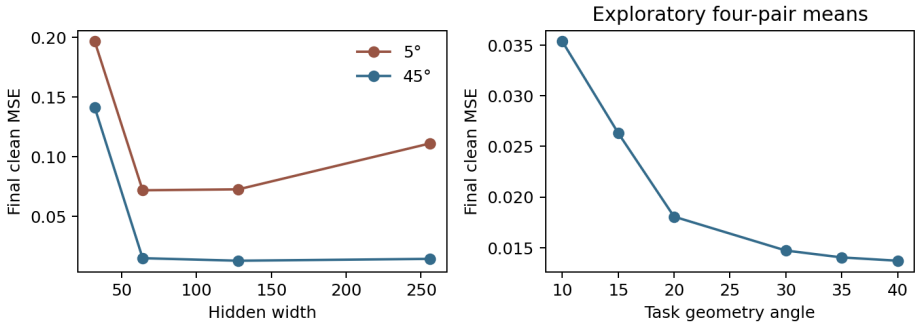

Figure 18: Exploratory mean final risks across width and intermediate task geometries. Each mean uses four training replicates. Uncertainty and the restricted scope are reported in the text.

At 5000 updates, the 5° minus 45° risk contrast was 0.46606 for batch 64 and 0.08354 for batch 1024. Holding the number of training windows equal to batch 256 times 5000, the contrast was 0.39171 for batch 64 times 20000 updates and 0.05506 for batch 1024 times 1250 updates. These are not pure gradient-noise interventions: update counts, data exposure and the fraction of the fixed 250-update warmup differ. Differences between contrast magnitudes are not themselves a tested statistical interaction.

A four-dimensional linear family and a two-dimensional quadratic family were also tested. The linear prior comprises the signed coordinate basis and its block rotation; the quadratic task is linear in features $(x_j^2 - 1)/\sqrt{2}$, while the network receives raw x. Mean prior contrasts were 0.13311 (de-

scriptive 95% CI -0.63889 to 0.90511) and -0.03788 (-0.70323 to 0.62747). These outcomes were too heterogeneous at the fixed budget to establish robust transfer, and the quadratic mean's negative sign is not reliable evidence of reversal. All exploratory intervals are unadjusted for multiple comparisons.

J.6 Pretrained language model interface

Two public Qwen2.5 base models (Qwen Team, 2024a;b), 0.5B and 1.5B, were evaluated at immutable repository revisions using BF16 greedy decoding and at most 12 generated tokens. Each received 832 fixed episodes: 32 continuous-circle functions times four contexts, two formats and three context conditions, plus 64 copy controls. The formats were numerical rows and sentences. Conditions were no demonstrations, 16 correct demonstrations and 16 demonstrations with permuted responses. Inputs were rounded before target generation. The first parsed number was scored by squared error capped at 16, with invalid outputs assigned 16. A complete independent audit regenerated data and verified every prompt, parsed output and loss.

The prespecified interface criterion required correct-context risk at most 75% of zero-predictor risk and a positive lower task-bootstrap confidence bound for the shuffled-minus-correct risk difference in both formats. This operational criterion was fixed before inference; it is not a general definition of ICL. The task, rather than the individual response, was the bootstrap unit (32 tasks, 2000 resamples).

Table 7: Correct-context capped MSE and shuffled-minus-correct differences. The zero predictor has risk 1.11580 and ridge using the displayed examples has risk 0.01017.

Model	Format	Correct MSE	Shuffled minus correct
Qwen 0.5B	Rows	1.11495	-0.00122
Qwen 0.5B	Sentences	1.23197	0.30553
Qwen 1.5B	Rows	0.88946	0.25038
Qwen 1.5B	Sentences	0.84927	0.46442

Numeric coverage is the fraction of outputs from which the parser extracts a finite number; it does not measure answer correctness or strict output-format compliance. Coverage was 1.000 in every model–format–condition cell. Valid-only uncapped MSE equaled capped MSE except for the 0.5B sentence clean cell, where it was 1.24084 rather than 1.23197; the complete capped and uncapped rows are retained in the audited `H2_TEXT` summaries included in the supplement. For 1.5B, shuffled-minus-correct differences had descriptive 95% intervals 0.06287–0.44849 and 0.21922–0.71927, showing a context effect. Nevertheless, neither model passed the full prespecified criterion. The simple copy task $y = x_1$ had zero MSE for 1.5B in sentences but MSE 1.12769 in rows, revealing substantial format dependence even for the control.

Accordingly, the conditional continued-pretraining stage was not activated. The experiment does not establish causal transfer of the H1 prior-geometry effect to pretrained language models. It also does not demonstrate absence of ICL in these models. The zero numerical predictor used for the gate is distinct from the models' zero-shot outputs. Exact revisions, versions, prompts and continuations accompany the returned data.

J.7 Fresh sentence-format prompting test

After the historical two-format gate above failed jointly, we froze a new sentence-only protocol before inspecting its outcomes. We evaluated `Qwen3-4B-Base` (Qwen Team, 2025) at immutable revision `906bfd4b4dc7f14ee4320094d8b41684abff8539`. A fresh generator sampled 64 independent two-dimensional linear tasks, eight contexts per task, and four paired conditions: correct demonstrations, shuffled responses, no demonstrations, and demonstrations from an orthogonal task. The same query and latent task were shared across conditions. Greedy BF16 decoding used a maximum of 12 new tokens. The first finite parsed number was scored by squared error capped at 16; invalid generations received loss 16.

The frozen joint decision required at least 95% numeric coverage in the clean condition, clean capped MSE at most 75% of no-context MSE, and positive Bonferroni-adjusted one-sided lower

task-bootstrap bounds for all three control-minus-clean contrasts. The task was the statistical unit; 50,000 deterministic resamples were used over 64 tasks. An independent implementation regenerated every task, prompt, tokenization, parsed value and loss without importing the inference generator. All 2,048 rows passed this audit.

Clean capped MSE was 0.47455, compared with 1.11773 for shuffled demonstrations, 1.30669 with no context, and 1.67922 for irrelevant demonstrations. Numeric coverage, defined as in the historical two-format test above, was 1.000 in all four conditions. The corresponding valid-only uncapped MSE values were 0.47455, 1.11853, 1.30710, and 1.74585; reporting both versions makes the effect of the cap explicit. Clean/no-context MSE ratio was 0.3632. Mean task-level benefits were 0.64318 for shuffled minus clean (95% bootstrap interval [0.50579, 0.79464]; adjusted lower bound 0.49467), 0.83214 for no-context minus clean ([0.66779, 1.01269]; lower 0.65473), and 1.20467 for irrelevant minus clean ([0.95153, 1.48343]; lower 0.93245). The frozen decision was positive.

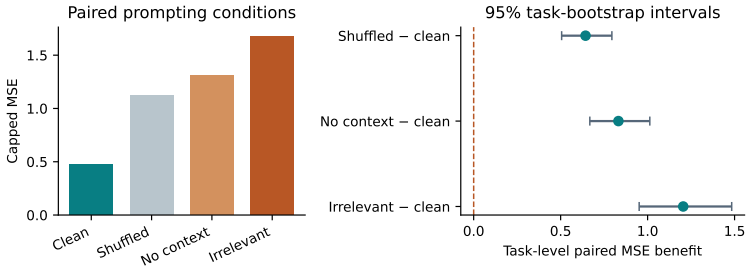

Figure 19: Fresh prompting test in one pretrained base model. Left: capped MSE for four paired conditions. Right: task-level control-minus-clean benefits with 95% bootstrap intervals. Correctly paired sentence demonstrations improve prediction, but this test does not reproduce the synthetic training-geometry intervention.

This is causal evidence that the selected pretrained model uses correctly paired sentence demonstrations on this known numerical generator. It is one model and one selected interface; the failed table-format result above remains part of the record. The test neither adapts the model nor shows that language pretraining learned the synthetic geometry effect, and it does not measure diversity of latent tasks in natural text.

A finite projected-language construction. The native 28 GoEmotions classification functions (Demszky et al., 2020) have full column rank on the corpus, so matching their complete mean response identifies the mixture weights. We therefore froze a narrower four-coordinate functional sketch before constructing new priors: label prevalence and three deterministic signed response averages on 8,192 cleaned real comments. Normalization, four projected means and ten projected second moments give a rank-15 constraint matrix over 28 task weights. A nullspace perturbation of the uniform prior produced two valid mixtures with total variation 0.24175. Projected first- and second-moment differences were below 7×10^{-19}, while projected third moments differed by 1.19×10^{-5}. Their full mixture response functions differed on held-out dev and test documents with RMS 0.02272 and 0.02246. An independent implementation redownloaded the exact source files and reproduced the result. This establishes non-identifiability of a specified finite projection, not matched full response laws or different neural behavior; no language model was trained on these priors.

J.8 Identifiability of tasks in text

Identifiability results for latent mixtures and hidden Markov models (Allman et al., 2009; Gassiat et al., 2013) motivate the following restrictions. A text-level extension requires a specified observation law and task semantics. Without restrictions, a one-component law P and an arbitrary mixture whose every component equals P are observationally identical. Latent component count is consequently not a corpus statistic in the unrestricted model. Identifiability results for conditional-independence mixtures or HMMs impose substantive structural and rank assumptions; they do not assign semantic task meanings automatically.

An operational alternative defines tasks through their conditional response laws on a fixed probe-input distribution Q. For numerical means, squared functional distance integrates the squared difference of conditional means over Q. This distance is unchanged by relabeling tasks or reparameterizing their latent coordinates, unlike the covariance of arbitrary text embeddings. Finite probes identify only a projection of the conditional law. Our pretrained-model test therefore begins with a known numerical generator encoded in two fixed textual formats; it does not recover the diversity of tasks in an unknown pretraining corpus.