SEPARATING SUPERVISION RANDOMNESS IN PRIOR-DATA FITTED NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior-data fitted networks (PFNs) learn predictive distributions from simulated tasks. At a fixed training budget, removing supervision randomness can improve source-prior fidelity while worsening prediction under prior shift, despite preserving the expected training objective. A four-arm intervention in a finite structural equation model separates additive component-target variation from within-component label randomness. Across five paired training seeds within one source prior, replacing posterior-resampled labels with component targets reduces source predictive KL by 0.00975 nats per query but increases coefficient-shift KL by 0.00413 at $\lambda = 0.25$. A signed loss estimator retains within-component label randomness while removing additive component-target variation and reproduces the reversal. At the same interpolation level, including the residual instead increases predictive KL under covariance-family shift. A separate six-prior, two-arm study shows that the pooled supervision ranking changes with shift magnitude and construction: full-predictive supervision wins at the smaller tested shift and sampled supervision at the larger one in both families. Source diagnostics show that residuals dominate native target variance. Source-selected output softening remains competitive in aggregate risk while moving predictions farther from the residual-trained predictor on source assessment inputs. These findings establish supervision-estimator choice as a consequential variable in finite-budget PFN training, requiring joint evaluation of source fidelity, shifted risk and predictive behavior.

1 INTRODUCTION

Prior-data fitted networks (PFNs) learn to predict from a context dataset by training over simulated tasks (Müller et al., 2022). This approach connects amortized Bayesian inference to the practical success of tabular foundation models (Hollmann et al., 2023; 2025). Their supervision can take the form of sampled labels or predictive probability targets. When the probability target is the conditional expectation of the sampled label, these choices preserve the expected cross-entropy objective while changing the stochastic training trajectory. When deployment tasks depart from the training prior, how does that choice affect source fidelity and shifted prediction?

Prior work shows that departures from a pretraining-prior Bayes rule can improve prediction in controlled regression (Raventós et al., 2023), that true probability targets can reduce variance (Menon et al., 2021), and that label-noise SGD can induce implicit regularization under specified assumptions (Damian et al., 2021). We ask which components of correctly sampled supervision change the source–shift trade-off when the expected training objective is held fixed. Sampled labels combine variation over a latent component with label randomness within that component. Replacing sampled labels with the full posterior predictive probability vector removes both sources of randomness at once and cannot distinguish their contributions.

We address this question with a four-arm intervention in a finite structural equation model, where the full posterior predictive Q and component-conditional target q_Z are numerically accessible. Writing the one-hot label vector as $L = Q + A + R$ separates component-target variation $A = q_Z - Q$ from the label residual $R = L - q_Z$. The four targets are Q, q_Z, $Q + R$ (Signed), and L (posterior-resampled Fresh). Signed supplies an experimental counterfactual: retaining the native residual while removing additive component-target variation. The construction yields two paired residual contrasts with the same expected objective.

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 a controlled test of which components of supervision randomness drive the source–shift trade-off. A four-arm study with five paired training seeds identifies the native label residual as sufficient to reproduce the reversal within one source prior. A separate six-prior study shows that the preferred supervision changes with the shift. Source-selected controls further distinguish competitive aggregate risk from reproducing the residual-trained predictive distribution. Supervision-switching and broader-prior experiments assess the scope of these findings.

2 Related work

PFNs and learned Bayesian prediction. Müller et al. (2022) introduce PFNs as predictors trained over a prior distribution of supervised-learning tasks. TabPFN demonstrates their effectiveness on small tabular classification problems (Hollmann et al., 2023), and later tabular foundation models extend their practical reach (Hollmann et al., 2025; Qu et al., 2025). Nagler (2023) provides a frequentist account of PFN prediction; here we study how the stochastic supervision used during pretraining affects finite-budget predictive behavior. Our controlled networks use a much smaller finite prior. Enumerating its components gives numerical access to the expected target and enables selective supervision interventions. Such access is not automatically available for a released foundation model.

Probability targets and stochastic optimization. Probability targets are central to distillation (Hinton et al., 2015); Menon et al. (2021) analyze the statistical benefits of true class probabilities and the consequences of approximate targets. Label randomness can also alter optimization beyond the expected objective. Blanc et al. (2020) study implicit regularization near zero-error solutions under squared loss; Damian et al. (2021) establish a preference of label-noise SGD for flatter minimizers under their assumptions. Gradient-noise geometry can matter as well (Zhu et al., 2019). These results motivate distinguishing the objective from its stochastic estimator. They do not directly characterize the categorical-loss Adam trajectories studied here.

Prior mismatch and useful approximation error. Raventós et al. (2023) show that sufficiently diverse pretraining in controlled regression produces behavior closer to ridge regression, which can outperform the finite-pretraining-prior Bayesian estimator on new tasks. Grünwald & van Ommen (2017) separately demonstrate limits of Bayesian prediction under misspecified linear models, a setting distinct from our task-prior shifts. Our contribution is the equal-expectation component/residual intervention. The output-softening control uses temperature scaling, an established confidence transformation (Guo et al., 2017), with a source-risk matching rule rather than conventional likelihood-based calibration.

3 Setting and supervision interventions

3.1 A finite predictive reference

Each task is generated by one of 48 latent components: eight covariance atoms crossed with six variable orders in a three-variable linear structural equation model. For each atom, order-specific Cholesky parameterizations preserve its population covariance. Independent innovations mix Gaussian and centered asymmetric-Laplace distributions with matched variances and mixture weights 0.25 and 0.75, respectively. Orders can consequently share covariance while differing in higher-order properties. The construction provides a tractable prediction problem with latent uncertainty; recovery of its causal structure is not required for the comparisons below.

Let $H = (D, x)$, where D contains 20 complete three-variable context observations and x contains the first two coordinates of a query. The third query variable is the prediction target. For component z and target bin I_b, define

$$q_z(b \mid H) = P(Y \in I_b \mid z, x), \qquad Q_b(H) = \sum_z P(z \mid D, x) q_z(b \mid H). \tag{1}$$

Given z, the query is independent of D, so the context affects prediction through the posterior component weights. These weights condition on both the context and the query covariates. We enumerate all components and evaluate the probabilities over 100 bins by numerical quadrature. Throughout the paper, Q denotes this numerical categorical predictive reference. Targets and trained

predictors are distinct: p_Q, p_{q_Z}, p_{Signed} and p_{Fresh} denote the four trained predictors. The prior, binning and integration details are given in Appendix B.

3.2 A FACTORIAL SUPERVISION INTERVENTION

At a fixed input H, draw $Z \sim P(z \mid H)$ and a one-hot target L from q_Z. Decompose its deviation from Q as

$$A = q_Z - Q, \qquad R = L - q_Z, \qquad T_{c\ell} = Q + cA + \ell R, \quad c, \ell \in \{0, 1\}. \tag{2}$$

Here A is component-target variation and R is the label residual. Figure 1 shows the four combinations. Replacing Fresh with q_Z removes only R from one-hot supervision. Replacing Q with Signed restores R while excluding additive A. Both residual-present arms still sample Z to construct R.

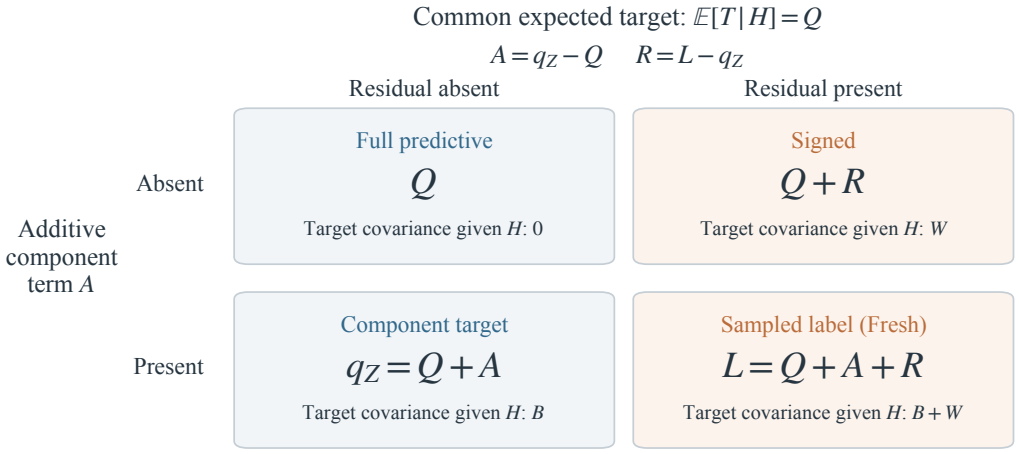


Figure 1. The four-arm factorial. Columns toggle the label residual $R = L - q_Z$; rows toggle component-target variation $A = q_Z - Q$. Every cell has conditional expected target Q. Target covariance is conditional on H. Signed removes additive A while retaining component sampling for R; it is a signed loss estimator with potentially negative target entries. A and R are conditionally uncorrelated, but need not be independent.

All four ideal target laws have conditional expectation Q, because

$$\mathbb{E}[A \mid H] = 0, \qquad \mathbb{E}[R \mid Z, H] = 0, \qquad \mathbb{E}[AR^\top \mid H] = 0. \tag{3}$$

For Q, q_Z, Signed and Fresh, the conditional covariances are 0, B, W and $B + W$, where

$$B = \mathrm{Cov}(q_Z \mid H), \qquad W = \mathbb{E}[\mathrm{diag}(q_Z) - q_Z q_Z^\top \mid H]. \tag{4}$$

Conditional uncorrelatedness follows from the zero conditional mean of R; its distribution still depends on Z. The decomposition follows from the law of total covariance; our contribution is the selective intervention it enables.

3.3 THE SIGNED LOSS AND COMMON EXPECTED OBJECTIVE

For logits f and probabilities $p = \mathrm{softmax}(f)$, the signed arm has loss

$$\mathcal{L}_{\text{Signed}} = -\langle Q + R, \log p \rangle = \mathcal{L}_Q - \langle R, f \rangle. \tag{5}$$

The equality follows from $\sum_b R_b = 0$. We implement it as the corresponding linear combination of cross-entropies, without clamping or renormalizing the target. The correction has zero conditional mean at fixed logits. Individual signed targets need not be probability vectors. For a fixed realized signed target, the loss as a function of the logits can be negative and unbounded below, although its conditional expectation is ordinary cross-entropy. In practice, all five Signed runs completed 200,000 updates, with no stability-triggered stops or excluded endpoints (Appendix B.4).

At fixed parameters and inputs, with logit Jacobian J, the target-induced gradient and its covariance satisfy

$$\nabla_\theta \mathcal{L}_T = J^\top (p - T), \qquad \mathrm{Cov}(\nabla_\theta \mathcal{L}_T \mid H) = J^\top \mathrm{Cov}(T \mid H) J. \tag{6}$$

Equal expected objectives do not imply identical realized gradients, Adam updates, or finite-budget training trajectories. Appendix A gives the derivation and finite-precision qualifications.

4 Experiments

4.1 Experimental setup

Training and evaluation. We train the four arms for 200,000 Adam updates with batch size 32, using a two-layer transformer with width 128. For each of five training seeds, initialization and input traversal are paired across arms; the seed also indexes the component and label draws. Each episode contains 20 context observations and seven queries, with one query processed per forward pass. All seven query losses contribute to the batch mean. The input bank is fixed, while separate target draws are stored for each visit and are never recycled. The sampled arm is *posterior-resampled Fresh*, abbreviated Fresh in figures and tables. Components are drawn independently for each query and visit. The component and label draws are paired across the q_Z, Signed and Fresh arms wherever those draws are used. Fresh preserves each query label's conditional distribution, but not the cross-query label dependence induced by a shared simulator component. Appendix B specifies the update. Two replacement target banks test target-draw sensitivity with one optimization seed and the same Q model.

The factorial uses one source prior. Two environment/context blocks supply discovery data and six withheld blocks supply the main evaluation. Each block independently draws both shift endpoints and an episode sample with 4,096 contexts and seven queries per context. Within a block, common random numbers pair latent-component and residual draws across the source and shifted regimes. We evaluate the same trained models on the source law and on two shift constructions. *Coefficient shift* interpolates structural coefficients toward independently drawn endpoints, holding innovation scales fixed. The *covariance-family shift* interpolates covariance parameters and reconstructs both coefficients and innovation scales. Changing coefficients can itself change covariance: these names identify constructions, not disjoint kinds of distribution change. The primary level is $\lambda = 0.25$, but the same interpolation level does not imply equal shift severity across the two constructions.

Predictive risk and uncertainty. Each regime e is scored against its own numerical predictive reference:

$$\mathcal{R}_e(p) = \mathbb{E}_{H \sim P_e}\big[D_{\mathrm{KL}}(Q_e(H) \| p(H))\big]. \tag{7}$$

KL measures excess expected logarithmic loss (Gneiting & Raftery, 2007), in nats per query. Factorial intervals use 2,000 paired bootstrap draws over training seeds and environment blocks, conditional on within-block contexts. Direct-effect intervals preserve arm pairing and are nominal 95% intervals, without a familywise adjustment across all contrasts. Numerical sensitivity uses two integration grids and the continuous ranked probability score (CRPS) for a specified reconstruction of the binned distribution. Table 4 distinguishes the resampling units, selection rules and interval adjustments across studies. Full training details appear in Appendix B.

4.2 Main finding: the four-arm residual intervention

The two residual interventions produce the same source/coefficient-shift reversal (Table 1, Figure 2). Replacing posterior-resampled Fresh labels by q_Z improves source fidelity but worsens coefficient-shift prediction. The signed intervention reproduces both directions: p_{Signed} has higher source KL and lower coefficient-shift KL than p_Q after additive component-target variation is removed.

Table 1. Predictive risk of the four trained predictors on six withheld blocks, grouped by supervision target. Values are mean KL in 10^{-3} nats per query over five paired training seeds; lower is better. Shifted risks use $\lambda = 0.25$. The component and residual columns identify the two target factors. Paired residual-effect intervals are shown in Figure 2.

Supervision	Component	Residual	Source	Coefficient shift	Covariance family
Full predictive (Q)	0	0	12.538	142.475	55.540
Component target (q_Z)	1	0	13.570	141.883	55.647
Signed ($Q + R$)	0	1	23.337	137.822	58.608
Fresh (L)	1	1	23.324	137.753	59.041

With component-target variation present, removing R reduces source KL by 0.009754 nats/query but increases coefficient-shift KL by 0.004131, with 95% interval $[0.002802, 0.005383]$. With component-target variation removed, adding the residual reduces coefficient-shift KL by 0.004653 nats/query. This agreement localizes the observed trade-off to inclusion of the label residual at its native amplitude in this prior and training recipe. The residual increases risk in both covariance-family comparisons, so its effect depends on the evaluation law.

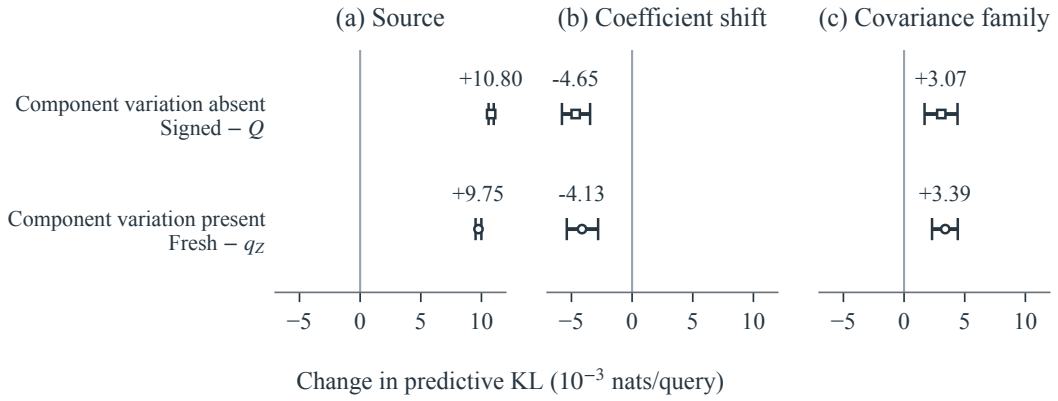


Figure 2. Adding the label residual reverses the source/coefficient-shift ordering. Component-target variation is absent in the top row ($\mathcal{R}_e(p_{\mathrm{Signed}}) - \mathcal{R}_e(p_Q)$) and present in the bottom row ($\mathcal{R}_e(p_{\mathrm{Fresh}}) - \mathcal{R}_e(p_{q_Z})$). Points and capped lines show paired mean differences and nominal 95% bootstrap intervals over five training seeds and six withheld blocks in one prior. All panels share an axis and zero line; negative values favor residual-present supervision. The two shifted panels use $\lambda = 0.25$.

Component-target effects are smaller at these native amplitudes. Under coefficient shift, adding A changes KL by -0.000592 without R and -0.000069 with it; both intervals include zero. On the source law, adding A increases KL when R is absent; its effect is unresolved when R is present. The interaction, defined as the residual effect with A present minus that with A absent, is -0.001045 on the source and $+0.000522$ under coefficient shift; only the source interval excludes zero. Appendix C reports interactions, direct contrasts and all seed effects.

Sensitivity to environments and target draws. The discovery and withheld cohorts agree in direction, and both replacement target banks reproduce Fresh's coefficient-shift advantage over q_Z, by 0.002612 and 0.003189 nats. These banks share inputs, the Q anchor, the optimization seed and evaluation blocks; they test target-draw sensitivity within the same prior. On both numerical grids and under reconstructed CRPS, both residual contrasts are positive on the source and covariance-family laws and negative under coefficient shift. Appendix C identifies the numerical summaries. These checks support the within-prior intervention effect across evaluation blocks, target draws and scoring choices.

4.3 Across six priors, the preferred supervision depends on the shift

The pooled ranking of sampled and full-predictive supervision changes with both shift magnitude and construction across six independently drawn source priors (Figure 3). This separate, earlier study uses four paired training seeds per prior. Changing this sampled supervision to Q removes both A and R, measuring their combined effect across priors; the factorial isolates the residual contrasts within one prior. Appendix F specifies the sampled target law.

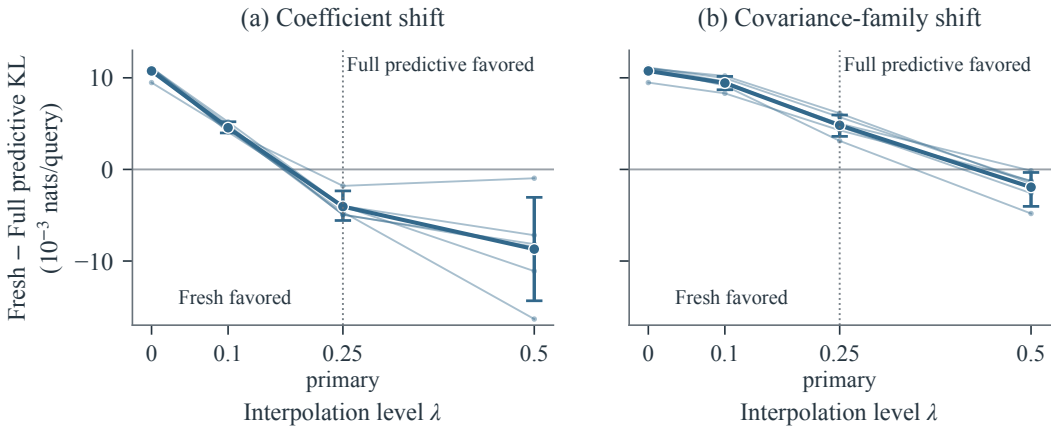


Figure 3. Broader two-arm evidence across six independent source priors: Fresh minus full-predictive KL, in 10^{-3} nats/query. Negative favors Fresh, as in Figure 2. Thin lines show the six prior means; bold points average them. Capped intervals at nonzero levels are 99.1667% hierarchical bootstrap intervals (Bonferroni family size six; nominal familywise 95%). The primary level is $\lambda = 0.25$; the common source point is descriptive. Lines only connect evaluated levels. Equal λ does not equate severity across families.

Using Fresh minus full-predictive risk throughout, the source difference is $+0.010748$ nats/query. At $\lambda = 0.25$, the coefficient-shift difference is -0.004054, with 97.5% hierarchical interval $[-0.005363, -0.002631]$. The covariance-family difference is $+0.004817$, with interval $[0.003824, 0.005804]$. These two co-primary intervals adjust for two comparisons; Figure 3 uses the wider six-comparison adjustment. The prior is the top-level resampling unit, and all six prior means share the corresponding signs.

The pooled ordering changes across the tested shift magnitudes in both families. At $\lambda = 0.1$, it favors full-predictive supervision; at $\lambda = 0.5$, it favors Fresh. All six nonzero-level pooled intervals in Figure 3 exclude zero after the six-comparison adjustment. Thus the broader finding is a change in the preferred supervision across tested shifts, extending beyond a single coefficient-shift advantage. Equal λ does not equate predictive mismatch between families, so the contrast at $\lambda = 0.25$ does not by itself isolate shift geometry.

4.4 Native noise scale and predictive alternatives

Within-component label residuals dominate native target variance. A source-only descriptive diagnostic on 256 assessment contexts and all seven queries gives mean $\operatorname{tr} B = 0.002074$ and $\operatorname{tr} W = 0.911261$, a ratio of means of 439.4. Component-target variation accounts for just 0.22707% of aggregate target variance. Thus removing the residual removes nearly all native conditional target variance on this panel, whereas removing additive component-target variation preserves nearly all of it. Empirically, the latter intervention also preserves the source/coefficient-shift reversal in Signed. The factorial identifies the effect of this dominant native residual; it does not separate noise magnitude from geometry. These traces describe target-space covariance, not parameter-gradient covariance. Appendix B reports the distributions and numerical checks.

Softening is competitive in risk but moves predictions farther from Fresh. We test whether a residual-free predictor can obtain competitive shifted performance through scalar output softening,

$$p_b^{(\tau)}(H) = \frac{p_b(H)^{1/\tau}}{\sum_j p_j(H)^{1/\tau}}, \qquad \tau \geq 1, \tag{8}$$

applied to p_{q_Z} or p_Q. The primary pair compares softened p_{q_Z} with p_{Fresh}; the secondary pair compares softened p_Q with p_{Signed}. For each seed, source calibration chooses the smallest evaluated temperature inside the risk-matching tolerance, not necessarily the center or root of the match. A separate source panel assesses that choice. Temperatures are fixed before shifted evaluation, and every failed match is retained.

Table 2. Source-selected controls. Differences are control minus residual-present KL in 10^{-3} nats/query. Panel A reports source match status on independent assessment; passing requires both source panels. Panel B reports coefficient-shift performance on six withheld blocks. The two q_Z/Fresh contrasts use primary 97.5% paired intervals (family size two); the Q/Signed comparison is descriptive at the same confidence level.

A. Independent source match assessment

Control comparison	Assessment difference	Seeds passing / 5
Softened q_Z arm − Fresh	-0.841	0
Early q_Z arm − Fresh	-0.201	2
Softened Q arm − Signed	-0.704	1

B. Shifted-performance comparison

Control comparison	Coefficient-shift KL difference	97.5% interval
Softened q_Z arm − Fresh	-1.385	$[-3.378, +1.014]$
Early q_Z arm − Fresh	$+1.735$	$[-0.053, +4.423]$
Softened Q arm − Signed	-0.872	$[-2.822, +1.653]$

Softened p_{q_Z} has lower independently assessed source risk than Fresh in all five seeds, with mean difference -0.000841 nats/query. Its coefficient-shift point estimate is also favorable, by 0.001385 nats/query, although the adjusted interval crosses zero (Table 2) and establishes neither superiority nor equivalence. For all five seeds, the source-risk difference is negative and its magnitude exceeds the prescribed matching tolerance of approximately 0.00046 nats/query. Softening therefore remains a credible predictive alternative despite failing the source-risk matching criterion.

The prediction-level comparison explains why aggregate risk and predictive similarity must be distinguished. On common source assessment inputs, softening the component predictor moves mean source KL from 0.013735 to 0.022940, closer to Fresh's 0.023781, while increasing $D_{\mathrm{KL}}(p\|p_{\mathrm{Fresh}})$ from 0.011769 to 0.020416 and total variation from 0.050676 to 0.066077 relative to the unsoftened predictor (Table 11). Softening therefore moves the component predictor farther from Fresh under both reported prediction-distance measures, despite remaining competitive in aggregate risk. Temperature was selected to target source risk, not to approximate Fresh's predictions; this source-only descriptive finding concerns that risk-selected transformation, not all temperatures or calibration-based explanations.

The secondary pair has the same qualitative risk outcome, with one of five source matches passing. Appendix D reports all assessment outcomes and the bootstrap procedure. Shifted comparisons reuse the factorial environments. Figure 4 plots the controls and factorial endpoints in a common source-risk–coefficient-shift-risk plane, using evaluation data separate from calibration and assessment.

We also select earlier checkpoints from the five original q_Z trajectories using source calibration alone, from forty checkpoints per seed under the original 200,000-update schedule. The selected checkpoints are from updates 50,000–65,000 and have higher mean coefficient-shift KL than Fresh for every seed: the pooled difference is $+0.001735$ nats/query, with primary 97.5% interval $[-0.000053, +0.004423]$, which includes zero. Seed 71 contributes an unusually large $+0.006147$ difference; the other four range from $+0.000404$ to $+0.001159$. Only two seeds pass both source-matching checks; the pooled comparison includes all five. Appendix E reports each selection and effect; its descriptive 95% interval is not the primary result.

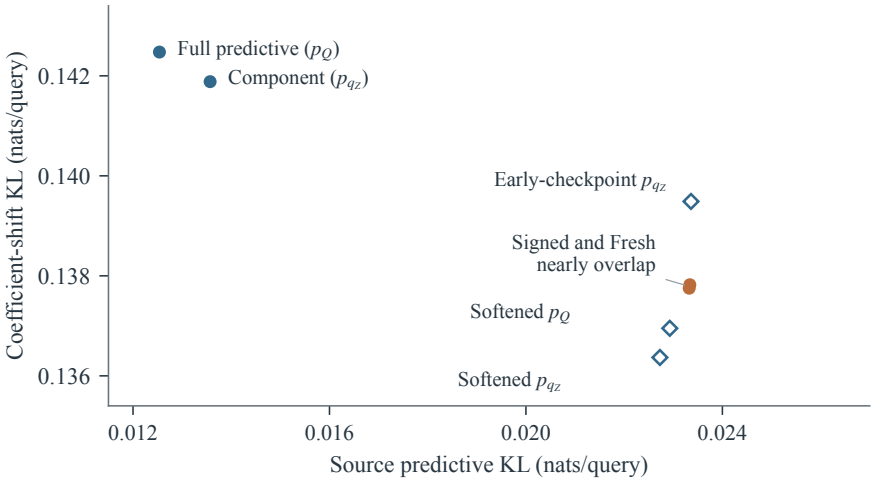


Figure 4. Source fidelity and coefficient-shift prediction in a common risk plane. Lower is better on both axes. Every point uses the same six withheld blocks, five seeds and grid-32 references; source values here are not calibration or assessment values. Points are descriptive pooled risks in nats/query, with no inferential ranking implied.

4.5 Supporting evidence: persistence under the tested schedule

The six-prior history study switches between Fresh and full-predictive supervision after 150,000 of 200,000 updates, retaining parameters, Adam state and the original cosine schedule. Switching to full-predictive supervision closes an average of 16.1% of the original within-prior source-KL gap. At fixed final supervision, an early Fresh history also retains lower coefficient-shift risk in all six prior means (Appendix F). However, the last 50,000 updates receive only 2.50% of the summed scheduled learning rates, with a mean rate of 4.99% of peak. This measures scheduled exposure, not Adam parameter displacement. Together with unequal total Fresh exposure, it limits the result to persistence at the tested horizon; it does not identify a privileged early-learning mechanism.

4.6 Practical scope

Two separate studies assess averaging-based supervision beyond the finite-prior setting. In broader-prior training from scratch, 16-sample Monte Carlo (MC16) and analytic targets average final observation noise conditional on the simulator and upstream state. Both improve synthetic negative log likelihood but establish no aggregate RMSE or likelihood gain across 92 real regression datasets. In a pretrained-continuation screen, all twelve recipes worsen development performance across thirteen datasets, and neither selected recipe qualifies for held-out evaluation. These averaging operations differ from the factorial intervention, and continuation uses a different loss. Appendix G gives the specifications and results.

5 Discussion

At a fixed training budget, removing supervision randomness can improve source-prior fidelity while worsening prediction under prior shift, despite preserving the expected training objective. The four-arm intervention identifies the native label residual as sufficient to reproduce this reversal without additive component-target variation. Across six priors, the pooled sampled-versus-full-predictive ranking changes with shift magnitude and construction. Source fidelity alone therefore cannot determine which supervision yields better shifted prediction.

The diagnostics sharpen this conclusion. Residuals supply nearly all native conditional target variance on the source diagnostic panel, so Signed retains the dominant stochastic term while removing additive component-target variation. Output softening offers a different route to competitive aggregate risk, yet the risk-selected transformation moves predictions farther from Fresh on common source assessment inputs. Obtaining competitive risk and reproducing a learned predictive distribution are distinct

outcomes. The intervention, noise-scale measurement and prediction comparison together show why supervision should be evaluated through both risk and predictive behavior.

Scope and open questions. The selective localization concerns one prior and recipe at native amplitudes; the six-prior study measures the combined effect of the two randomness terms. Posterior-resampled Fresh preserves per-query target laws but changes cross-query dependence relative to ordinary simulator supervision. The shifted comparisons rank learned predictors and do not establish superiority over the source Bayesian predictive on shifted inputs. Supervision estimators should therefore be evaluated by the predictors they produce under deployment shift, not only by their fidelity to the source-prior predictive.

REPRODUCIBILITY STATEMENT

Appendices A–C specify the target laws, model, training recipe, numerical references and factorial statistics. Appendices D–G describe controls and supporting experiments. The anonymous source package includes an executable CPU example in `reviewer/`: the finite-prior generator, target-identity tests, four-arm training and scoring, recipe configurations, and reconstruction of reported point contrasts from saved means. Its README gives commands, dependencies and provenance. The small run verifies the execution path; it does not reproduce the headline training effect. Full training banks and checkpoints are not included. Appendix H distinguishes these materials from reconstruction of saved results.

AI USE STATEMENT

AI assistants were used to aid and polish manuscript writing; retrieve, discover and verify related literature; and assist with research ideation and execution, including planning, mathematical formulation, experiment implementation and auditing, and result interpretation. Proposed claims were checked against experiment records and primary sources during preparation. The authors take responsibility for all analyses, claims and references.

REFERENCES

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 483–513. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/blanc20a.html.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27449–27461. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e6af401c28c1790eaef7d55c92ab6ab6-Paper.pdf.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Peter Grünwald and Thijs van Ommen. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, 12(4):1069 – 1103, 2017. doi: 10.1214/17-BA1085. URL https://doi.org/10.1214/17-BA1085.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network, 2015. URL http://arxiv.org/abs/1503.02531.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6.

Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, Huai-Hong Yin, Tao Zhou, Jun-Peng Jiang, and Han-Jia Ye. Talent: A tabular analytics and learning toolbox. *Journal of Machine Learning Research*, 26(226): 1–16, 2025. URL http://jmlr.org/papers/v26/25-0512.html.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7632–7642. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/menon21a.html.

Samuel Müller, Noah Hollmann, Sebastian Pineda-Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25660–25676. PMLR, 2023. URL https://proceedings.mlr.press/v202/nagler23a.html.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/qu25d.html.

Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=SxsyLjIfWB.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 14228–14246. Curran Associates, Inc., 2023. doi: 10.52202/075280-0626. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2e10b2c2e1aa4f8083c37dfe269873f8-Paper-Conference.pdf.

Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning methods on tabular datasets, 2025. URL https://arxiv.org/abs/2407.00956.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7654–7663. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhu19e.html.

A TARGET IDENTITIES AND THE SCOPE OF EQUAL EXPECTATION

For each fixed input H, let $a_z = P(Z = z \mid H)$. Since $\mathbb{E}[L \mid Z, H] = q_Z$, the two deviations in Eq. (2) have mean zero and

$$\mathbb{E}[AR^\top \mid H] = \mathbb{E}[(q_Z - Q)\mathbb{E}[R^\top \mid Z, H] \mid H] = 0.$$

Consequently,

$$\mathrm{Cov}(T_{c\ell} \mid H) = c^2 B + \ell^2 W, \tag{9}$$

$$B = \sum_z a_z (q_z - Q)(q_z - Q)^\top, \tag{10}$$

$$W = \sum_z a_z \{\mathrm{diag}(q_z) - q_z q_z^\top\}, \tag{11}$$

$$B + W = \mathrm{diag}(Q) - QQ^\top. \tag{12}$$

The last equality is the law of total covariance for a categorical label. Conditional uncorrelatedness does not make the two terms independent, and the binary interventions do not equate B and W. The contribution is the selective experimental intervention enabled by these identities, not the identities themselves. For fixed real scalars, the extension

$$T_{\alpha,\beta} = Q + \alpha A + \beta R, \qquad \mathbb{E}[T_{\alpha,\beta} \mid H] = Q, \qquad \mathrm{Cov}(T_{\alpha,\beta} \mid H) = \alpha^2 B + \beta^2 W$$

preserves the same expected objective. A source-defined choice such as $\beta^2 = \mathbb{E}_s \,\mathrm{tr}\, B / \mathbb{E}_s \,\mathrm{tr}\, W$, with $\alpha = 0$, would match the mean target variance of a residual-only arm to that of q_Z. Such matching would equate neither covariance orientation nor higher moments. Gradient-space comparison requires $J^\top B J$ and $J^\top W J$ at a common model state and on common inputs; comparing different trained states would mix noise differences with changes in J. No scaled-arm training result is reported here.

Every target sums to one, including Signed. For cross-entropy,

$$\mathcal{L}_T(f) = -T^\top f + \log \sum_b e^{f_b}, \qquad \mathbb{E}[\mathcal{L}_T(f) \mid H] = \mathcal{L}_Q(f).$$

The gradient with respect to logits is $p - T$, giving Eq. (6). For Signed, the zero-sum residual cancels the log-partition term in its correction. If some target entry $T_j < 0$, sending $f_j \to -\infty$ while holding the others fixed sends $-T_j f_j \to -\infty$. Thus the loss for a fixed signed target need not be bounded below as a function of the logits. This does not contradict its expected cross-entropy identity at fixed logits. In particular, an expectation identity is not a justification for optimizing a single signed realization to completion.

Population objective versus realized bank. The identities concern conditional target laws. Distinguish a fixed input bank, independent target draws stored for successive visits, and reuse of a previously realized target. The experiment uses the first two and does not recycle a target draw. Precomputation alone does not bias the intended estimator. Let $\mathcal{F}_{t-1}$ contain the optimizer history and previously consumed targets, excluding unused stored draws. Under the ideal sampler and target-independent input traversal, each current query target satisfies

$$\mathbb{E}[T_t \mid H_t, \mathcal{F}_{t-1}] = Q(H_t).$$

The numerical qualifications below apply to its implementation. Conditional on an entire realized target bank, exact averaging to Q is not guaranteed. Nor does the conditional-mean property imply identical realized gradients, Adam updates or parameter trajectories. Replacement banks test sensitivity to target draws; they do not isolate variance as the sole mediator of a learned difference.

Numerical storage adds a second distinction. Component targets are rounded to float32 and normalized once; the numerical Q reference is cast to the loss precision. Labels are sampled from the original float64 component laws. A guard compares the mixture of rounded, normalized component laws and the rounded predictive, with maximum-coordinate tolerance 2×10^{-6}, rather than asserting bitwise identity. The mathematical common-objective construction describes the underlying conditional laws, with numerical discrepancies subject to these guards and grid checks.

B FINITE-PRIOR CONSTRUCTION AND IMPLEMENTATION

B.1 STRUCTURAL MODEL AND OBSERVATIONS

The atom prior and order prior are independent and uniform. The factorial's eight covariance atoms use generator seed 2026092301. Log standard deviations are drawn uniformly from $[\log 0.6, \log 1.5]$; off-diagonal correlation magnitudes lie in $[0.3, 0.8]$ with equiprobable signs. Candidate correlation matrices must be positive definite and pass coefficient and innovation-scale constraints for all six orders. Production experiments used the same serialized float64 covariance matrices across machines; the later diagnostic regeneration is described separately below.

For covariance Σ and permutation π, let

$$C = \mathrm{chol}(\Sigma_{\pi,\pi}), \quad C_u = C\,\mathrm{diag}(C_{ii})^{-1}, \quad U = C_u^{-1}, \quad b_i = C_{ii}/\sqrt{2}.$$

The strictly lower-triangular structural coefficients are $-U_{ij}$. Acceptance requires $|U_{ij}| \le 1.5$ off diagonal and $b_i \in [0.3, 1.3]$. With independent residuals e, observations in permuted coordinates are $v_\pi = eC_u^\top$, subsequently returned to the original order. Each residual follows

$$(1-\epsilon)\mathcal{N}(0, 2b_i^2) + \epsilon\,\mathrm{AL}_{\mathrm{centered}}, \qquad \epsilon = 0.75.$$

For the asymmetric-Laplace component, set $c_i = \sqrt{2b_i^2/17}$, $a_i = 4c_i$ and $u = e_i + (a_i - c_i)$. The density is $\exp(-u/a_i)/(a_i + c_i)$ for $u \ge 0$ and $\exp(u/c_i)/(a_i + c_i)$ otherwise. The centering and matched variance preserve the intended covariance for each ordering.

An episode contains 20 context observations and seven query observations. The predictor receives the context and the first two coordinates of one query, predicting its third coordinate. For each query, the forward pass concatenates a task token, the 20 context tokens and that query token only. Other query covariates cannot enter that prediction. The conditional input $H = (D, x)$ therefore agrees with the reference's conditioning set.

B.2 NUMERICAL POSTERIOR AND TARGET SAMPLING

Let $f_z(x, y)$ be the component joint query density,

$$n_{zb}(x) = \int_{I_b} f_z(x, y)\,dy, \qquad d_z(x) = \int_{\mathbb{R}} f_z(x, y)\,dy.$$

With context weights $w_z = P(z \mid D)$, the query-conditioned weights, component laws and full predictive are

$$a_z = \frac{w_z d_z}{\sum_j w_j d_j}, \qquad q_z(b) = n_{zb}/d_z, \qquad Q_b = \frac{\sum_z w_z n_{zb}}{\sum_z w_z d_z}.$$

The executable path computes the context posterior from the first 20 observations and then the query operator from that query's two covariates. Zero-weight components cannot be sampled. Probability guards require finite, nonnegative, normalized laws. Mixing the component laws with context-only weights would omit the query update and is not the construction used here.

The 100 bins have the 99 internal borders of an equally spaced 101-point grid on $[-8, 8]$. The two outer bins extend below -7.84 and above 7.84, respectively; each interior bin has width 0.16. Training uses 32 Gauss–Legendre nodes per interior bin and 128 tail nodes after a rational transformation. Evaluation compares 32/128 and 128/512 interior/tail node settings. These finite comparisons assess numerical sensitivity, without certifying exact integration over all possible inputs.

For the outer bins, Gauss–Legendre nodes $t \in (-1, 1)$ map to $u = (t+1)/2 \in (0, 1)$ and then to $y = -7.84 - u/(1-u)$ on the left or $y = 7.84 + u/(1-u)$ on the right. The rational map has unit scale and absolute Jacobian $(1-u)^{-2}$; transformed weights include the additional factor $1/2$ from the first map. The implementation is supplied in `reviewer/pfn_dag_verify/pilot_shared.py`.

Each training seed indexes precomputed component/label draws for seven visits. For each query and visit separately, the sampler draws Z from the query-conditioned a by its cumulative probabilities, then samples labels from that same q_Z. Fresh and Signed use the first draw of the stored 16-label stream. The component draw is not the episode's generating latent state. Signed is implemented as

$$\mathrm{CE}(Q, p) + \mathrm{CE}(e_Y, p) - \mathrm{CE}(q_Z, p).$$

No operation clamps negative entries or converts the signed target into a probability vector.

One training update. For each of 32 episodes in the current batch, retrieve its next visit index v and process all seven queries j. Retrieve the paired $(Z_{v,j}, L_{v,j})$ draws and form $T_{c\ell,v,j} = Q_j + c(q_{Z_{v,j}} - Q_j) + \ell(L_{v,j} - q_{Z_{v,j}})$. Compute a separate forward sequence for every query and minimize

$$\widehat{\mathcal{L}}_{c\ell} = \frac{1}{32 \cdot 7} \sum_{i=1}^{32} \sum_{j=1}^{7} -\langle T_{c\ell,v_i,j}, \log p_\theta(D_i, x_{ij}) \rangle$$

with one Adam step. Inputs retain the simulator's episode dependence and shared context. Targets are sampled independently by query conditional on those inputs, so Fresh preserves the correct per-query marginal law but not the shared generating-component dependence among ordinary simulator query labels. Component and label draws remain paired across the corresponding supervision arms. Independent draws stored in advance are used once each; repeated visits reuse inputs, not a previous realized target.

For query gradients $g_1, \ldots, g_m$, the covariance of their average includes

$$\mathrm{Cov}\left(\frac{1}{m}\sum_j g_j\right) = \frac{1}{m^2}\left\{\sum_j \mathrm{Cov}(g_j) + \sum_{j \neq k} \mathrm{Cov}(g_j, g_k)\right\}.$$

The single-query covariance decomposition does not describe these cross-query terms. A sampled-label baseline retaining the simulator episode's shared component would therefore be required to bridge posterior-resampled Fresh to ordinary simulator supervision at the same training budget. That baseline has not been evaluated.

B.3 Network and training recipe

Context rows use a learned linear $3 \to 128$ embedding; query covariates use a separate $2 \to 128$ embedding. A learned task token comes from a three-entry table, always using index 2 in this experiment. The sequence contains that token, the 20 context rows and one query. There is no attention mask, positional embedding or separate feature embedding. The final query representation passes through a linear $128 \to 100$ output head. Initialization uses PyTorch 2.4.0+cu121 module defaults after setting the seed to $1000s + 7$ for training seed s; initialization is paired across arms.

The encoder uses `torch.nn.TransformerEncoderLayer` with ReLU activation, post-normalization (`norm_first=False`), layer-normalization epsilon 10^{-5}, enabled biases and `batch_first=True`. The enclosing `TransformerEncoder` has no additional final normalization. The model definition is supplied in `reviewer/pfn_dag_verify/corrected_models.py`.

The network has two transformer layers, width 128, feedforward width 256, four attention heads, zero dropout and 100 output bins. All arms use float32 model and loss arithmetic and Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\varepsilon = 10^{-8}$, no weight decay and no gradient clipping. With zero-based update t, horizon $T = 200{,}000$, warmup $u = 200$ and peak $\eta = 10^{-3}$, the learning rate is

$$\eta_t = \begin{cases} \eta(t+1)/u, & t < u, \\ \frac{\eta}{2}\left[1 + \cos\{\pi(t-u)/(T-u)\}\right], & t \geq u. \end{cases}$$

The input bank has 1,000,000 fixed episodes. Batch size 32 over 200,000 updates produces 6,400,000 context visits: six complete traversals and 400,000 visits in the seventh. Stored targets cover those visits without recycling an earlier visit. Initialization, input order and architecture are paired across arms. Checkpoints retain model, optimizer, random-number states and bank traversal position. For each seed, the q_Z arm reuses the corresponding endpoint trained with the same component-target recipe, input bank and target-draw stream.

The two replacement target banks use distinct target seeds at optimization seed 71. Their input bank, optimizer seed, Q_{71} model and evaluation panels are shared. Their scope is consequently more limited than retraining all five seeds across all banks. Actual-visit residual accounting covers the original and replacement banks, including the incomplete final traversal.

B.4 Evaluation regimes and reconstructed CRPS

Both constructions retain the source orderings, uniform atom/order weights and residual distribution family. Coefficient endpoints independently redraw every atom/order's strict lower-triangular coefficients from $\mathrm{Uniform}[-1.5, 1.5]$. At $\lambda = 0.25$, each becomes $0.75\beta_0 + 0.25\beta_1$, with source innovation scales retained. This coefficient change can itself change covariance.

Covariance-family endpoints retain each source atom's correlation signs, redraw correlation magnitudes and standard deviations, and satisfy the finite-prior validity constraints at $\lambda \in \{0.1, 0.25, 0.5, 1\}$. Correlations interpolate linearly and standard deviations logarithmically. For every ordering, the coefficients and innovation scales are then reconstructed from the interpolated covariance. The names describe constructions, not mutually exclusive statistical changes, and equal λ does not mean equal severity. Both context and query distributions follow the evaluation regime, whose numerical Q_e supplies the risk reference.

Two distinct shifted-input reference questions. The factorial reports $\mathbb{E}_{H\sim P_t} D_{\mathrm{KL}}(Q_t(H)\|p(H))$. Source-reference fidelity on those inputs, $\mathbb{E}_{H\sim P_t} D_{\mathrm{KL}}(Q_s(H)\|p(H))$, and the shifted risk of the source predictor, $\mathbb{E}_{H\sim P_t} D_{\mathrm{KL}}(Q_t(H)\|Q_s(H))$, answer different questions. They would distinguish proximity to the source rule on shifted inputs from performance relative to that rule under the shifted law. The current factorial evaluation does not include Q_s on the shifted inputs. The temporal companion analysis below concerns different priors and endpoints.

Descriptive source noise-scale diagnostic. The conditional covariance identities imply $\mathrm{tr}\, B = \sum_z a_z \|q_z - Q\|_2^2$ and $\mathrm{tr}\, W = \sum_z a_z(1 - \|q_z\|_2^2)$. A post hoc source-only diagnostic recomputes these quantities on every sixteenth assessment context (rows $0, 16, \ldots, 4080$), giving 256 contexts and 1,792 queries, each with 20 context observations. This systematic subset is used for descriptive numerical diagnostics, not as a new confirmatory sample. Both the 32/128 and 128/512 integration grids are evaluated. Table 3 gives the higher-grid summaries.

Table 3. Source target covariance scale on 1,792 queries. Quantiles are over queries, with no independence-based uncertainty claim. Effective posterior components are $1/\sum_z a_z^2$.

Quantity	Mean	Median	95th percentile
$\mathrm{tr}\, B$	0.002074	0.000947	0.008117
$\mathrm{tr}\, W$	0.911261	0.922433	0.952217
$\mathrm{tr}\, B/\,\mathrm{tr}(B+W)$	0.002285	0.001026	0.009177
Effective posterior components	1.756625	1.650844	3.022571

The ratio of mean residual to component variance is 439.40. Component variation accounts for 0.22707% of aggregate variance, which differs slightly from the mean per-query fraction of 0.22848%. Mean-trace matching would scale a residual-only arm by $\beta = \sqrt{\mathbb{E}_s\, \mathrm{tr}\, B / \mathbb{E}_s\, \mathrm{tr}\, W} = 0.047706$. This is a diagnostic scale, not a tested training setting. Posterior concentration and the small component-target variance make an amplitude explanation plausible; they do not show that it explains the learned risk differences. No gradient covariance or scaled-arm training was evaluated.

The diagnostic uses the archived source generator and numerical reference implementation, verified against their recorded hashes. The original frozen covariance binary is absent locally, and local deterministic regeneration produces a different byte hash. On every included query, the regenerated full predictive agrees with the corresponding cached reference to within 1.23×10^{-15} at both grids. A separate replay of the original 64-context pilot exactly matches all 35 stored component-target array hashes, covering 15,680 sampled component vectors at float32 precision. This verifies selected pilot components, not all component probabilities and posterior weights on the assessment inputs used for B and W; full-predictive agreement alone does not establish that stronger equivalence. The largest between-grid differences in $\mathrm{tr}\, B$ and $\mathrm{tr}\, W$ are 4.66×10^{-6} and 2.32×10^{-5}, respectively. The archive includes the pilot results in `evidence/component-provenance.json` and detailed provenance in `ARTIFACT_VALIDATION.md`.

Signed stability. All five Signed seeds completed 200,000 updates, with no seed stopped by the stability gate and no endpoint excluded. The collected logs contain 960 finite batch-loss observations,

ranging from 2.5535891056 to 3.0516715050. Seeds 71–72 have collected logs from updates 21,000–200,000 after deliberate continuation of 20,000-update pilots; seeds 73–75 cover 1,000–200,000. The implementation checks loss and gradient finiteness every update. These observations do not quantify the frequency of negative target entries, nor do they provide logit or gradient-norm summaries.

CRPS sensitivity reconstructs a law from the categorical prediction using uniform density on the interior bins and exponential tails. Conditional on an outer bin, the reconstructed variable is $-7.84-E$ or $7.84+E$, respectively, where E is exponential with scale $s=0.16f$ for tail factor $f>0$. The baseline scale is the adjacent interior-bin width 0.16 (rate 6.25); factor 0 collapses outer-bin mass to its finite border. For predictive CDF F and reference CDF G, the implementation analytically integrates expected CRPS as $\int[(F-G)^2+G(1-G)]\,dy$. The primary tail factor is 1, with factors 0 and 2 used for sensitivity checks. This score is exact for the chosen reconstructions, not the original continuous structural predictive, and is used only as a qualitative check alongside categorical KL.

C Factorial results and statistical definitions

Nominal 95% intervals use the 2.5th and 97.5th bootstrap percentiles without adjustment. The 97.5% intervals use the 1.25th and 98.75th percentiles for a Bonferroni family of two; the six-comparison curve uses the 0.41667th and 99.58333rd percentiles, giving 99.1667% intervals. Both adjusted conventions target nominal 95% familywise coverage for their stated families, not across the whole paper. Bootstrap intervals based on five seeds or six priors do not remove uncertainty about how well those small cohorts represent new runs or priors.

Let $r_{c\ell,e}=\mathcal{R}_e(p_{c\ell})$ denote the trained arm's risk in regime e, with $p_{00}=p_Q$, $p_{10}=p_{qz}$, $p_{01}=p_{\mathrm{Signed}}$ and $p_{11}=p_{\mathrm{Fresh}}$. The conventional interaction is

$$I_e=r_{11,e}-r_{10,e}-r_{01,e}+r_{00,e}.$$

It is the residual effect with A present minus that with A absent. Table 5 computes it from unrounded arm means and within each saved joint bootstrap draw, not from rounded Table 1 entries. Source nonadditivity is resolved; both shifted interaction intervals include zero.

The original scalar summary and decision-rule history are retained in Appendix H.2. Numerical summaries of both residual contrasts under KL and reconstructed CRPS on both integration grids are supplied in `evidence/audit-reporting.json`. Each contrast is positive on source and covariance-family inputs and negative on coefficient-shift inputs; this statement concerns residual-effect signs, not the complete four-arm ranking.

The direct-contrast intervals in Table 6 are calculated from differences within each of the original 2,000 stored joint bootstrap draws, preserving arm pairing. They are not new experiments or independent resamples. The bootstrap globally resamples the five paired seed settings and independently resamples the six environment blocks, averaging their crossed risks. It holds the contexts and queries within each block fixed. Numerical-grid results are not added as replicate observations.

At optimization seed 71, the two replacement banks yield $\mathcal{S}=-0.010392$ and -0.012558, and Fresh-over-q_Z coefficient benefits of 0.002612 and 0.003189 nats/query. They share a Q anchor, prior and evaluation panels. Their estimates remain separate from the five-seed original-bank intervals.

Table 4. Statistical protocol. Differences pair methods on the same inputs. Higher-level units are shown explicitly; query counts do not replace them. Fixed selection means intervals condition on the selected model, temperature or checkpoint.

Study / estimand	Replication and pairing	Resampling / confidence	Selection / status
Factorial: arm KL contrasts and I_e	One prior; five paired seeds; six withheld blocks (two earlier discovery blocks excluded).	2,000 draws: seeds and blocks independently; contexts fixed. Nominal 95%, unadjusted.	Fixed 200k endpoints. Primary metric on withheld follow-up; direct contrasts and interaction descriptive.
Six-prior: mean Fresh − full-predictive KL	Six independent priors; four paired seeds/prior; twelve directions. Equal prior weights.	5,000 draws: priors, seeds, directions, contexts. Co-primary $\lambda = .25$: 97.5% each (two comparisons). Curve: 99.1667% each (six).	Fixed 200k endpoints. Primary levels and secondary curve; study precedes factorial.
Source-selected controls: coefficient KL difference	One prior; five paired seeds; six previously examined blocks. Softening and early checkpoint versus Fresh.	5,000 draws: seeds, blocks, contexts; queries grouped. Primary 97.5% each (two comparisons); 95% descriptive.	Temperatures and checkpoints fixed. Rules frozen before new shift scoring, after factorial. Failed matches retained; Q/Signed secondary.
Source matching: per-seed risk mismatch	Separate calibration and assessment panels; paired contexts within each.	5,000 paired assessment-context draws; nominal 95%. Tolerance pass/fail is a separate criterion.	Source-only selection fixed before assessment; no reselection. Descriptive per-match uncertainty.
Source history: normalized switching effects	Six fixed priors; four fixed seeds; four histories sharing inputs.	2,000 draws: contexts within each prior. 97.5% each in recorded two-comparison families; denominators fixed.	Fixed histories/endpoints; intervals conditional on priors, seeds and trained states.
Shifted history: fixed-final-target KL contrasts	Same six priors and four seeds; twelve directions; 96 endpoints.	5,000 draws: priors, seeds, directions, contexts. Co-primary 97.5% each (two comparisons); companion analyses descriptive.	Fixed endpoints, inherited optimizer/schedule. Subsequent scoring of existing cohort.
Broader-prior scratch: NLL / geometric RMSE ratio	Three fixed paired seeds; 4,096 synthetic tasks or 92 real datasets.	Paired task or dataset bootstrap; nominal 95%, unadjusted. Average contexts/training seeds within dataset.	Fixed 18k endpoint. Separate practical study; intervals conditional on trained seeds.
Continuation: geometric development nIPS	Thirteen development datasets; twelve recipes; no continuation evaluation on 68 held-out datasets.	Development summary only; no confirmatory interval or benchmark ranking.	Development-selected recipes and resource-selected endpoints; no admission to confirmation.

Table 5. Conventional factorial interaction in 10^{-3} nats/query, using the original 2,000 paired draws. Intervals are nominal 95% and unadjusted; this decomposition is descriptive.

Regime	I_e	95% interval
Source	-1.045	$[-1.351, -0.732]$
Coefficient shift	$+0.522$	$[-0.696, +1.853]$
Covariance family	$+0.327$	$[-0.388, +1.093]$

Table 6. All four factorial simple-effect contrasts on withheld blocks. Values and nominal 95% intervals are in 10^{-3} nats per query. Negative favors the first named arm. Multiple displayed contrasts have no familywise adjustment.

Regime	Contrast	Difference	95% interval
Source	Signed $-$ Q	$+10.800$	$[+10.589, +11.015]$
	Fresh $-$ q_Z	$+9.754$	$[+9.517, +9.994]$
	$q_Z - Q$	$+1.032$	$[+0.821, +1.205]$
	Fresh $-$ Signed	-0.013	$[-0.225, +0.174]$
Coefficient	Signed $-$ Q	-4.653	$[-5.784, -3.463]$
	Fresh $-$ q_Z	-4.131	$[-5.383, -2.802]$
	$q_Z - Q$	-0.592	$[-1.267, +0.162]$
	Fresh $-$ Signed	-0.069	$[-1.056, +0.969]$
Covariance family	Signed $-$ Q	$+3.067$	$[+1.690, +4.408]$
	Fresh $-$ q_Z	$+3.394$	$[+2.306, +4.429]$
	$q_Z - Q$	$+0.107$	$[-0.208, +0.480]$
	Fresh $-$ Signed	$+0.433$	$[-0.123, +1.044]$

Table 7. Every training seed's residual effects on the six withheld blocks, in 10^{-3} nats/query. Entries are source / coefficient-shift risk differences. These five optimization replicates share one prior.

	Signed $-$ Q	Fresh $-$ q_Z
Seed	Source / Coefficient	Source / Coefficient
71	$+10.850/-4.039$	$+9.880/-4.604$
72	$+10.913/-5.193$	$+9.313/-4.579$
73	$+10.946/-5.117$	$+9.950/-3.498$
74	$+10.759/-5.261$	$+9.657/-2.952$
75	$+10.531/-3.654$	$+9.972/-5.021$

Table 8. Discovery means, kept separate from withheld estimates. Two blocks, five paired training-seed settings, 200,000 updates. Units are nats per query.

Regime	Q	q_Z	Signed	Fresh
Source	0.012445	0.013573	0.023203	0.023272
Coefficient shift	0.154651	0.153557	0.147688	0.147559

D SCALAR-TEMPERATURE CONTROL

The source calibration and source assessment panels are distinct from one another and from the source panels in the six withheld blocks. Temperature is selected separately for each of the five seeds and two pairs. Let $p_{\mathrm{ref},s} = p_{\mathrm{Fresh},s}$ for the primary pair and $p_{\mathrm{ref},s} = p_{\mathrm{Signed},s}$ for the secondary pair. Selection uses a 65-point logarithmic grid on $[1, 16]$, then at most 40 bisections of the first target-risk bracket. The chosen temperature is the smallest evaluated value whose calibration risk is within

$$\delta_s = \max\{5 \times 10^{-5},\ 0.02\,\mathcal{R}_{\mathrm{cal}}(p_{\mathrm{ref},s})\}$$

of the reference's calibration risk. It is not necessarily the numerically exact root when a smaller temperature already meets the tolerance. Assessment cannot reselect it. Every calibration match passes; Table 9 reports the independent assessment. The same tolerance is used for that assessment.

Table 9. All temperatures and independent source-assessment matches. Mismatch and tolerance are in nats/query; mismatch is softened-control minus residual-present reference KL. Primary is softened p_{q_Z} versus p_{Fresh}; secondary is softened p_Q versus p_{Signed}. All mismatches are negative.

Seed	Pair	τ	Assessment mismatch	Tolerance	Match
71	primary	1.17660	−0.000869	0.000466	no
71	secondary	1.18606	−0.000792	0.000465	no
72	primary	1.16400	−0.001172	0.000454	no
72	secondary	1.18921	−0.000200	0.000471	yes
73	primary	1.17660	−0.000845	0.000467	no
73	secondary	1.18606	−0.000738	0.000460	no
74	primary	1.17030	−0.000713	0.000461	no
74	secondary	1.18290	−0.001020	0.000457	no
75	primary	1.17660	−0.000607	0.000468	no
75	secondary	1.18290	−0.000768	0.000466	no

The transformed log probabilities are evaluated as $\log p/\tau - \mathrm{logsumexp}(\log p/\tau)$. All seeds score all six withheld blocks in source, coefficient and covariance regimes, at both numerical grids. For each of the two pairs, we retain control, reference and paired difference arrays. There are 180 comparison cells (two pairs, five seeds, six blocks, three regimes), each containing both grids and 4,096 context-mean scores; all seven query losses contribute to each context mean. No failed match is removed from the transfer analysis.

Table 10. Completed softened-control minus residual-present comparisons, in nats per query. Source rows use withheld source panels; Table 2 uses independent assessment. Only the softened q_Z/Fresh coefficient contrast belongs to the primary family with the early-checkpoint contrast; remaining rows are descriptive. Brackets give 97.5% intervals. Neither complete pair qualifies as source matched.

Pair	Regime	Difference	97.5% interval
Softened q_Z arm − Fresh	Source	−0.000596	[−0.000967, −0.000225]
	Coefficient	−0.001385	[−0.003378, +0.001014]
	Covariance family	−0.000571	[−0.002620, +0.001774]
Softened Q arm − Signed	Source	−0.000408	[−0.000736, −0.000101]
	Coefficient	−0.000872	[−0.002822, +0.001653]
	Covariance family	+0.000442	[−0.001567, +0.002766]

The 5,000-draw bootstrap independently resamples paired seeds and blocks, then contexts within each selected occurrence of a block, keeping query groups and paired methods together. Repeated selections of a block receive fresh context resamples. The two primary comparisons are softened q_Z minus Fresh and early-checkpoint q_Z minus Fresh; both are complete. Their 97.5% intervals retain the frozen family size of two. The Q/Signed comparison remains secondary and descriptive. Temperature selection itself is held fixed, so these intervals do not include all calibration-selection uncertainty. They also reuse environments already examined in the factorial.

Failed matching is neither an equivalence result nor grounds to discard a control with competitive point estimates. Matching aggregate shifted risk is distinct from reproducing the residual-trained predictions. Neither result alone identifies the optimization mechanism induced by residual training.

Descriptive source prediction comparison. A post hoc diagnostic uses the archived predictions for all five seeds on the original independent source assessment panel: 4,096 contexts and seven queries per context, with the original fixed temperatures and the 128/512-node references. No temperature is reselected and no new shifted scores enter this analysis. Table 11 reports means over the paired seeds and queries. Softening brings aggregate source risk closer to Fresh while increasing prediction divergence in both $D_{\mathrm{KL}}(p\|p_{\mathrm{Fresh}})$ and total variation. Softening therefore moves the component predictor farther from Fresh under both reported source-prediction distances. The temperature was selected to match source risk; this comparison does not assess temperatures selected to approximate Fresh's predictions.

Table 11. Source prediction diagnostics on common assessment inputs. Entropy and KL are in nats; TV is total variation distance to Fresh. Values are descriptive means, not an equivalence test.

Predictor	Source KL	Entropy	$D_{\mathrm{KL}}(p\|p_{\mathrm{Fresh}})$	TV to Fresh
p_{q_Z}	0.013735	2.749597	0.011769	0.050676
p_{Fresh}	0.023781	2.759640	0	0
Softened p_{q_Z}	0.022940	2.871511	0.020416	0.066077

The unsoftened component predictor has lower source KL than Fresh on 79.0% of assessment queries on average over seeds; after softening that fraction is 35.9%. Aggregate risk and the locations of improvements consequently provide different summaries. The source-only diagnostic does not determine shifted error overlap or whether softening can substitute for residual training in a fresh confirmatory comparison. Artifact-validation details are consolidated in Appendix H.2.

E EARLY-CHECKPOINT CONTROL

We rerun the five original q_Z training configurations with unchanged initialization, input order, target banks, numerical precision, optimizer and 200,000-update schedule, saving states every 5,000 updates to obtain forty candidates per seed. All five rerun endpoints reproduce the original model, Adam state, NumPy/Torch/CUDA random-number states, input permutation and target-visit position exactly. Here exact equality means recursive equality of decoded state fields, including tensor and array dtypes, shapes and value bytes; serialized checkpoint files need not have identical hashes. The later artifact audit compares these saved states without rerunning training. A portable checker and the recorded outcomes are supplied in `reviewer/replay_validation.py` and its fixture. The selected states retain the original full-horizon cosine schedule; neither weights nor predictors are interpolated.

The source panels are shared with the temperature control and each contain 4,096 contexts with seven queries. For each seed, we choose the checkpoint minimizing the absolute calibration-risk difference from its paired Fresh endpoint, breaking ties by earlier update and then fixed checkpoint identifier. The selection is fixed before independent source assessment and shifted evaluation. The same tolerance δ_s applies on both source panels; assessment never reselects the checkpoint.

Table 12. Every source-selected earlier checkpoint. Risk differences and tolerances are in 10^{-3} nats/query, relative to the paired Fresh endpoint. The assessment interval resamples paired contexts with the checkpoint fixed. Only seeds 74 and 75 pass both source gates.

Seed	Update	Cal. Δ	Assess. Δ	Assessment 95% interval	Tolerance	Cal./assess.
71	50k	$+0.675$	$+0.353$	$[-0.275, +0.977]$	0.466	no/yes
72	65k	$+0.242$	-0.475	$[-1.052, +0.109]$	0.454	yes/no
73	60k	-0.143	-0.628	$[-1.270, -0.030]$	0.467	yes/no
74	60k	$+0.125$	$+0.137$	$[-0.463, +0.731]$	0.461	yes/yes
75	60k	-0.245	-0.390	$[-1.006, +0.242]$	0.468	yes/yes

All five selected states are evaluated on the same six withheld blocks, three regimes and two numerical grids, giving 90 complete model-pair/panel cells. The coefficient-shift risks are 0.139488059 for the checkpoint controls and 0.137752737 for Fresh. Their difference is $+0.001735322$, with primary 97.5% interval $[-0.000053396, +0.004422951]$. The ordinary descriptive 95% interval,

$[+0.000088208, +0.004086347]$, does not replace the frozen primary interval. Every seed and block mean favors Fresh, but the adjusted ranking remains unresolved. Seed 71 has the largest difference, $+0.006146948$; the other four range from $+0.000404267$ to $+0.001159331$.

Table 13. Every seed's coefficient-shift control-minus-Fresh difference, in 10^{-3} nats/query, on the same six withheld blocks. The large early-checkpoint effect for seed 71 is retained. These are descriptive seed means; adjusted pooled inference remains that of Table 2.

Seed	Softened component target	Early-checkpoint component target
71	-0.824	$+6.147$
72	-1.218	$+1.159$
73	-1.763	$+0.404$
74	-2.509	$+0.483$
75	-0.612	$+0.483$

Seed 71 fails calibration; seeds 72 and 73 fail assessment. All are retained in the five-seed result. The small pooled withheld-source difference, $+0.000036767$, does not satisfy the required per-seed matching checks. The covariance-family difference, $+0.000863739$, is descriptive. The bootstrap uses the paired seed/block/context procedure above, with calibration and checkpoint selections fixed. These are previously examined environments under one source prior. Because selection is restricted to 5,000-update intervals, failed matching does not rule out a match at another checkpoint.

The largest aggregate grid-32/grid-128 contrast difference is below 8.8×10^{-9}, against the frozen 10^{-4} tolerance. Checkpoint replay and numerical reconstruction details appear in Appendix H.2.

F SIX-PRIOR TWO-ARM COMPARISON AND TEMPORAL HISTORY

F.1 OVERALL SUPERVISION REVERSAL ACROSS PRIORS

The six independently redrawn priors use source seeds 990093001–990093006; the original factorial source prior is excluded. Each has eight atoms, three variables and $\epsilon = 0.75$. Four paired training seeds, 71–74, train Fresh and full-predictive (Q) supervision with the architecture and optimizer recipe in Appendix B, giving 48 fixed endpoints. Initialization, input banks and traversal are paired within each prior and seed.

For each query and training visit, Fresh draws a new categorical label directly from the cached full posterior predictive vector Q. Marginalizing the posterior-component sampler in Appendix B gives the same conditional label law. Thus replacing this sampled target by Q removes the combined $A + R$ deviation in distribution; this study does not separately sample or intervene on its two terms.

Each prior has twelve direction indices, each with a separately drawn endpoint for each shift family, and levels $\lambda \in \{0, 0.1, 0.25, 0.5\}$. Each family/direction/level panel contains 512 contexts, 20 context rows and seven queries. Common random numbers pair the latent and residual draws across families and levels at a given direction index; streams differ across priors and directions. At $\lambda = 0$, the two families share a source panel for each direction, counted once, giving twelve unique source panels per prior. The main text reports the two co-primary $\lambda = 0.25$ contrasts. The covariance-family endpoint preserves the source correlation-sign pattern while interpolating correlation matrices and log standard deviations, and can change residual scales.

The estimand weights priors equally, then averages over seeds, directions and contexts. The 5,000 paired hierarchical bootstrap draws resample priors at the top level, followed by seeds, directions and contexts within each selected prior. The two co-primary 97.5% intervals adjust for two comparisons. Figure 3 displays 99.1667% intervals adjusted for the six nonzero family/level comparisons. Table 14 exposes all six primary prior effects in the Fresh-minus-full-predictive direction. This comparison removes both A and R and does not replicate the factorial localization.

Evaluation uses 128 interior and 512 tail nodes, with reference-only refinement on four input-selected contexts per panel, up to 1,024/4,096 when required. The recorded checks pass. This procedure is distinct from the factorial's two-grid check and is not a global error certificate.

Table 14. All six prior effects at $\lambda = 0.25$, in 10^{-3} nats/query. Negative favors Fresh. Each entry averages four paired seeds and twelve directions within that prior.

Prior seed	Coefficient Fresh $-$ Q	Covariance family Fresh $-$ Q
990093001	-4.919	$+5.706$
990093002	-4.012	$+3.117$
990093003	-3.989	$+4.682$
990093004	-4.652	$+4.271$
990093005	-1.796	$+4.998$
990093006	-4.959	$+6.128$

Table 15. Pooled Fresh-minus-Q differences across six priors, in 10^{-3} nats/query. The 99.1667% hierarchical intervals adjust for six comparisons and correspond to Figure 3. Unrounded values are in `evidence/audit-reporting.json`.

Shift family	λ	Difference	99.1667% interval
Coefficient	0.10	$+4.550$	$[+3.982, +5.204]$
	0.25	-4.054	$[-5.575, -2.346]$
	0.50	-8.698	$[-14.338, -3.049]$
Covariance family	0.10	$+9.435$	$[+8.694, +10.153]$
	0.25	$+4.817$	$[+3.608, +5.942]$
	0.50	-1.940	$[-4.034, -0.332]$

F.2 Source-prediction switching

The switch study preserves the six-prior, four-seed cohort and changes supervision after update 150,000. Let F denote Fresh supervision and T denote full-predictive supervision. The first and second letters identify the initial and final stages: FF and TT continue, whereas FT and TF switch. The final 50,000 updates inherit both parameters and optimizer state and follow the original schedule. Unchanged continuations are replayed against the original endpoints as an integrity check. There are 96 endpoints and 48 unchanged replay checks; the replays add no scientific observations.

The source-only switching analysis uses a separate cohort: panels 01–23 of 24 newly generated source panels per prior, each containing 512 contexts. Panel 00 was excluded because it had been inspected for engineering and numerical checks. These panel identifiers index independent source samples, not shift directions. Conditioning-context and visible-input hashes verify uniqueness within this cohort and disjointness from the twelve source panels above. This different evaluation cohort accounts for the source-gap difference between the overall supervision comparison (0.010748) and the switching study (0.010785, rounded).

For history a and prior w, define source risk over four fixed seeds and $N = 23 \times 512 = 11{,}776$ contexts by

$$R_{a,w} = \frac{1}{4N} \sum_{s=1}^{4} \sum_{i=1}^{N} \ell_{a,w,s,i}, \qquad G_w = R_{\mathrm{FF},w} - R_{\mathrm{TT},w},$$

where each context loss averages query KLs and the denominator G_w is frozen before treatment analysis. The two summaries are

$$r = \frac{1}{6} \sum_{w=1}^{6} \frac{R_{\mathrm{FF},w} - R_{\mathrm{FT},w}}{G_w}, \qquad h = \frac{1}{6} \sum_{w=1}^{6} \frac{R_{\mathrm{FT},w} - R_{\mathrm{TT},w}}{G_w}.$$

Their point estimates are 0.160615 and 0.839385 and satisfy $r + h = 1$. In each of 2,000 bootstrap draws, N contexts are sampled with replacement within each fixed prior, pairing indices across histories and seeds. Priors and seeds are not resampled, and denominators retain the original G_w. Thus $r^* + h^* = \frac{1}{6} \sum_w G_w^* / G_w$, generally not one. The reported 97.5% intervals, $[0.158615, 0.162574]$ and $[0.829935, 0.848844]$, use the 1.25th and 98.75th percentiles with linear interpolation. Their endpoints need not complement because numerators vary while denominators remain fixed. They are conditional on the six priors, seeds, inherited states and continuation label streams, and should not be read as uncertainty over new training histories.

Table 16. Pooled temporal endpoint source KL in nats/query. F denotes Fresh; T denotes full-predictive (Q) supervision. Letter positions identify first and final stages. These descriptive pooled means do not replace the per-prior ratio aggregation.

History	FF	FT	TF	TT
Source KL	0.023061	0.021329	0.013093	0.012276

The outcome concerns the fixed final horizon. Summing the stated schedule over zero-based updates $0, \ldots, 199{,}999$, the last 50,000 updates receive 2.49698% of the total summed learning rates. The rate at update 150,000 is 14.67247% of peak, and its mean over the final stage is 4.99401% of peak. These are scheduled-exposure diagnostics, not measures of parameter displacement under Adam. The result supports persistence under the tested schedule, not an irreversible approximation floor or a privileged early-learning mechanism. Learning-rate restarts, optimizer-state resets or matched-exposure designs would be needed to strengthen that interpretation.

F.3 Shifted prediction of the existing switch endpoints

The temporal-transfer analysis evaluates all 96 endpoints under coefficient shift at $\lambda = 0.25$, using twelve existing directions per prior, 512 contexts per direction and seven queries per context. All histories and seeds share the inputs and numerical references. The full cohort contains 72 distinct panels and 1,152 model–panel score cells. Its primary metric is $D_{\mathrm{KL}}(Q_t \| p)$ under the shifted law. A companion metric, $D_{\mathrm{KL}}(Q_s \| p)$ on the same shifted inputs, measures source-prior fidelity on those inputs; it is distinct from source-population risk.

Table 17. Temporal endpoint risks on coefficient-shift inputs, in nats per query. The shifted reference Q_t defines the primary predictive risk. The source reference Q_s supplies a companion fidelity measure on those same inputs.

History	Shifted predictive KL	Source-reference KL on shifted inputs
FF	0.143866	0.042755
FT	0.145063	0.041232
TF	0.146608	0.030191
TT	0.147921	0.029331

The two co-primary contrasts hold final supervision fixed: $h_T = R_{\mathrm{FT}} - R_{\mathrm{TT}}$ and $h_F = R_{\mathrm{FF}} - R_{\mathrm{TF}}$, using shifted predictive risk. They are -0.002857207 and -0.002741993, with 97.5% intervals $[-0.003949574, -0.001702001]$ and $[-0.003864617, -0.001524248]$. Both are negative in every prior mean; one of the 24 prior/seed h_F means reverses sign. The contrast changes both target-randomness components during early training and retains inherited Adam state. Total Fresh exposure is 200k, 150k, 50k and zero updates for FF, FT, TF and TT respectively. Thus the comparison does not isolate training order at equal exposure, a uniquely effective early period, or parameter history separately from optimizer history.

The estimand weights priors, seeds, directions and contexts equally, first averaging seven queries within each context. Each of 5,000 hierarchical bootstrap draws resamples six priors. Within each selected prior occurrence, it resamples four seeds and twelve directions; within each selected direction occurrence, it resamples 512 contexts. Seeds are paired across directions; contexts are paired across selected seeds, histories and metrics. Repeated prior and direction selections receive fresh lower-level resamples. The 1.25th and 98.75th percentiles preserve the two-comparison primary family. Other contrasts and source-fidelity intervals are descriptive.

Numerical qualification uses the reference panels described above. On four input-selected refinement contexts per panel, the contrast check weights the refined-reference discrepancy by the absolute paired log-probability contrast, averaging over seeds and queries. For both primary test-KL contrasts, these diagnostic-subset bounds on the inter-grid contrast discrepancy have maxima below 5.23×10^{-7} and means below 1.72×10^{-7} across panels, against thresholds 5×10^{-4} and 10^{-4}. They do not bound integration error for all 512 contexts or for the exact predictive reference.

Saved-score reconstruction and its boundaries are reported in Appendix H.2; the numerical-reference checks retain the diagnostic-context scope above.

G Practical scope studies

G.1 Broader-prior training from scratch

This study is separate from the finite-prior factorial. A row transformer is trained on a 70% MLP / 30% tree simulator mixture, using an archived TabICL v0.1.4 prior implementation (Qu et al., 2025), commit `1badfa744f3098926acae9fabf63e0cec5dd7237`. It has twelve layers, width 512, feedforward width 1,024, four attention heads and 1,000 bins. Three paired seeds, 4101–4103, train each arm for 18,000 updates at effective batch size 512, giving 9,216,000 fresh tasks per run. Adam uses learning rate 3×10^{-4}, betas $(0.9, 0.999)$, no weight decay, 2% linear warmup, cosine decay and gradient-norm clipping at 1. Forward computation uses bfloat16 and losses float32.

Each task has 1,024 rows and 2–100 features; the context size is sampled uniformly from 102–922 per effective batch. Initialization, task identities, covariates, context labels and queries are paired across supervision arms. Eligible tasks have a noncausal simulator or terminal causal target. MC16 averages losses from sixteen independent draws of final target observation noise, holding simulator, upstream state, covariates and context fixed. Analytic supervision integrates that Gaussian noise into bin probabilities. Ineligible tasks use identical sampled losses in all arms. This is generating-simulator-conditional averaging, not posterior integration or the finite-prior q_Z intervention.

Bins are calibrated on 4,096 independent simulated tasks with seed 918237; density components are uniform inside bins with half-normal tails. Context-fitted normalization and deterministic invalid-draw rejection define the effective distribution. Some synthetic categorical-feature conversion can inspect the complete feature column before context normalization; we do not claim that every synthetic preprocessing operation is context-fitted.

Evaluation uses 4,096 held-out synthetic tasks and all 92 eligible regression datasets in the frozen TALENT inventory (Liu et al., 2025; Ye et al., 2025). The inventory uses `dataset-latest.zip` from the Hugging Face dataset repository `LAMDA-Tabular/TALENT`, at revision `7bd276bcb7f6b4c0998025855528bd76bd88f13d`; this is a dataset-repository revision, not a toolbox-code revision. The exact 92-dataset manifest, download URL and archive hash are included in `evidence/broader-prior-inventory.json`. Eligible datasets have 2–100 features in total, including numeric and ordinal-encoded categorical features, and at least 32 training rows. Official partitions are retained. Five context draws, seeds 8101–8105, use at most 1,024 training rows; preprocessing is fitted to that context. The 18,000-update checkpoint is primary. Real-data RMSE is averaged over contexts and optimization seeds within each dataset before taking an equally weighted geometric ratio across datasets. NLL differences receive equal dataset weight. Nominal paired bootstrap intervals resample datasets or synthetic tasks, conditional on the three trained seeds, without multiplicity adjustment.

Table 18. Broader-prior study. NLL differences are in nats/query; real-data geometric RMSE changes are percentages. Negative favors the first arm. Nominal 95% paired intervals condition on three trained seeds, with no multiplicity adjustment.

Evaluation	Comparison	Estimate	95% interval
Synthetic NLL	MC16 − sampled	-0.001538	$[-0.002046, -0.001035]$
	Analytic − sampled	-0.000780	$[-0.001266, -0.000289]$
Real RMSE (%)	MC16 − sampled	$+0.1833$	$[-0.2836, +0.6782]$
	Analytic − sampled	-0.0788	$[-0.7267, +0.5037]$
Real NLL	MC16 − sampled	-0.005859	$[-0.013324, +0.001315]$
	Analytic − sampled	-0.005200	$[-0.013275, +0.002353]$

The real-data likelihood estimates use 2,000 paired dataset bootstrap draws (seed 20260915); both intervals contain zero. Synthetic paired NLL is evaluated directly through bin cross-entropies, cancelling the common within-bin density contribution and retaining all tasks. The saved real-score reconstruction contains 92 datasets, five context seeds and three optimization seeds per arm, or 4,140 dataset-by-context-seed-by-training-seed-by-arm score records.

G.2 PRETRAINED-CONTINUATION DEVELOPMENT SCREEN

The continuation model is TabICL regressor v2 (QU et al., 2026), checkpoint `tabicl-regressor-v2-20260212.ckpt`, implemented with TabICL 2.2.0 at commit `0dbff3ec8fc68c123c87af77b0ea8b25cd2d23f3`. It contains 28,544,991 parameters, of which 28,544,983 are trainable. Eight rotary-frequency parameters remain fixed, as in the pretrained implementation. Continuation passes all model parameters to AdamW without additional layer freezing. AdamW uses learning rates 3×10^{-7} or 10^{-6}, betas $(0.9, 0.999)$, epsilon 10^{-8}, weight decay 0.01, and 105 updates of linear warmup followed by a constant rate. The effective batch contains 64 tasks, accumulated in eight microbatches of eight, with 1,024 rows, at most 100 feature slots and up to 128 queries. Training averages empirical pinball loss over 999 quantiles; it does not use the factorial's categorical loss.

Each update mixes 32 native-prior tasks and 32 broader-prior tasks. The archived native-task factory calls the released TabICL `graph_scm` generator with `PriorConfig()` defaults and the continuation size constraints above. It generates groups of four tasks sharing the requested feature count, sequence length and context boundary, with a separate mechanism per task. This specifies the available factory; original configuration receipts for the historical cached prefix are incomplete. Native here does not mean a reconstruction of the checkpoint's full pretraining distribution. The model, stream and label seeds are 2026092400, 8190 and 2026092430.

For an eligible broader-prior task, let y_i and μ_i be the cached target and conditional mean. Noise scale a multiplies the observation residual for both context and query rows: $v_i = \mu_i + a(y_i - \mu_i)$. For $a \neq 1$, context-only normalization is recomputed with mean m and population standard deviation $s = \max\{\mathrm{sd}(v_{\mathrm{context}}), 10^{-6}\}$, yielding $y'_i = (v_i - m)/s$, $\mu'_i = (\mu_i - m)/s$ and $\sigma' = a\sigma/s$. Scale 1 retains the cached coordinates. Only query supervision is freshly sampled as $\mu'_q + \sigma'_q\varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$, using query-indexed streams; context labels retain the transformed cached draws. The sampled query label is the first of the MC16 draws. Native and ineligible tasks retain sampled supervision.

Twelve recipes cross noise scales $0.5, 1, 2$, sampled or MC16 targets, and the two learning rates. Thirteen development datasets in twelve families give 156 scored recipe–dataset cells. A recipe proceeds to held-out evaluation only if it improves development normalized integrated pinball score (nIPS) by at least 0.2% over the unchanged pretrained model U. For $\tau \in \{0.01, 0.02, \ldots, 0.99\}$, nIPS is twice the mean pinball loss $\rho_\tau(y - \widetilde{q}_\tau)$ over examples and quantiles, divided by $\max\{\mathrm{sd}(y), 10^{-12}\}$ using population standard deviation. Predictions are made nondecreasing by setting $\widetilde{q}_\tau = \max_{\tau' \leq \tau} q_{\tau'}$. The screen aggregates datasets geometrically. This finite quantile-grid score is not exact continuous CRPS.

Within each six-recipe strategy, selection minimizes equal-dataset geometric nIPS. A 0.05% tie tolerance favors lower learning rate, then noise scale closest to one in log distance, then lower noise. Endpoint update counts are fixed by resource estimates. The exact dataset inventories, family assignments and recipe specifications are preserved in the accompanying benchmark and screen manifests. The development datasets are 2dplanes, Ailerons, BNG(lowbwt), BNG(mv), IEEE80211aa-GATS, Large-scale_Wave_Energy_Farm_Sydney_100, Large-scale_Wave_Energy_Farm_Sydney_49, Moneyball, NASA_PHM2008, chscase_foot, gas_turbine_CO_and_NOx_emission, qsar_aquatic_toxicity and stock.

All twelve recipes worsen development nIPS. The selected sampled and MC16 recipes both use noise scale 1 and learning rate 3×10^{-7}. Their endpoints were set to 721 and 714 updates using estimated resource costs; the two runs were not matched on exactly measured compute. The unchanged nIPS is 0.098684381; selected sampled and MC16 values are 0.099542965 and 0.099561074. Relative increases are 0.8700% and 0.8884%, with RMSE increases of 0.4502% and 0.4887%. Held-out evaluation additionally requires at most 0.5% RMSE harm; neither recipe passes the nIPS condition.

Neither recipe met the criterion for held-out evaluation, so no continuation recipe was evaluated on the held-out inventory of 68 datasets in 59 families. Unchanged-baseline inference was already cached for those datasets; it supplies no continuation comparison. The result concerns these development recipes and does not establish a general failure of MC averaging. The inventory revision history appears in Appendix H.2.

H CHRONOLOGY AND REPRODUCIBILITY RECORDS

Table 4 records the cohorts and resampling units. The two replacement banks are separate sensitivity checks with one optimization seed and shared inputs, Q model and evaluations; they add no prior-level replication. Numerical grids and replay checks add no scientific replicates.

The six-prior overall supervision comparison and temporal continuation precede the four-arm component/residual factorial. Two factorial environment blocks are discovery data; six are withheld under the recorded follow-up procedure. The two target-bank replacements are a further sensitivity check at one optimization seed. The scalar-softening and early-checkpoint analyses follow the factorial and use the already studied environments. Their selection algorithms and assessment rules were fixed before the newly computed shifted scores were inspected. The first readout seed was fixed in advance, and all five seeds completed regardless of its result. The source noise-scale and prediction-distance diagnostics were added after these analyses and are descriptive. These stages do not constitute one wholly preregistered experiment.

H.1 SAVED ARTIFACTS AND COMPUTE

The figure/table builder reads unrounded means and seed arrays from `factorial-export.json`, the original joint draws from `withheld-bootstrap.npz`, per-prior trajectories from `six-prior.json`, and control risk arrays from `temperature-summary.json` and `early-summary.json`. Temporal results retain their separate score summary, bootstrap draws and numerical-bound arrays. The common-panel risk plot checks that the controls' reference-arm risks agree with the factorial's withheld risks; no source calibration or assessment risks enter that plot. The broader-prior readout and continuation benchmark/screen manifests retain dataset effects, inventories, family assignments and selection rules.

The implementation records distinguish the simulator/reference code (`corrected_models.py`, `corrected_oracle.py`), visit-target sampler (`fresh_data.py`, `targets.py`), and factorial training, environment, scoring and aggregation programs (`factorial.py`, `environments.py`, `scoring.py`, `report.py`). Typesetting dependencies are separate from experiment dependencies. The anonymous `reviewer/` package includes the original structural-model and reference modules and exact extractions of the production target, forward-pass and loss functions. A documented adapter supplies the fixed bin edges locally. The small example reduces data and updates and omits cluster orchestration; its README and provenance record identify these changes and the actual software versions. Identity tests and saved-means contrast reconstruction accompany it. The package excludes production banks and checkpoints, so executing the small example does not reproduce full-run effects.

The factorial used NVIDIA A6000 allocations. Five packed production allocations lasted 2.73–3.12 hours each; multiple arms shared each allocation and some resumed from 20,000 updates, so these are not per-arm full-run times. Evaluation-reference generation for eight blocks and three regimes, each with 4,096 contexts, seven queries and both grids, used 24 jobs of 290–340 seconds on 16 CPU cores, totaling approximately 33.46 allocated CPU-core-hours. Twenty withheld-scoring jobs used 71–86 seconds each, totaling approximately 0.42 allocated GPU-hours. The factorial campaign ledger totals approximately 22.82 allocated GPU-hours including auxiliary checks, failures and diagnostics, but excludes predecessor campaigns and is not the whole-paper cost. The continuation screen consumed approximately 31.00 allocated GPU-hours including failures, diagnostics, restarts and inference.

H.2 AUDIT AND ARTIFACT VALIDATION

The original scalar statistic, stored as `H`, is denoted here by $\mathcal{S} = (r_{01} - r_{10})_{\text{coef}} - (r_{01} - r_{10})_{\text{source}}$. Discovery $\mathcal{S} = -0.015499$; withheld $\mathcal{S} = -0.013829$, with 95% interval $[-0.014997, -0.012572]$. All five seed-specific and six block-specific withheld values are negative. Because $\mathcal{S}$ includes source differences, its interval does not replace the coefficient simple-effect intervals in Figure 2.

The plan specified component and residual decision branches before withheld evaluation. The component branch required $\mathcal{S} \geq 0.001$ and both coefficient-shift benefits of adding A to be at least 0.001 nats/query; the residual branch required $\mathcal{S} \leq -0.001$ and both benefits of adding R to meet that threshold. Both required the Fresh/Q source–shift crossover and consistent seed signs.

The component branch failed and the residual branch passed discovery, withheld evaluation and replacement-bank checks. The full staged criteria, including block-sign and numerical checks, are preserved in `ARTIFACT_VALIDATION.md`. These historical decision rules differ from the direct contrasts emphasized in the paper.

Saved-artifact checks reconstruct factorial means and joint contrasts, control selections and bootstrap intervals, and temporal endpoint aggregates. Sampler/loss inspection verifies query conditioning and per-query forward isolation, and the source-control audit verifies disjoint calibration and assessment contexts. These checks reconstruct saved statistics and selected states; they do not independently rerun all training, model inference or integration. The anonymous source package's `ARTIFACT_VALIDATION.md` preserves the detailed reconstruction counts, numerical discrepancies and checkpoint identifiers.

The combined control exports contain 15 source-matching records, 540 detailed transfer rows and 18 pooled summary rows. The detailed rows comprise 360 temperature and 180 checkpoint rows, counting both grids; the pooled rows comprise twelve temperature and six checkpoint summaries. Pooled summaries and numerical grids are not additional observations.

The early-checkpoint aggregation uses a validation-only numerical tolerance; selections, scores and inference rules are unchanged. The continuation screen followed an earlier attempt excluded from confirmatory analysis because development and confirmation sets overlapped at the dataset-family level. The revised held-out inventory contains 68 datasets in 59 families, superseding the planned count of 71; the earlier attempt supplies no confirmatory evidence. Detailed validation receipts remain in the accompanying artifact record.