SLOW PREDICTION CONCENTRATION IN TABULAR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

How do tabular foundation model predictions change with more labelled rows at inference? We measure prediction spread across contexts and the effect of deleting a single row for six released checkpoints from four model families. On disjoint higgs contexts of 1024–16,384 rows, four checkpoints have fitted probability spread slopes between -0.08 and -0.14, compared with -0.51 for logistic regression. At 16,384 rows, their accuracy is 8–10 percentage points higher than logistic regression's. An exact-tree XGBoost reference refit on the same supports also concentrates faster. These slopes describe the observed range. They give no asymptotic guarantee. Deleting a support row can also alter the fingerprint features of retained rows. For TabPFN v2.5, we compare fingerprints on and off across two datasets, six support seeds, four sizes (64–512), and 50 fixed, unfiltered queries per dataset. Disabling fingerprints makes mean pairwise influence decay faster on both probability and log-odds scales. All four paired contrasts survive Holm correction. The log-odds exponent rises by 0.239 on higgs and 0.550 on magic. Faster influence decay does not by itself imply better prediction or a population-wide effect. Whether spread eventually concentrates faster at larger context sizes remains open.

1 INTRODUCTION

Tabular foundation models predict from labelled rows supplied at inference, without updating their weights. How much a prediction depends on those rows determines how much it varies across contexts of the same size. For a single attention layer with bounded inputs and finite weight norms, Nagler (2023) derives an $n^{-1/2}$ concentration rate across context draws. Our models are deeper, so we use this rate as a reference rather than a prediction for their behavior.

We study two kinds of context dependence in six released checkpoints from four families. *Prediction spread* is the standard deviation of a query's predicted probability across draws of the labelled context, or support set. *Single row influence* measures the change when one support row is removed; it is related to finite sample influence in robust statistics (Hampel, 1974). We fit each quantity against support size n and compare the checkpoints with classical learners refit on the same data.

Across 64–1024 rows on higgs and magic, spread decays more slowly than $n^{-1/2}$ for all six checkpoints. For 1024–16,384-row disjoint higgs supports, four have slopes between -0.08 and -0.14; logistic regression has a slope of -0.51 $[-0.53, -0.48]$ (Table 1). At the largest size, the checkpoints are 8–10 percentage points more accurate than logistic regression. We also refit an exact-tree XGBoost model on the same supports. Its accuracy is closer to the checkpoints', but its spread falls faster through 32,768 rows (Figure 1). Disjoint supports rule out overlap between context draws as the source of the result; XGBoost shows that it is not limited to a linear comparison. On twelve further datasets, two checkpoints reach the reference rate only on pol and wilt, subject to the qualifications in Section 4.5.

Preprocessing matters for single row influence. TabPFN's default fingerprint feature hashes preprocessed rows, and deleting one support row can change the hashes assigned to rows that remain. We test this mechanism with one v2.5 checkpoint on higgs and magic, holding queries, supports, deletions, and CUDA settings fixed. Disabling the fingerprint flag increases the mean pair decay exponent on both probability and log-odds scales (Table 2). Frozen preprocessing and a fixed salt provide

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.

further, exploratory controls. Earlier influence measurements select correctly classified queries, so they estimate a different quantity (Section 2). A separate accuracy comparison across fourteen datasets does not detect a loss from disabling fingerprints, though it does not establish equivalence.

Our measurements cover several recent checkpoints, disjoint support comparisons, and a specific preprocessing mechanism that affects retained rows. The observed slopes do not rule out faster asymptotic concentration, nor does the fingerprint experiment explain the spread slopes. Randomizing weights does not cleanly isolate architecture: accuracy and label response collapse in those controls (Section 4.4). We report failed predictions, exclusions, and gaps alongside the results they qualify. Appendix K identifies the supporting records.

Related work. Biloš et al. (2026) study readout mechanisms in three tabular model families at nearly fixed context size; they do not fit single row influence decay. Their Mitra readout label does not transfer to our checkpoint under the registered diagnostic (Appendix I). Other robustness and interpretability studies hold n fixed (Peroni et al., 2026; Miftachov et al., 2026; Rundel et al., 2024) or vary it without fitting a rate (Hu & Ghelichi, 2026; Djilani et al., 2026). Context-size variance has also been measured before: Nagler (2023) and Koshil et al. (2024) use synthetic data, while Liu & Ye (2025) study Adult and Bank through 12,800 rows. McCarter (2025, Appendix A.1) links default fingerprints to irregular probability functions. Ye et al. (2025, Appendix D) report limited benchmark gains from adding fingerprints and polynomial features together. Appendix A discusses stability and adversarial analyses of text and synthetic in-context learning (Anwar et al., 2025).

2 Method

At each support size n, we predict the same 50 fixed, unfiltered held out queries using 24 to 48 support draws. The 32,768-row experiment uses 16 draws; Nagler's synthetic testbed uses 200 to 500 draws and 100 queries. For each query, we take the standard deviation of its predicted class-1 probability across draws, then average over queries. We fit the log of this *spread* against $\log n$ by ordinary least squares. A slope at or below $-1/2$ *reaches the reference rate*. We also report whether the 95% support-resampling interval lies above, contains, or lies below $-1/2$. A measurement is excluded from the fit if its spread is less than three times the standard deviation under permutations of one fixed support. Initial draws may overlap, so the extended higgs experiments use disjoint supports. Appendix D describes the overlap correction and resampling intervals.

For the historical influence experiments, we draw a support S for each dataset, seed, and n, then take the first 50 queries that the model classifies correctly. Thus query sets can differ by configuration, model, seed, and support size. Historical fingerprint contrasts pair seeds but need not pair queries. The common-query experiment instead uses the first 50 unfiltered held out rows for every configuration, size, and seed (Appendix G.1). We sample k support rows uniformly without replacement for deletion, using the same sampled rows across models within a measurement. For output f and query q, the absolute response to deleting row i is $d_i(q) = |f(S, q) - f(S \setminus i, q)|$. We measure both predicted probability and the log-odds margin $\log p_1 - \log p_0$. The deletion mean $Q_{\text{mean}} = \text{mean}_i\, d_i$ and maximum $Q_{\max} = \max_i d_i$ are computed per query, then averaged over queries.

Deletion changes both row content and context length. Our main influence statistic compares two deletions at the same length:

$$Q_{\text{pair}}(q) = \text{mean}_{i \neq j} |f(S \setminus i, q) - f(S \setminus j, q)|.$$

Its maximum over pairs is $Q_{\text{pair}}^{\max}(q)$, with both averaged over queries. These statistics remove any additive response shared by all deletions. Multiplicative margin drift across n remains. With $k = n$, the maximum is exact over the observed deletion pairs, not over all possible replacement rows or supports. With $k = 32$, it becomes a sampled maximum. Mean statistics estimate the full support mean without the maximum's downward sampling bias, although their fitted exponents still have sampling error.

We report influence decay as α, the negated per-seed OLS slope of $\log Q$ against $\log n$, so positive α means declining influence. Results are means and standard errors over six seeds, or three in the legacy $k = 128$ sweeps. In historical analyses, an exponent is classified above or below $1/2$ only when its unrounded t statistic has magnitude at least 3. Between model separations require 3 pooled SE. With six seeds, the reference distribution is t_5. The common query experiment uses

paired fingerprint-off minus shipped mean pair exponent contrasts: two datasets and two scales form one four-test Holm family at 0.05, with ordinary, non-simultaneous 95% intervals. Appendix E reports the registered fit windows and the Holm correction over the mapped comparison family. Query selection, Kish's effective sample size (Kish, 1965), and fixed query checks are detailed in Appendix B.

Measurement checks and configurations. We test order sensitivity with 20 support permutations, assess randomized controls with label flips, and measure repeat variation and numerical precision. For order sensitive models, position-clean deletion moves the target row to the end of the base call before removing it, preserving retained row positions. This corrects the base versus deletion statistics. Pairwise statistics use the unchanged deleted calls and retain any order sensitivity. Appendix B reports the corresponding floors and checks.

We measure TabICLv2 (Qu et al., 2025; 2026), Mitra (Zhang et al., 2025; AutoGluon team, 2025), TabFM (Google Research, 2026), and TabPFN v1, v2.5, and v3 (Hollmann et al., 2023; 2025; Grinsztajn et al., 2025; 2026). The primary datasets are higgs (Baldi et al., 2014) and magic (Bock, 2004). All models receive the same numeric preprocessing and, in matched comparisons, the same support arrays. Unless labelled otherwise, measurements use fp32 and one estimator or view. Mitra uses a deterministic harness with fine-tuning off. TabPFN's *shipped* configurations retain default preprocessing, including the fingerprint feature, but are not the vendor's full ensemble configuration. Table 4 gives checkpoint versions, settings, training data disclosures, and model specific limits. Appendix L lists the measured ranges.

3 Influence and Concentration

For supports differing in one row, Nagler (2023) assumes $|q_\theta(y \mid x, D_n) - q_\theta(y \mid x, D'_n)| \le Ln^{-\alpha}$ for almost all such supports and all sufficiently large n (Eq. 5). If $\alpha > 1/2$, Theorem 5.2 gives vanishing support draw fluctuations on the probability scale. For the specified single layer attention model with bounded inputs and finite weight norms, Theorem 6.2 gives the $n^{-1/2}$ reference rate. The constant is unrestricted by our measurements, and our 12- to 24-block checkpoints lie outside that theorem's architectural scope.

For an order invariant predictor, the deletion pair maximum probes the replace-one difference on observed supports of size $n-1$. Even at $k = n$, it does not establish Eq. 5's uniform bound. The pair mean is a descriptive statistic. Its exponent alone is not the condition in Theorem 5.2. Under the i.i.d. assumptions of Efron–Stein, a root-mean-square replace-one difference decaying as $n^{-\alpha_{\mathrm{rms}}}$ bounds spread by a constant times $n^{1/2-\alpha_{\mathrm{rms}}}$. Thus $\alpha_{\mathrm{rms}} > 1/2$ is also a sufficient route to vanishing spread. Both the assumptions and the empirical checks are given in Appendix C. Probability and margin results remain separate: the logistic link can increase the measured probability scale exponent as predictions approach zero or one.

A slower rate need not imply poor prediction. Under regularity conditions, parametric learners have $O(1/n)$ single row influence. Local learners using fewer effective observations can decay more slowly. Nagler's Theorem 5.4 requires locality for vanishing bias over its function class. For comparison, minimax risk results for transformers on smoothness classes imply a bound on query-averaged SD of $n^{-\beta/(2\beta+d)}$ (Ma et al., 2026; Ching et al., 2026). For β between 2 and 4, the exponent ranges from 0.06 to 0.11 at higgs's $d = 28$ and from 0.14 to 0.22 at magic's $d = 10$. These rates help compare the measurements. They do not identify the computation or smoothness of a shipped checkpoint.

4 Experiments

4.1 Prediction spread across support draws

Over 64–1024 rows, a range within every checkpoint's supported limit, all twelve model–dataset spread intervals lie above $-1/2$ (Table 1).

Logistic regression refit on the same draws reaches the rate on higgs: -0.52, with an interval containing $-1/2$, contrary to our prediction. On magic, the default penalty gives -0.45. Weakening

Table 1: Prediction spread slopes and support resampling 95% intervals on the probability scale. Initial windows use 64–1024 rows. Disjoint higgs windows use 1024–16,384 from a 3,440,379-row pool (†: Mitra ends at 8192). R is draws per size; 48/24 means 24 at 16,384. p: Linux environment; conf.: excluded by the spread to floor rule. Tables 7, 8, and 14 give extensions, configurations, and overlap corrections.

	higgs		magic		higgs, disjoint		
Model	slope	support CI	slope	support CI	slope	support CI	R
Mitra trained	-0.06	$[-0.11, +0.02]$	-0.30	$[-0.34, -0.26]$	$-0.43^{\dagger}$	$[-0.46, -0.39]$	24
TabICLv2 trained	-0.28	$[-0.31, -0.24]$	-0.37	$[-0.41, -0.33]$	-0.10^{p}	$[-0.12, -0.08]$	24
TabPFN v2.5	-0.12	$[-0.16, -0.07]$	-0.34	$[-0.37, -0.30]$	-0.08	$[-0.11, -0.07]$	48/24
TabPFN v3	-0.10	$[-0.15, -0.04]$	-0.38	$[-0.43, -0.33]$	-0.14^{p}	$[-0.16, -0.12]$	24
TabPFN v1	-0.07	$[-0.11, -0.02]$	-0.27	$[-0.30, -0.24]$	-0.41^{p}	$[-0.44, -0.39]$	48/24
TabFM trained	-0.12	$[-0.16, -0.08]$	-0.43	$[-0.47, -0.40]$	-0.12^{p}	$[-0.14, -0.10]$	24
logreg (refit)	-0.52	$[-0.54, -0.50]$	-0.45	$[-0.48, -0.42]$	-0.51	$[-0.53, -0.48]$	48
logreg (refit, C=10^6)	-0.64	$[-0.65, -0.62]$	-0.57	$[-0.60, -0.53]$	-0.51	$[-0.53, -0.49]$	48

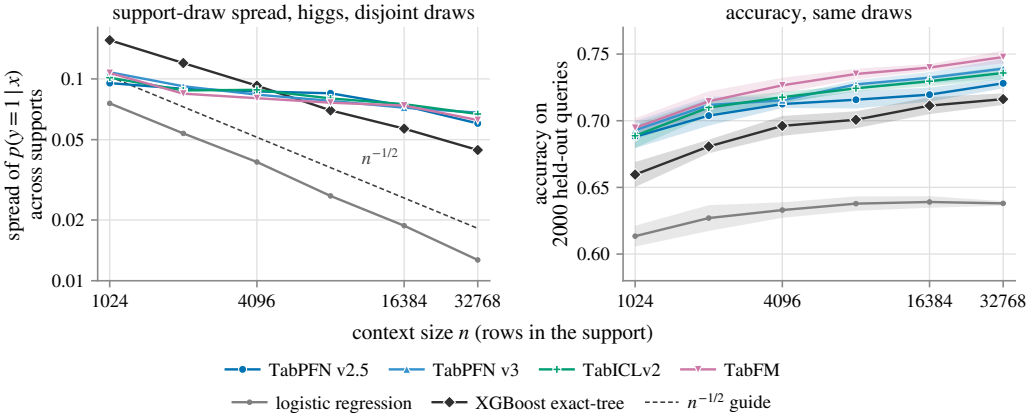


Figure 1: Prediction spread and accuracy on disjoint higgs supports. Left: query mean probability SD. The dashed $n^{-1/2}$ guide passes through the checkpoints' geometric mean level at 1024. XGBoost exact tree clears the $3\times$ order floor rule through 32,768. The original histogram experiment is retained in Figure 3. Right: mean accuracy and 95% t intervals across eight supports, conditional on 2000 fixed queries. Both classical learners are refit on the same draws. Exact tree XGBoost uses 100 depth-4 trees, learning rate 0.1, and one CPU thread (Table 9).

that penalty to $C = 10^6$ or scaling C with n yields intervals below $-1/2$, including under the $\kappa = 1$ overlap sensitivity analysis (Appendix F). The $\sqrt{n}$-nearest-neighbour and XGBoost refits decay more slowly: -0.27 and -0.24 on higgs, both -0.29 on magic, with all intervals above $-1/2$. Five of six checkpoints are flatter than both refits on higgs. TabICLv2's interval overlaps theirs.

Vendor default TabPFN v2.5, TabICLv2, and TabFM slopes differ by at most 0.05 on matched windows. Mitra's seeded stack, with fine-tuning off, reproduces its harness's spread pattern. Table 7 retains the initial cost limited higgs run and its completion. V3 has no vendor default spread run.

On disjoint higgs supports, v2.5, v3, TabICLv2, and TabFM have slopes from -0.08 to -0.14 over 1024–16,384 rows. TabPFN v1 gives -0.41 beyond its 1024-row cap, and Mitra gives -0.43 up to its 8192-row cap. All intervals lie above $-1/2$. Logistic regression gives -0.51 on the matched window and -0.52 over 1024–65,536, the latter interval below $-1/2$. At 16,384 rows, the four checkpoints exceed its accuracy by 0.08 to 0.10 on 2000 held out queries. Mitra exceeds it by 0.07 at 8192 (Figure 1). At the same 16,384-row size, v3's query mean spread is 0.072, against logistic regression's 0.019. Histogram XGBoost reaches accuracy 0.714 there. Exact tree reaches 0.711, versus 0.720–0.740 for the four checkpoints. These are point comparisons, without a checkpoint versus XGBoost accuracy test.

The four checkpoints also have flatter slopes than the disjoint $\sqrt{n}$-NN and histogram XGBoost refits, whose valid extension fits give -0.25 and -0.37. The paired slope difference intervals exclude

zero. V1 and Mitra do not share this ordering (Table 8). The separate exact tree configuration gives -0.362 $[-0.390, -0.355]$ over 1024–32,768, with zero measured order floor at all six sizes. Its paired comparisons and size omission sensitivity are reported in Appendix D. This finite range slope does not establish an asymptotic rate. V3, TabFM, and TabICLv2 retain slopes between -0.10 and -0.16 through 32,768. V2.5, v1, and Mitra steepen late without establishing a transition to a faster rate. Post hoc size omission checks retain all twenty support only intervals for the four principal curves above $-1/2$, despite window dependence (Appendix D). Joint query and support resampling and the stored 2000-query matrices retain the contrast. The latter use only eight supports (Table 6).

On magic, disjoint partitions over 1024–4096 give post hoc slopes -0.45 for v1 and -0.31 for v2.5. Overlapping draws to 8192 appear steeper, but none of the examined overlap adjusted checkpoint intervals lies wholly below $-1/2$ (Table 14). These sensitivities have no established coverage.

On Nagler's synthetic testbed, v1 does recover the reference rate after the bias curve bends: -0.30 over 64–1024, then -0.54 over 1024–4096, with the latter interval below $-1/2$. Noise dimensions delay this transition. The additional real datasets show no corresponding dimension ordering (Spearman -0.50, a point estimate; Appendix D).

A post hoc two-regime fit did not identify a transition to faster decay in v2.5's disjoint higgs series (Appendix D).

4.2 Preprocessing and single row influence

TabPFN's default preprocessing appends a fingerprint column by hashing preprocessed row bytes. Deleting one row refits the scaler and SVD, changing features and hashes of retained rows. At the $n = 256$ diagnostic, no retained fingerprint survived a deletion, even with the hash salt fixed. This support deletion mechanism complements earlier observations of irregular probability functions and limited accuracy gains from joint feature engineering ablations (McCarter, 2025; Ye et al., 2025). Deterministic hashing, available from package 6.4.0 onward, does not prevent the response to a changed support.

The controlled common query experiment measures one v2.5 checkpoint on two datasets, six support seeds, and four sizes (64–512), with the first 50 unfiltered queries per dataset fixed across all configurations. All 192 measurements were completed. Disabling only the fingerprint flag steepens mean pair decay on both scales. All four primary paired contrasts survive Holm correction (Table 2). The exploratory frozen stack configuration D also steepens decay while preserving retained fingerprints and query transforms exactly. Fixing only the salt does not preserve fingerprints. Its exponent contrasts are inconclusive (Appendix G.1). Accuracy change intervals span zero, and log-loss point estimates worsen on both datasets: faster influence decay establishes neither better prediction nor equivalence. These estimates condition on the fixed query sets. The historical filtered query comparisons and broader frozen stack corroboration are in Appendix G.

Table 2: Prospective common-query TabPFN v2.5 follow-up: primary mean-pair influence exponents over $n = 64, 128, 256, 512$, $k = 32$, six paired seeds, and the same 50 unfiltered queries per dataset. Exponents are means (SE); Δ is fingerprint off minus shipped. Intervals are ordinary t_5 intervals, not simultaneous. The four two-sided contrasts form one Holm family. CUDA fp32, one estimator; all 96 primary and 96 exploratory-control cells completed. Larger α means faster decay in this window, not an asymptotic guarantee or improved accuracy.

Dataset	Scale	Shipped α	Off α	Paired Δ [95% CI]	Holm p
higgs	probability	0.383 (0.096)	0.611 (0.110)	+0.228 [+0.152,+0.304]	0.00105
higgs	margin	0.369 (0.124)	0.608 (0.141)	+0.239 [+0.161,+0.317]	0.00105
magic	probability	0.574 (0.023)	1.128 (0.051)	+0.554 [+0.412,+0.695]	0.00062
magic	margin	0.272 (0.049)	0.822 (0.046)	+0.550 [+0.411,+0.688]	0.00062

Across twelve additional datasets, historical seed paired conditional sampled maximum changes are positive (median $+0.33$). Seven fingerprint-off exponents exceed $1/2$ at 3 SE, and nine do so for mean deletion influence. Table 27 retains all results and Appendix H the two failed predictions. The distinction between sampled and all-deletion maxima matters: v2.5's fingerprint-off magic exponent falls from 0.73 at $k = 32$ to 0.22 at $k = n$, while its pair mean remains above $1/2$.

Spread measurements require no deletion, but their predictions still use the preprocessing stack. The fingerprint-off and shipped spread measurements agree over $n \leq 1024$ on both primary datasets and on the higgs extension to 8192. On magic, however, the fingerprint reduces spread beyond 2048, steepening the extension fit from -0.43 to -0.51. The two largest disjoint higgs sizes have no fingerprint-off comparison. The same hash mechanism amplifies row order dependent floating-point differences in the SVD and scaler. Appendix G reports the associated preprocessing interventions.

Mitra illustrates a separate inference stack confound. With AutoGluon 1.5.0's stochastic settings, two completed seeds gave a Q_{mean} slope of $+0.58$. These measurements are excluded. Unseeded augmentation, per-call support permutations, and bfloat16 arithmetic produced a noise floor with slope $+0.68$, reaching the signal by $n = 256$. Removing those stages, with weights unchanged, gave exact repeat determinism and order noise 4.7×10^{-6} (Appendix G).

4.3 Historical single row influence

Table 3 separates mean pairwise influence from the exact maximum over deletion pairs. The mean is insensitive in expectation to the number of sampled deletions. The maximum is not. The registered primary statistic, the $k = 32$ sampled maximum, remains in Table 15, with both full and registered fit windows. Fingerprint-off rows change one preprocessing flag in the same checkpoint. Whether that feature was used in pretraining is undisclosed.

TabICLv2's margin scale pair mean exponent is above $1/2$ on both datasets. The other checkpoints' higgs fits are unclassified ($|t| \leq 2.7$). Magic separates shipped TabPFN below $1/2$ from fingerprint-off TabPFN, Mitra, and TabFM above it at the uncorrected 3-SE threshold. Among mapped magic mean type checkpoint comparisons in this table, only TabICLv2 survives Holm correction (Appendix E).

At the margin scale exact deletion maximum, every checkpoint configuration in Table 3 is below $1/2$ on magic, including fingerprint-off TabPFN. On higgs, shipped v2.5 and v3 are below. The other five rows are unclassified, with Mitra measured only over 64–256. Logistic regression's magic maximum exponent also falls, from 0.98 at $k = 32$ to 0.65 at $k = n$, missing our predicted band. Thus part of the decrease comes from maximizing over more rows, although the checkpoint exponents remain 0.3 to 0.6 below this anchor. The completed histogram XGBoost and MLP all-deletion maxima are also below $1/2$ on both datasets under their separate four-test correction (Table 19). Slow maximum influence decay is therefore not confined to foundation models.

On the probability scale, only shipped v2.5 classifies below $1/2$ among magic's checkpoint all-deletion maxima. TabFM classifies above, outside the mapped family (Table 18). Magic pair means instead approach 1 for fingerprint-off TabPFN, TabICLv2, Mitra, and TabFM, versus about $1/2$ for shipped v2.5/v3 (Table 17). Appendix C reports RMS comparisons, including TabICLv2's probability exponent 0.74 on the spread experiment's own higgs queries over 512–2048. Larger sizes were not measured.

The shorter registered fits generally retain the direction of full window classifications but lose precision. Only 27 of the 154 mapped comparisons survive the Holm correction. The shipped TabPFN pair mean and sampled maximum below boundary classifications do not survive correction on their own. Mitra's higgs curve is not a power law, and its full window exponent does not replicate on magic. Tests did not support the proposed explanations based on readout type, layer criticality, depth, or prior mismatch. Influence decay also failed to predict label flip attack cost (Appendix I).

4.4 Weight randomized controls

We compared trained checkpoints with copies whose tensor entries were shuffled, preserving architecture and each tensor's Frobenius norm. Full scope shuffles affect every tensor. Trunk scoped shuffles leave input embedders trained. Every family has deletion measurements for at least one shuffle scope, and Mitra and TabICLv2 also have spread measurements (Appendix J). On higgs, two copies have spread slopes at or below $-1/2$, with intervals containing it. On magic, both intervals contain $-1/2$ and the copies decay faster than their trained counterparts (Table 7).

These copies do not provide a clean architectural baseline. Across the five copies tested by flipping every support label, the median absolute margin response is 3.9–252$\times$ smaller than for trained

Table 3: Margin scale influence under historical query selection, mean (SE) over six seeds, 64–512 rows (64–256 for Mitra-higgs at $k = n$). Left: mean pair influence at $k = 32$. Right: the maximum over all observed deletion pairs, not a global replacement supremum. t compares unrounded exponents with $1/2$. Bold marks $|t| \geq 3$. h: survives the mapped Holm correction; p: Linux environment; ‡: Mitra's window sensitive higgs fit. Fingerprint-off minus shipped rows give seed paired conditional differences (SE), with t against zero: each configuration selects its own correctly classified queries. Unavailable cross-machine pairs are dashed. Classical models are refit per support. The registered XGBoost histogram and MLP all-deletion completion is detailed in Table 19. Appendix E gives configurations, multiplicity, and the registered $k = 32$ maximum results. Table 17 gives probability scale means.

		mean $\alpha(Q_{\text{pair}})$, margin, $k = 32$			all-deletion $\alpha(Q_{\text{pair}}^{\max})$, $k{=}n$ (t)	
Model	config.	higgs	magic	t vs $\frac{1}{2}$	higgs	magic
TabPFN v2.5	fp off	0.563 (0.135)	0.830 (0.053)	$+0.5/+\mathbf{6.3}$	0.22 (-2.0)	0.22 ($-\mathbf{5.0}$)
TabPFN v2.5	shipped	0.374 (0.123)	0.285 (0.046)	$-1.0/-\mathbf{4.7}$	0.06 ($-\mathbf{4.4}$)	0.03 ($-\mathbf{21.2}^{\text{h}}$)
TabPFN v3	fp off	0.536^{p} (0.122)	0.813 (0.037)	$+0.3/+\mathbf{8.6}$	0.17^{p} (-2.5)	0.25^{p} ($-\mathbf{4.1}$)
TabPFN v3	shipped	0.252^{p} (0.090)	0.313 (0.058)	$-2.7/-\mathbf{3.2}$	0.10^{p} ($-\mathbf{5.0}$)	0.09^{p} ($-\mathbf{8.9}^{\text{h}}$)
TabPFN v2.5	fp off minus shipped	$+0.19$ (0.02)	$+0.55$ (0.06)		$+0.16$ ($+2.8$)	$+0.18$ ($+3.8$)
TabPFN v3	fp off minus shipped	$+0.28^{\text{p}}$ (0.09)	$+0.50$ (0.04)		$+0.07^{\text{p}}$ ($+0.5$)	$+0.17^{\text{p}}$ ($+4.7$)
TabICLv2	1 estimator	1.021 (0.064)	0.973 (0.039)	$+\mathbf{8.1}/+\mathbf{12.3}^{\text{h}}$	0.55 ($+1.2$)	0.33 ($-\mathbf{4.4}$)
Mitra	det. harness	$0.339^{\ddagger}$ (0.137)	0.777 (0.061)	$-1.2/+\mathbf{4.6}$	-0.05 (-2.8)	0.17 ($-\mathbf{6.6}$)
TabFM	1 view	0.645 (0.085)	0.764 (0.082)	$+1.7/+\mathbf{3.2}$	0.31^{p} (-2.3)	0.31 ($-\mathbf{3.4}$)
logreg (refit)		1.248 (0.028)	0.992 (0.062)	$+\mathbf{26.4}^{\text{h}}/+\mathbf{7.9}$	1.09 ($+\mathbf{27.0}$)	0.65 ($+1.9$)
XGBoost (refit)	hist.	0.145 (0.035)	0.068 (0.031)	$-\mathbf{10.2}^{\text{h}}/-\mathbf{14.1}^{\text{h}}$	0.09 ($-\mathbf{17.3}$)	0.06 ($-\mathbf{22.0}$)
MLP (refit)		-0.031 (0.035)	0.397 (0.056)	$-\mathbf{15.2}^{\text{h}}/-1.8$	-0.03 ($-\mathbf{15.7}$)	0.29 ($-\mathbf{3.3}$)

models, and accuracy stays between 0.485 and 0.51 (Figure 5). The two higgs copies reaching the reference rate vary about $20\times$ and $160\times$ less across supports at $n = 64$ than their trained counterparts. Their concentration therefore accompanies weak label response and near-chance prediction. Related studies also find that in-context learning emerges with pretraining (Olsson et al., 2022; Garg et al., 2022; Chan et al., 2022). The observed exponent differences cannot separate learned and architectural contributions.

Label response alone is insufficient as a validity check. Convex mixtures of trained and shuffled Mitra weights can remain within the registered factor of two of the trained label flip response, reach the reference spread rate, and classify near chance (Table 13). The original copies had no registered label response pass threshold. Their failure as attribution controls is a descriptive conclusion. Comparisons among trained checkpoints do not require these controls.

4.5 Twelve additional datasets

A registered selection rule chose twelve further OpenML datasets and fixed predictions before measurement. All twelve failed TabPFN's order probe and used position-clean deletion. Spread was measured for TabPFN v2.5 and TabICLv2 on the same 24 draws at each size over 64–512 (Appendix H).

V2.5 is slower than the reference rate on ten datasets, with every interval above $-1/2$. It reaches the rate on pol, with its interval below $-1/2$, and at the point estimate on wilt, whose interval contains $-1/2$ and whose overlap corrected slope is -0.49. TabICLv2 also has slower point estimates on ten, although jm1's interval contains the boundary, and intervals below $-1/2$ on pol and wilt. The two models' dataset rankings have Spearman correlation 0.52, a point estimate (Table 24).

Pol and wilt have the highest accuracy and lowest probability spread in this set. V2.5's margin scale spread is flat on pol and decays more slowly than $n^{-1/2}$ on wilt, with both intervals above $-1/2$. This scale difference is consistent with the shrinking logistic link factor $p(1-p)$ at accuracy near 0.97. TabICLv2's margin spread was not measured. The $\sqrt{n}$-NN refit does not reach the rate on either dataset. XGBoost does on pol.

For v2.5's shipped preprocessing deletion experiment, nine of twelve margin scale sampled maximum exponents are below $1/2$, four at 3 SE. None is classified above. The point estimate exceptions are jm1, numerai28.6 (near chance), and covertype. Probability scale exponents are higher on nine

datasets, with six at or above $1/2$ and none classified below. Pol is classified above at 4.2 SE. The logistic link diagnostic accounts for 96% of its mean influence scale gap (Appendix H).

The trained models' spread rankings do not meet the predicted correlation of $+0.5$ with the $\sqrt{n}$-NN refit: across fourteen datasets, the correlations are $+0.35$ and $+0.34$. Default penalty logistic regression reaches the rate at the point estimate on only two of the twelve additional datasets, churn and pol, against nine predicted. Only churn's interval lies below $-1/2$. XGBoost is slower on eleven (Table 25). Thus the short window also limits concentration of the parametric reference on most of these datasets.

5 Limitations and conclusion

These results support specific precautions when measuring context dependence. Mitra's augmentation and precision settings must be controlled. Seeded vendor stack measurements cover spread, not deletion influence. Repeat determinism alone does not detect the fingerprint effect, and shuffled weights do not provide a valid attribution baseline when label response and accuracy collapse.

Slow concentration matters when extrapolating uncertainty across context sizes (Fortini et al., 2026). For example, scaling an interval width from 1024 to $16{,}384$ rows by $n^{-1/2}$ gives a width about $2.7\times$ smaller than scaling by v3's fitted $n^{-0.14}$ spread rate. This is a rate-based illustration, not a measured coverage failure. Nor does prediction spread imply equally variable accuracy: at $16{,}384$ higgs rows, only v2.5 has an across support accuracy SD above logistic regression's 0.0052 (0.0063), contrary to the registered prediction (Table 12).

The study covers binary classification, six checkpoints, and four families. Higgs and magic support the main comparisons. Only two checkpoints were tested for spread on all fourteen datasets. The weight mixtures and label flip diagnostics cover higgs only. Spambase is excluded from the main comparisons after a registered control band failure. Bank's deviation disappears at a matched window, and TabPFN v2 was not measured. Training data disclosures are incomplete. V2.5's default checkpoint was fine-tuned on real datasets including higgs and the source of eye_movements. Its synthetic only counterpart gives similar measurements on the registered comparisons, with the failed accuracy prediction retained (Appendix B).

Fitted exponents summarize the stated ranges. Mitra's higgs influence curve is not a power law, with local slopes spanning 1.36. Pair differences remove a common additive length response but retain multiplicative margin drift, measured up to $2.4\times$. A scale normalized statistic was not run. Finite deletion probes cannot establish a global stability guarantee (Kim & Barber, 2023). Shipped TabPFN measurements differ across machines by up to 0.40 relative, and one full window classification changes with platform (Appendix K). No checkpoint was measured beyond $32{,}768$ rows, and the study does not cover regression or models retrained in response to these measurements.

For four checkpoints on disjoint higgs supports, higher measured accuracy coexists with slow spread decay over 1024–$16{,}384$ rows. The fingerprint intervention resolves the selection confound for its fixed 50-query sets. It does not establish population wide effects or explain the spread rates. Its frozen stack control is broader, and its six-seed uncertainty is conditional on the finite inputs. The supplement includes the 48 original TabICLv2 RMS deletion matrices, which allow those exponents to be checked from predictions. Other historical foundation model influence results remain reproducible from stored measurement summaries, but their raw deletion matrices are not in the submission ZIP. Larger context behavior and attribution of spread to internal computation remain outside the conclusions supported by these measurements.

AI use statement

LLMs assisted with code implementation and debugging, literature searches, critique of experimental validity, and drafting/polishing parts of the appendix and supplementary material. The authors reviewed these contributions and take responsibility for the final paper.

Reproducibility Statement

The supplementary code and run reference bundle provide measurement scripts, stored results, registrations, environment specifications, and a claim to run index. The code README gives commands for checking the bundle and reproducing the principal tables and figures from stored results. The common query follow-up includes genuine base/deletion predictions, feature arrays and identities, frozen protocols, source snapshots, and offline analysis commands. Rerunning the model separately requires licensed weights. Appendix B describes the models and data. Appendix L states the measured ranges. Appendix K reports numerical reproducibility and gaps in the records. The theoretical assumptions and aggregation proof are in Appendix C.

References

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Fracesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanhan, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel Bowman, Ethan Perez, Roger Grosse, and David Duvenaud. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 129696–129742, 2024. doi: 10.52202/079017-4121.

Usman Anwar, Johannes von Oswald, Louis Kirsch, David Krueger, and Spencer Frei. Understanding in-context learning of linear models in transformers through an adversarial lens. *Transactions on Machine Learning Research*, 2025. arXiv:2411.05189.

Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah D. Goodman. Bayesian scaling laws for in-context learning. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2410.16531.

AutoGluon team. Mitra: A tabular foundation model (`autogluon/mitra-classifier`). Hugging Face model card, https://huggingface.co/autogluon/mitra-classifier, 2025. Model card last updated 19 November 2025; accessed 3 September 2026.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:4308, 2014. doi: 10.1038/ncomms5308.

Marin Biloš, James T. Wilson, Anderson Schneider, and Yuriy Nevmyvaka. A mechanistic study of tabular foundation models. *arXiv preprint arXiv:2605.21288*, 2026.

R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2004. URL https://doi.org/10.24432/C52C8B.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18878–18891, 2022. doi: 10.52202/068431-1371. arXiv:2205.05055.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7654–7664, 2022. doi: 10.18653/v1/2022.acl-long.527. arXiv:2202.12172.

Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and minimax optimal in-context nonparametric regression with transformers. *arXiv preprint arXiv:2601.15014*, 2026.

Mohamed Djilani, Thibault Simonetto, Karim Tit, Florian Tambon, Salah Ghamizi, Maxime Cordy, and Mike Papadakis. On the robustness of tabular foundation models: Test-time attacks and in-context defenses. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2026. arXiv:2506.02978.

Fabian Falck, Ziyu Wang, and Christopher C. Holmes. Is in-context learning in large language models Bayesian? A martingale perspective. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12784–12805. PMLR, 2024. arXiv:2406.00793.

Minghong Fang, Neil Zhenqiang Gong, and Jia Liu. Influence function based data poisoning attacks to top-N recommender systems. In *Proceedings of The Web Conference 2020 (WWW)*, pp. 3019–3025. ACM, 2020. doi: 10.1145/3366423.3380072. arXiv:2002.08025.

Sandra Fortini, Kenyon Ng, Sonia Petrone, Judith Rousseau, and Susan Wei. Uncertainty decomposition for Bayes-filtered transformers via Bayesian predictive inference. *arXiv preprint arXiv:2602.04596*, 2026.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 30583–30598, 2022. doi: 10.52202/068431-2217. arXiv:2208.01066.

Google Research. TabFM 1.0 (`google/tabfm-1.0.0-pytorch`). Model card and code, https://huggingface.co/google/tabfm-1.0.0-pytorch, https://github.com/google-research/tabfm; announced in the Google Research blog post of 30 June 2026; no technical report as of September 2026, 2026.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, Clara Cornu, Lilly Charlotte Wehrhahn, Alessandro Bonetto, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306. arXiv:1906.06755.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2207.01848.

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*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

James Hu and Mahdi Ghelichi. Noise immunity in in-context tabular learning: An empirical robustness analysis of TabPFN's attention mechanisms. *arXiv preprint arXiv:2604.04868*, 2026.

Byol Kim and Rina Foygel Barber. Black-box tests for algorithmic stability. *Information and Inference*, 12(4):2690–2719, 2023. doi: 10.1093/imaiai/iaad039.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, 1965.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

Mykhailo Koshil, Thomas Nagler, Matthias Feurer, and Katharina Eggensperger. Towards localization via data embedding for TabPFN. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.

Mary Letey, Yue M. Lu, Cengiz Pehlevan, and Jacob Zavatone-Veth. Sequential correlations change in-context learning: Effective context length and architectural mismatch. In *ICML 2026 Workshop on High-dimensional Learning Dynamics (HiLD)*, 2026. arXiv:2607.03660.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2301.07067.

Hejia Liu, Mochen Yang, and Gediminas Adomavicius. Robustness is important: Limitations of LLMs for predictions on tabular data. *PNAS Nexus*, 5(6):pgag197, 2026. doi: 10.1093/pnasnexus/pgag197. arXiv:2508.19563.

Si-Yang Liu and Han-Jia Ye. TabPFN unleashed: A scalable and effective solution to tabular classification problems. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40043–40068, 2025. arXiv:2502.02527.

Tianyi Ma, Tengyao Wang, and Richard J. Samworth. Optimal in-context adaptivity and distributional robustness of transformers. *arXiv preprint arXiv:2510.23254*, 2026. v3, May 2026.

Calvin McCarter. What exactly has TabPFN learned to do? *arXiv preprint arXiv:2502.08978*, 2025. Originally an ICLR 2024 Blogposts Track post, https://iclr-blogposts.github.io/2024/blog/what-exactly-has-tabpfn-learned-to-do/; the arXiv version adds the TabPFN-v2 appendix.

Valentyn Melnychuk, Vahid Balazadeh, Stefan Feuerriegel, and Rahul G. Krishnan. Frequentist consistency of prior-data fitted networks for causal inference. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.12037.

Ratmir Miftachov, Bruno Charron, and Simon Valentin. Inspectable tabular foundation models via in-context kernel learning. In *ICML 2026 Workshop on Foundation Models for Structured Data (FMSD)*, 2026. arXiv:2602.02162 (the arXiv version's title: Interpretable Tabular Foundation Models via In-Context Kernel Regression).

Samuel Müller, Noah Hollmann, and Frank Hutter. Bayes' power for explaining in-context learning generalizations. *arXiv preprint arXiv:2410.01565*, 2024.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2305.11097.

Tai Nguyen and Eric Wong. In-context example selection with influences. *arXiv preprint arXiv:2302.11042*, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Matthew Peroni, Franck Le, and Vadim Sheinin. Robust tabular foundation models. In *AAAI 2026 Workshop on Shaping Responsible Synthetic Data in the Era of Foundation Models (RSD)*, 2026. arXiv:2512.03307.

Prior Labs. TabPFN-3 changelog. https://docs.priorlabs.ai/changelog/tabpfn-3, 2026. Accessed 2026-09-01.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847, 2025. arXiv:2502.05564.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.11139.

David Rundel, Julius Kobialka, Constantin von Crailsheim, Matthias Feurer, Thomas Nagler, and David Rügamer. Interpretable machine learning for TabPFN. In *Explainable Artificial Intelligence (xAI 2024)*, volume 2154 of *Communications in Computer and Information Science*, pp. 465–476. Springer, 2024. doi: 10.1007/978-3-031-63797-1_23. arXiv:2403.10923.

Jake A. Soloff, Rina Foygel Barber, and Rebecca Willett. Bagging provides assumption-free stability. *Journal of Machine Learning Research*, 25(131):1–35, 2024. arXiv:2301.12600.

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. Poisoning attacks on LLMs require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025.

Tongxi Wang and Zhuoyang Xia. Stability of in-context learning: A spectral coverage perspective. *arXiv preprint arXiv:2509.20677*, 2025.

Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at TabPFN v2: Understanding its strengths and extending its capabilities. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 150356–150388, 2025. doi: 10.52202/085713-4527. arXiv:2502.17361.

Di Zhang, Ningxu Zhang, and Zimeng Liu. Bigger is safer: Provable robustness in in-context learning scales with capacity. *arXiv preprint arXiv:2602.17743*, 2026.

Xiyuan Zhang, Danielle C. Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W. Mahoney, Cuixiong Hu, Huzefa Rangwala, George Karypis, and Bernie Wang. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 17831–17876, 2025. doi: 10.52202/085713-0535. arXiv:2510.21204.

A Related work

Closest empirical precedents. Nagler's Section 6.5 and Figure 1 measure bias and variance using 500 simulated datasets and 100 test queries with TabPFN v1. Koshil et al.'s Section 4.1 and Figure 1 extend this analysis to 1000 simulated datasets, 1024 test queries, and contexts through 8192, including embedding-based localization. Liu and Ye's Section 2.3 and Figure 2 study error, bias and variance on Adult and Bank through 12,800 contexts. Thus neither context size variance analysis nor real data evidence is new by itself. McCarter's Appendix A.1 attributes bumpy predicted probability functions in simple TabPFN-v2 examples to default fingerprints. Ye et al.'s Appendix D jointly ablates fingerprints and polynomial features. Its limited benchmark benefit is neither a fingerprint only ablation nor equivalence evidence. We credit those observations and examine the more specific path from support dependent preprocessing to retained row hash changes under deletion.

Biloš et al. (2026) identify TabPFNv2 and Mitra as attention weighted label voters at a late layer, and TabICLv2 as a nearest class prototype reader. They validate these mechanisms causally, demonstrate attacks, and compare refit baselines at essentially fixed context size. Their Appendix J varies n from 1000 to 1500 as a stability check. They do not measure single row influence decay or a breakdown point. Their Mitra label fails the registered transfer test on our checkpoint (Appendix I). We measured TabPFN v2.5 rather than their v2 checkpoint, and our rate comparisons require no readout classification.

For text models, Wang & Xia (2025) define stability through prediction divergence between independently drawn demonstration sets, closely related to our spread measurement, but report threshold crossings and divergence quantiles at $K \leq 32$. Influence-based example selection also uses a fixed demonstration count (Nguyen & Wong, 2023). Training data influence methods estimate a related object through gradients (Koh & Liang, 2017). Our deletions use direct prediction calls. Bayesian scaling laws for text in-context learning fit performance rather than influence (Arora et al., 2025). Li et al. (2023) derive generalization results from algorithmic stability under attention weight constraints, without measuring a tabular model exponent. Falck et al. (2024) test Bayesian uncertainty contraction in language models and find violations. Our study measures context dependence in tabular checkpoints and evaluates the validity of weight randomized controls.

Near-constant poisoning budgets in language model pretraining (Souly et al., 2025) and many-shot jailbreaking (Anil et al., 2024) concern related questions about how data effects change with scale. Letey et al. (2026) derive an effective context length for correlated prompts in linear attention. Their mechanism is prompt correlation rather than the single row sensitivity measured here. Anwar et al. (2025) find breakdown point fragility for linear attention and GPT-2-style transformers on synthetic linear regression.

B Method and protocol

B.1 Influence statistics

For each query, the deletion response d_i defined in Section 2 gives Q_{mean}, $Q_{\max}$, and $\mathrm{CV}^2(q) = \mathrm{Var}_i(d_i)/(\mathrm{mean}_i d_i)^2$. The runners average $\mathrm{CV}^2(q)$ over queries, then report $n_{\text{eff}} = n/(1 + \mathrm{mean}_q\, \mathrm{CV}^2(q))$ using Kish's form (Kish, 1965). This is not the query mean of $n/(1 + \mathrm{CV}^2(q))$. For position-clean TabPFN, the stored CV statistics retain the original base responses. Only Q_{mean} and $Q_{\max}$ are replaced by their position-clean values. Exponents are fitted per seed by OLS on log statistic versus log n and summarized by their mean and SE (six seeds, or three for the $k = 128$ runs). Tables explicitly distinguish signed slopes from their negation, α.

(i) Uniform row sampling makes Q_{mean} unbiased for the full support mean at any k. The usual choice is $k = 32$.

(ii) A sampled $Q_{\max}$ underestimates the maximum over all rows. Its bias depends on k, which is held fixed across n and models except in labelled exact maximum runs.

(iii) Entropy and participation statistics on a k-subsample have ceilings determined by k (entropy at $\log k$). The CV-based summary avoids that fixed ceiling, although its finite sample estimate can still depend on k. Its fitted exponent separates no pair of trained models (Appendix J).

(iv) A sampled maximum lower-bounds the exact maximum at each n, but this level ordering alone does not order their fitted slopes. If the influential tail broadens with n, a fixed-k fit can overstate decay. The 5-NN rerun illustrates the effect (Appendix F). The $Q_{\max}$ and $Q_{\text{pair}}^{\max}$ exponents agree within 1.5 pooled SE except in five runs: extended shuffled Mitra (2.25), extended refit MLP (3.1), extended XGBoost (1.65), and the two shipped v2.5 exact maximum magic runs (2.1 under macOS, 2.0 in the Linux environment). Every other trained foundation model run agrees within 1.4 pooled SE.

(v) In historical influence runs, queries are the first 50 correctly classified by the measured model from the sampled candidate pool. Both the pool and filter are renewed for each (n, seed) measurement. The shipped and fingerprint-off runs apply this correctness filter separately (`fixed_queries=false`). Pairing (n, seed) does not pair the selected queries. Across 64–512, stored base margin sign sequences differ in 23 of 24 higgs cells and 20 of 24 magic cells, proving unequal ordered selections in those cells. Matching signs in the remaining cells do not prove matching identities. Historical arm D is an exception: it uses arm 0's selected queries while freezing the fitted preprocessing during deletion. The registered fixed query check instead uses the $(n = 64, \text{seed}\,0)$ selection throughout (Appendix E).

(vi) With six seeds, t_5 rather than normal critical values apply; 3 SE corresponds to approximately $p = 0.03$ two-sided. Historically we describe differences of at least 3 pooled SE

as separations. The common query contrasts instead use their four-test Holm family (Section G.1).

Context length. Deleting one row changes both the support content and its length, $n \to n-1$. An additive length response $m(n-1) - m(n)$ affects every deletion and can dominate absolute responses. For example, $m(n) \propto \sqrt{n}$ produces a common response of order $n^{-1/2}$, even without a slowly decaying row specific effect. Such a response can also make influence look uniform across rows and produce $n_{\text{eff}} \sim n$.

The pair statistic compares predictions on two supports of size $n-1$ differing in one row, matching the replace-one structure of Nagler's Eq. (5) and classical stability (Bousquet & Elisseeff, 2002). For order sensitive predictors, the deleted calls can also differ in retained row positions. The pair statistic cancels a common additive length response without an additional forward pass. The maximum at $k = n$ is exact over these deletion pairs, while Eq. (5) ranges over replacement datasets more generally. A synthetic common shift check changes Q_{mean} linearly and leaves Q_{pair} unchanged to four decimals. For the measured higgs runs, the pair mean exponent is 0.339 ± 0.137 for Mitra versus 0.344 for Q_{mean}, and 1.021 for TabICLv2 versus 1.057. Control exponents are unchanged. Table 3 uses pair statistics. Other tables name Q_{mean} or $Q_{\max}$ where used.

The mean signed deletion response estimates the common component within each measurement. Its magnitude contributes 24%–46% of Q_{mean} on trained higgs runs and up to 67% on trained TabPFN on magic. The corresponding shares are 14–27% for randomized TabICLv2, TabFM, and Mitra, and 61% for randomized TabPFN. For trained Mitra, the common component has slope about -0.15 when fitted to mean levels, or -0.08 ± 0.17 over per-seed fits. The similar pair statistic exponents above show that this component does not account for the measured influence pattern by itself.

B.2 Numerical checks

Order sensitivity. If the model is order sensitive at the scale of the signal, deletion (which reindexes rows) is confounded. TabICLv2: mean per-query margin SD over 20 support permutations is 6.5×10^{-6} (max 1.5×10^{-5}) against maximum influence 0.43. Mitra, under the deterministic harness of Section 4.2: exact repeat determinism and order SD 4.7×10^{-6}.

Control validity. An interpretive control must retain a response to its context labels. Every tested randomized copy has a much smaller label flip response than its trained counterpart, preventing attribution of the exponent difference to learned structure alone (Section 4.4). Copies permute each float tensor within a stated scope, preserving its values while changing their arrangement. Appendix J lists the scopes and measurements. Appendix K describes the supporting records.

Determinism and precision. We check that repeated fit plus predict calls on identical arrays agree and that arithmetic noise is below the influence signal. The original Mitra stack failed these checks: unseeded augmentation, fresh support permutations, and bfloat16 autocast produced call noise about 60 percent of the signal at $n = 64$, increasing with slope $+0.68$. Augmentation contributed most of that noise (Appendix G).

B.3 Models and data

Order sensitivity and datasets. V2.5's order floor reaches 6.8×10^{-2} on higgs. All twelve additional datasets also fail the order check. Magic's matched size floors are at most 3.2×10^{-4}, with four of five below 10^{-5}. V1 is invariant at floating-point precision. The v2.5 effect arises from hashing small order dependent changes in fitted SVD and scaling (Appendix G). Liu et al. (2026) also report TabPFN order sensitivity. Higgs mean floors at 64–1024 are 0.018, 0.038, 0.022, 0.009, and 0.007, or 12–33% of matched size pair influence through 512. Moving one row to the end has an effect comparable to a full permutation. Scheme B was selected before rerunning the TabPFN influence cells. The earlier randomized cell is the sole tabulated exception. Pair statistics reproduce bit-for-bit because they depend only on deleted calls. Scheme B changes the base to deletion comparison, giving higgs $\alpha(Q_{\text{mean}}) = 0.305$ versus 0.299. A quadrature subtraction of the matched floor changes $\alpha(Q_{\text{pair}})$ by 0.016 ($0.374 \to 0.358$) and $\alpha(Q_{\text{pair}}^{\max})$ by 0.001, passing the registered check. TabPFN's `fit()` reloads the checkpoint, so its randomized arm reapplies the fixed weight permutation after every fit and checks both disagreement and repeat determinism.

Table 4: Checkpoints, environments, measurement configurations, and datasets. The order check reports the largest per-query margin SD over 20 support permutations at the specified size. Unless labelled as vendor defaults, measurements use fp32, one estimator or view, and exact repeat determinism. Appendix L lists the grids.

Model	Checkpoint, package, torch	Configuration	Datasets and grids	Order probe
TabICLv2 (Qu et al., 2025; 2026)	`tabicl-classifier-v2-20260212` (Hub `4dcd344e`); `tabicl` 2.1.1; torch 2.11.0	3 column, 3 row, 12 ICL blocks; $n_{\text{estimators}}$=1 (no ensemble aggregation); MPS	higgs, magic, spambase; adult, bank, credit-g (k=128 sweep); twelve additional datasets (spread); 64–512, extended to 2048; spread to 1024; disjoint higgs to 32,768 (Linux)	1.5×10^{-5}: pass
Mitra (Zhang et al., 2025; AutoGluon team, 2025)	`autogluon/mitra-classifier` (Hub `c425e9fa`); `autogluon-tabular` 1.5.0; torch 2.9.1	12-layer Tab2D; deterministic harness (no augmentation, caller row order, fp32 CPU); fine-tuning off	higgs, magic, spambase; 64–512, higgs to 2048; spread to 1024; disjoint higgs to 8192 (the package's support cap)	4.7×10^{-6}: pass
TabFM (no technical report; Google Research, 2026)	`google/tabfm-1.0.0-pytorch`; `tabfm` at upstream `b8a8b09` (v1.0.1 plus 14 commits; the PyPI 1.0.0 wheel lacks the dtype keyword); torch 2.13.0	3 column, 3 row, 24 ICL blocks, 1.64B parameters; single view, no feature shuffle, no class shift, norm "none"; fp32 (`dtype=None`); MPS	higgs, magic; 64–512, scope ladder arms to 2048; spread to 1024; disjoint higgs to 32,768	2.7×10^{-6}: pass
TabPFN v2.5 (Hollmann et al., 2023; 2025; Grinsztajn et al., 2025)	`tabpfn-v2.5-classifier-v2.5_default.ckpt` (10.7M; the package's default classifier file, which the vendor's card lists as fine-tuned on real data and byte-identical to `v2.5_real-large-samples-and-features`; the synthetic only classifier is `v2.5_default-2`, measured in the synthetic only comparison below); `tabpfn` 8.2.0 (the fingerprint salt $n \times d$ since 6.4.0); torch 2.13.0	$n_{\text{estimators}}$=1, inference precision float32; MPS; shipped default and fingerprint-off	higgs, magic, spambase, twelve additional datasets (deletion and spread); 64–512, extended to 2048; spread to 8192; disjoint higgs to 32,768	**fails**: 3.5×10^{-2} at the n=256 check size on higgs (worst 6.8×10^{-2} at n=128) and on all twelve additional datasets; magic $\leq 3.2 \times 10^{-4}$
TabPFN v3 (Grinsztajn et al., 2026)	`tabpfn-v3-classifier-v3_default.ckpt` (53.2M by state dict count); the same package, settings and device	as v2.5	magic (MPS); higgs (Linux)	shares the fingerprint step; rerun position-clean
TabPFN v1	pip 0.1.8, torch 2.5.1 (its own pinned environment; checkpoint from the repository's v1.0.0 tag, pinned by SHA-256)	default	spread arms on higgs and magic to 8192 (past its 1024-row cap, the package's limit); Nagler's testbed; disjoint higgs (Linux)	1.7×10^{-6}: order invariant at float level
refits	scikit-learn logistic regression ($C = 1$, $C = 10^6$, $C = n$), XGBoost (100 trees), $\sqrt{n}$-NN, 5-NN, MLP, random forest	refit per support on the stored draws and queries	as the arm they anchor	n/a

All models receive identical numeric arrays. Primary datasets are higgs (OpenML 23512 v2, the 50,000-row UCI HIGGS simulation subsample) and magic (1120 v1). Spambase is flagged by its control band failure. Adult, bank, and credit-g appear only in the TabICLv2 $k = 128$ sweep. Additional dataset IDs are electricity 151, jm1 1053, phoneme 1489, numerai28.6 23517, churn 40701, wilt 40983, credit/GMSC 44089, binarized covertype 44121, pol 44122, house_16H 44123, MiniBooNE 44128, and eye_movements 44130, using versions served on 2026-09-04. Rows with missing values are dropped before an 80/20 stratified split with `random_state=0`.

Training data disclosures. V2.5's default classifier was fine-tuned on the 43 real datasets listed by Grinsztajn et al. (2025), Appendix C.1. That list includes OpenML higgs, which supplies every higgs query and about 1.2% of the disjoint support pool, and the source of eye_movements 44130. It does not list magic or the other eleven extension datasets. A Kaggle term deposit entry shares bank's source. Its rows were not compared. The v3, v1, TabICLv2, Mitra, and TabFM disclosures describe synthetic pretraining and name neither primary dataset as training data. TabICLv2 used 60 TabPFNv2 development datasets for ablation validation, a model selection use of real data whose list could not be retrieved. Mitra selects its prior mixture using a performance vector on unnamed real datasets. Its optional task fine-tuning, enabled by default for 50 steps in AutoGluon 1.5.0, is disabled

in all spread and deletion measurements. The separate accuracy comparison includes a labelled fine-tuning arm. TabFM has no technical report. Its card and release post describe hundreds of millions of synthetic datasets from structural causal models. The earlier Real-TabPFN corpus includes both higgs and magic. Evaluation suites are distinct from training data: TabArena-v0.1 excludes higgs by its selection rule and does not list magic. These disclosures cannot establish training history from the weights, especially for TabFM, where the evidence is a model card.

Synthetic only v2.5. On the same draws, `v2.5_default-2` gives spread slope -0.066 $[-0.09, -0.05]$ over 1024–16,384, versus the fine-tuned checkpoint's -0.082 $[-0.11, -0.07]$. This passes the registered difference tolerance 0.05 and remains slower than the reference rate. Accuracy at 16,384 is 0.7198 $[0.7160, 0.7235]$, versus 0.7196. Although within the registered 0.02 tolerance, it misses the joint prediction of not exceeding the fine-tuned checkpoint, by 0.0002. Initial window slopes are -0.103 on higgs and -0.347 on magic, versus -0.119 and -0.338. Both differences pass 0.05. Mean pair exponents are 0.336 (0.146) and 0.367 (0.045), versus 0.374 (0.123) and 0.285 (0.046). Both pass the three-pooled-SE criterion.

Paired synthetic only minus fine-tuned spread slope differences are $+0.024$ $[+0.01, +0.03]$ on the initial disjoint window, $+0.020$ $[+0.01, +0.03]$ over 1024–8192, $+0.017$ $[+0.01, +0.02]$ over 1024–16,384, and $+0.011$ $[+0.00, +0.02]$ over 1024–32,768. Levels are 1.04–1.08 times the fine-tuned values from 1024 to 32,768, and accuracy differs by at most 0.0021 at each size. Linux accelerator measurements use CUDA builds of the pinned packages (Appendix K).

C THEORY INTERFACE

Efron–Stein. For a support S of n i.i.d. rows and a fixed query,

$$\operatorname{Var} f(S) \le \tfrac{1}{2} \sum_i \mathbb{E}[f(S) - f(S^{(i)})]^2 = \tfrac{n}{2}\, \mathbb{E}[\Delta^2],$$

where $S^{(i)}$ replaces row i by an independent copy and the last equality assumes exchangeable coordinate effects (or defines Δ using a uniformly selected coordinate). For an order invariant predictor, $f(S \setminus i) - f(S \setminus j)$ has the replace-one form for a support of size $n-1$ under this sampling model. If its RMS is $O(n^{-\alpha_{\mathrm{rms}}})$, the bound on SD is $O(n^{1/2-\alpha_{\mathrm{rms}}})$, so $\alpha_{\mathrm{rms}} > 1/2$ suffices for vanishing spread. The margin to probability bound follows from $|\sigma(a) - \sigma(b)| \le |a-b|/4$. Our finite pool, without replacement measurements and order sensitive pair calls are empirical comparisons with this i.i.d. bound, not exact implementations of all its assumptions.

At each n, the pair mean is no larger than the RMS, which is no larger than the pair maximum. This level ordering does not by itself order fitted exponents. The original 48 TabICLv2 RMS deletion matrices are included and reproduce their recorded cell statistics and fits. Table 18 retains the recorded summaries for other arms whose raw matrices are outside this supplement. On the margin scale, RMS exponents exceed $1/2$ at 3 SE for TabICLv2 on both datasets (0.96 and 0.83), fingerprint-off v3 on magic (0.70, $t = +4.0$), and logistic regression on higgs and magic (1.26, $t = +53.5$; 0.92, $t = +9.9$). TabICLv2's probability scale exponents are 0.98 and 1.05. Across both scales its distances are 7–15 SE. On magic, shipped v2.5's platform rerun (0.27, $t = -6.5$) and shipped v3 (0.30, $t = -3.8$) fall below on the margin scale. Other margin scale comparisons are unclassified.

Probability and margin scales. To first order, $|dp| = p(1-p)|dm|$. The fitted slope of $\mathbb{E}_q[p(1-p)]$ therefore provides a diagnostic for the gap between probability and margin exponents, although a product of query means need not equal the mean response. For TabPFN v2.5 over $n \le 2048$, the diagnostic accounts for 86–89% of the Q_{mean} gap: $+0.071$ of $+0.083$ on higgs and $+0.166$ of $+0.187$ on magic. Residuals are $+0.012$ (0.002) and $+0.021$ (0.011), respectively.

A registered accuracy experiment used 200 fixed unfiltered queries, both primary datasets, and six seeds. Every foundation model improved with n without a 2-SE drop. From 64 to 1024 rows, v2.5 gained $+0.15$ on higgs and $+0.08$ on magic despite its influence exponent below $1/2$. These gains were 6.5 and 4.7 SE under the registered unpaired scoring, or 8.1 and 4.0 SE under seed paired scoring. TabICLv2 also improved (4.6 and 4.8 SE) with an influence exponent near 1. These four model–dataset comparisons show that accuracy improvement and influence decay need not move together. They establish no general relation.

The measured checkpoints have 12- to 24-block trunks. Multi-layer transformer theorems generally bound prediction risk rather than a replace-one difference. Ma et al. (2026) obtain minimax risk $n^{-2\beta/(2\beta+d)}$ for an empirical risk minimizing softmax transformer of depth $L \geq 3$. Ching et al. (2026) obtain the same rate with $\lceil C \log(en) \rceil$ linear attention blocks. Combining $\mathrm{Var} \leq \mathrm{MSE}$ with Jensen's inequality bounds query-averaged conditional SD by $n^{-\beta/(2\beta+d)}$, retaining the task and pretraining averages in the respective risk guarantees. This does not give a bound for every query or fixed pretrained checkpoint. Theorem 3.2 of Li et al. (2023) instead gives a replace-one stability bound of order $1/n$ with explicit norm dependent constants. Our norm audit does not establish those assumptions: the trained Mitra trunk has a median spectral norm $1.84\times$ the random matrix edge at the Frobenius norms of its shuffled copy.

Hahn (2020) prove $O(1/n)$ single symbol sensitivity for soft attention transformers at fixed depth under Lipschitz assumptions on position-wise functions (Lemma 5). Chiang & Cholak (2022) show that layer normalization with $\epsilon = 0$ can violate that premise (Proposition 2), while $\epsilon > 0$ restores it but can delay the decay over finite lengths (Figure 2). Prior induced bias that persists as data accumulate is also documented by Melnychuk et al. (2026). These results motivate finite range checks without identifying the cause of the rates measured here.

The following elementary bound separates direct removal of an aggregation weight from changes in the remaining representations, in the spirit of Nagler's Lemma 6.1.

Proposition 1. *For a fixed query x^* and a support of size $n \geq 2$, let $f(S, x^*) = g(\sum_i a_i(S)\, v_i(S))$ with $0 \leq a_i < 1$, $\sum_i a_i = 1$, and $\|v_i\| \leq D$. Let $\bar{v}(S) = \sum_i a_i(S) v_i(S)$, and assume g is L-Lipschitz on a convex set containing the original and all deleted support aggregates. Write $\tilde{v}_i = (1-a_i)^{-1} \sum_{j \neq i} a_j v_j$ and $\varepsilon_i = \|\bar{v}(S \setminus i) - \tilde{v}_i\|$ and $Q_{\mathrm{mean}} = n^{-1} \sum_i |f(S, x^*) - f(S \setminus i, x^*)|$. Then $\tilde{v}_i - \bar{v} = \frac{a_i}{1-a_i}(\bar{v} - v_i)$ exactly, and*

$$Q_{\mathrm{mean}} \;\leq\; \frac{2LD}{n}\,\kappa(S) \;+\; \frac{L}{n}\sum_i \varepsilon_i, \qquad \kappa(S) = \sum_i \frac{a_i}{1-a_i} \;\leq\; \frac{1}{1-\max_i a_i}.$$

Proof. Subtracting $\bar{v}$ from $\tilde{v}_i$ gives the stated identity. Since $\|\bar{v}\| \leq D$, the Lipschitz and triangle inequalities give

$$|f(S, x^*) - f(S \setminus i, x^*)| \leq L\|\bar{v} - \tilde{v}_i\| + L\varepsilon_i \leq 2LD\frac{a_i}{1-a_i} + L\varepsilon_i.$$

Average over the n deletions and use $\sum_i a_i/(1-a_i) \leq \sum_i a_i/(1-\max_j a_j) = 1/(1-\max_j a_j)$. □

The proposition bounds influence levels. Its first term is $O(1/n)$ when κ stays bounded, but the remainder also needs control before this becomes a bound on total influence decay. It does not bound a finite window fitted slope. The logistic regression exponents above 1 (1.335 for the sampled pair maximum, 1.214 for Q_{mean}) are therefore not anomalous.

For Mitra's label channel attention, the maximum weight over layers is at most 0.035 for $64 \leq n \leq 1024$ (last layer, 0.023), giving $\kappa \leq 1.036$. Across that range, this upper bound variation corresponds to a slope allowance of 0.013. The all channel maximum, 0.21 at layer 3 and $n = 256$, gives $\kappa \leq 1.27$ and an allowance of 0.085. These values constrain the attention term. They do not establish the rate of the total deletion response.

The registered Mitra remainder measurement (AP5, higgs, last cross-row layer, $k = 32$, six seeds) gives shares ρ of 0.664 (0.045), 0.734, 0.833, 0.874, and 0.836 (0.017) from 64 to 1024 rows. The relative remainder level ε falls from 0.0618 to 0.0085, and the attention term δ from 0.0265 to 0.0015. Here $\rho = \varepsilon/(\varepsilon + \delta)$ is the remainder's share of the two component norms, 0.66–0.87. It is not a fraction of the actual output change: the vector components can cancel, and the readout need not preserve their norms. Both terms shrink, with fitted exponents 0.62 ± 0.10 and 1.02 ± 0.01, respectively.

A nonzero representation change remainder prevents us from using this bound to infer aggregation structure from an observed influence exponent.

Empirical bound comparisons. Over 64–512, the TabICLv2 quantity $\sqrt{(n-1)/2}\,\Delta_{\mathrm{rms}}$ exceeds the measured support SD except for one near-equality. The ratio falls from $2.1\times$ to $1.6\times$ on higgs and from $1.6\times$ to 1.0–$1.2\times$ on magic. At magic's 256-row measurement, it is $0.997\times$ using the seed mean of Δ and $1.007\times$ using the seed mean of Δ^2. This comparison uses different query sets: the first 50 unfiltered held out queries for spread and 50 correctly classified queries for deletion. It is therefore descriptive, not a test of the bound. The ratio decreases because the empirical bound estimate decays faster than measured spread. Under the bound's assumptions, both fitted rates could not continue indefinitely. RMS influence beyond 2048 is unmeasured.

The registered ES experiment measures TabICLv2 deletion influence on the spread experiment's own 50 higgs queries, using $k = 512$ deletions, six seeds, and $n \in \{512, 1024, 2048\}$. The empirical quantity $\sqrt{(n-1)/2}\,\Delta_{\mathrm{rms},p}$ exceeds disjoint support SD by factors 1.51, 1.35, and 1.31 (1.60, 1.41, and 1.27 for the stored overlapping draws). The probability scale RMS exponent is 0.74 (0.06), meeting the predicted upper limit of 0.75. The margin scale exponent is 0.70. This is flatter than the clean-correct 64–512 fits near 1. If the bound's fitted exponent of about 0.25 and the disjoint spread exponent of 0.10–0.11 continued, the $1.31\times$ ratio at 2048 would close near 12,000–14,000 rows, within three octaves. Under the bound's assumptions, at least one rate would then have to change. No deletion measurement covers that extrapolation. The correctly classified query check at $n = 512$, seed 0, reproduces the stored 2026-08-22 RMS measurement to 10^{-6} on both pair statistics and scales.

D SPREAD MEASUREMENT DETAIL

Table 7 reports signed spread slopes with support- and query resampling intervals. The registered exclusion rule requires spread to exceed the order floor by at least $3\times$. V2.5's ratios are 9.6–63 on higgs and at least 208 on magic. Its extension measurements at $n \geq 2048$ have ratios at least 36. V3's minimum initial window ratios are 3.6 on higgs and 3.8 on magic. TabICLv2's higgs slope, -0.28, misses the prediction of reaching $-1/2$, to which the registration assigned probability 0.75 (Appendix K).

Sensitivity analyses. Three post hoc checks use stored probability matrices (`spread_sensitivity.py`). None changes a registered classification.

- Joint resampling of support and query indices (10^4 replicates, 50 queries) widens initial window intervals by up to 0.08 on either side. Only TabFM on magic changes interpretation: the joint interval $[-0.55, -0.34]$ contains $-1/2$, whereas the registered support interval is $[-0.47, -0.40]$. The resampling fraction $P(\mathrm{attain})$ is 0.12. All other trained model joint intervals stay above $-1/2$.
- At $n = 64$, trained Mitra's query mean SD is 0.096 versus 0.0051 for its shuffled copy. Trained TabICLv2's is 0.195 versus 0.0012 for its full scope copy. Across the twelve trained model–dataset combinations, spread is 0.10–0.20 at 64 and 0.04–0.10 at 1024.
- Replacing the mean of per-query SDs by the square root of mean per-query variance changes higgs slopes by at most 0.011. Every magic fit becomes flatter by 0.07–0.11. Thus the registered aggregation gives the faster magic decay, without changing which intervals exclude the reference rate.

Leave-one-context-size-out sensitivity (post hoc). We omit each size in turn from the principal disjoint higgs windows and refit the query mean of per-query sample SDs against n on log-log axes (Table 5). The inputs are the stored probability matrices. The original $V \geq 3F$ floor rule and disjoint support checks are unchanged. Each fit uses 10^4 support index resamples with RNG seed 0 and the published scorer's degenerate replicate redraw rule. When all retained sizes have the same draw count, a replicate uses one shared index vector across sizes. If any counts differ, every retained size is resampled independently at its own count. Thus omitting 16,384 changes v2.5 and v1 from independent resampling (48/24 draws) to paired resampling (48 draws throughout), as required by that convention. The 50 queries stay fixed. Equal draw counts alone do not justify cross-size dependence: the archived runner redraws supports at each size. We preserve the historical convention and add an independent at every size sensitivity below, with queries shared across sizes. None of these intervals has a simultaneous coverage or calibrated coverage claim.

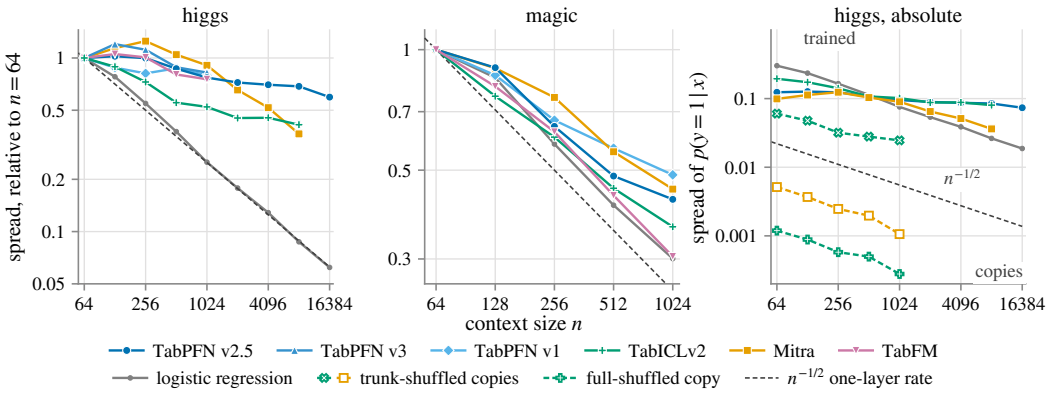


Figure 2: Prediction spread against support size (log-log axes). Left and middle: curves normalized at $n = 64$, with an $n^{-1/2}$ guide. Higgs v2.5 and logistic regression use disjoint supports through 16,384. Mitra and the displayed TabICLv2 run end at 8192. Other curves use the original draws through 1024. Table 1 includes the longer Linux TabICLv2 run. Right: absolute higgs spread, including randomized Mitra and TabICLv2 copies (dashed, hollow markers). At 64, the Mitra and full scope TabICLv2 copies have about 20 and 160 times less spread than trained. At 16,384, v2.5's SD is 0.074, versus logistic regression's 0.019. Tables 7 and 8 give fits and uncertainty.

Across the twenty omissions for v2.5, v3, TabICLv2, and TabFM, slopes range from -0.155 to -0.055, with every support interval above $-1/2$. Table 5 shows the window dependence, largest for TabFM. All five logistic regression intervals contain $-1/2$, and v1's and Mitra's remain above it. The separately registered exact tree check is in Table 10. `reproduce.py sensitivity` produces every omitted size fit and interval. None establishes a power law or extrapolates to larger contexts.

Table 5: Post-hoc leave-one-context-size-out probability-spread slopes on disjoint higgs supports. Each column names the omitted size; 'none' reproduces the published full fit. Windows are 1024–16,384, except †: Mitra and XGBoost use 1024–8192. Mitra retains its package cap; XGBoost's larger measurements fail the floor rule. V1's extension exceeds its 1024 cap. Detailed 95% support-only intervals, retained sizes, draw counts, exclusions, and resampling modes are in `expected_results/spread_leave_one_out.csv`.

	Omitted context size					
Model	none	1024	2048	4096	8192	16,384
Mitra†	−0.428	−0.420	−0.433	−0.435	−0.405	—
TabICLv2	−0.100	−0.082	−0.111	−0.100	−0.100	−0.102
TabPFN v2.5	−0.082	−0.087	−0.085	−0.082	−0.092	−0.055
TabPFN v3	−0.141	−0.116	−0.147	−0.141	−0.141	−0.155
TabPFN v1	−0.409	−0.441	−0.402	−0.409	−0.412	−0.384
TabFM	−0.122	−0.066	−0.137	−0.122	−0.121	−0.154
logreg	−0.506	−0.512	−0.506	−0.506	−0.503	−0.504
logreg, $C = 10^6$	−0.509	−0.514	−0.509	−0.509	−0.506	−0.509
XGBoost†	−0.367	−0.373	−0.366	−0.367	−0.361	—
$\sqrt{n}$-NN	−0.251	−0.253	−0.250	−0.251	−0.251	−0.251

Post hoc query and support sensitivities. Joint resampling of queries and supports preserves the disjoint-HIGGS contrast over 1,024–16,384 rows (Table 6). We retain the recorded query pairing across sizes and compare the historical support resampling rule with independent support resampling at every size. Neither has calibrated coverage for the finite pool design. Reusing the stored 2,000-query accuracy matrices gives probability spread slopes of -0.109, -0.158, -0.139, and -0.137 for v2.5, v3, TabICLv2, and TabFM, versus -0.509 for logistic regression and -0.366 for exact tree XGBoost. These matrices use only eight supports at each size, whose stored indices match the first eight main supports exactly. The comparison also changes support count and, for some models, execution platform or query batch, so it does not isolate query count alone. In particular,

Table 6: Post-hoc disjoint-HIGGS probability-spread slopes over 1,024–16,384 rows, with joint query-and-support percentile intervals. Supports are resampled independently at every size; a single query resample is shared across sizes. The 50-query analyses use the original 24–48 supports; the 2,000-query analyses use eight supports. The respective bootstrap counts are 10,000 and 10,000; coverage is uncalibrated. The latter reuse accuracy-run probabilities, not accuracy variances, and differ in support count and, for some models, execution platform/query batch.

Model	50 queries	2,000 queries
TabPFN v2.5	-0.082 $[-0.127, -0.049]$	-0.109 $[-0.163, -0.055]$
TabPFN v3	-0.141 $[-0.178, -0.105]$	-0.158 $[-0.214, -0.104]$
TabICLv2	-0.100 $[-0.140, -0.063]$	-0.139 $[-0.193, -0.083]$
TabFM	-0.122 $[-0.163, -0.084]$	-0.137 $[-0.192, -0.079]$
logistic regression	-0.506 $[-0.534, -0.478]$	-0.509 $[-0.580, -0.444]$
exact-tree XGBoost	-0.369 $[-0.406, -0.342]$	-0.366 $[-0.417, -0.317]$

TabICLv2's main 50-query slope is -0.100 using all 24 supports and -0.145 using only their first eight. These checks support the finite range contrast without establishing an asymptotic rate or a relationship between concentration and predictive quality. Reproduction details and the support count decomposition are supplied in `followup/robustness/`.

Configurations and exclusions (Tables 7 and 8). Vendor default slopes differ from the single estimator/view fits by at most 0.03 for v2.5 (VE) and 0.05 for TabICLv2 and TabFM (AP1). TabICLv2 uses Latin-square feature/class shuffles. TabFM uses random views, class shifts, and bfloat16 autocast. There is no v3 vendor default spread run or Mitra spread run with fine-tuning.

Mitra's seeded vendor stack higgs run stopped at 256 under the registered two-hour limit, leaving the floor prediction unevaluable. Six other units passed. On the matched 64–256 range, the harness slope is $+0.14$, with spread levels within 0.003 of the stack's at each size. The Linux completion gives -0.06 over 64–1024, with its interval above $-1/2$. The separate Mitra extension gives -0.18 $[-0.21, -0.13]$ over 64–4096. Local slopes are -0.39 from 1024 to 2048 and -0.43 from 2048 to 4096.

Classical learners (CA) are refit on the same 48 draws and queries. Random forest bootstrap noise fails the $V \geq 3F$ rule at every higgs size and all but the three smallest magic sizes. The MLP optimization floor excludes its higgs extension. The supplementary registration ledger and detailed record retain these omitted fits and the seeded stack reproduction outcomes. Extension rows use complete grids whose base measurements reproduce the initial runs. Fingerprint-off extensions are detailed in Appendix G.

Disjoint supports and reference comparisons. The disjoint support pool contains the original 40,000 training rows followed by UCI HIGGS rows matching none of the OpenML subsample. The 50 spread queries are unchanged. Figure 1 shows the absolute spread levels on these supports. TabICLv2's macOS run stopped at 8192 under its computational limit. The Linux run reaches 32,768. V1 measurements from 2048 onward exceed its 1024-row cap. The TabICLv2 full shuffled copy's long windows and histogram XGBoost's 16,384 and 32,768 measurements reach only $2.7\times$ and $2.4\times$ their order floors, so no slope is reported for those windows.

The randomized Mitra and TabICLv2 copies' disjoint 64–1024 point slopes do not reach $-1/2$, although their intervals contain it. Mitra's longer window point does reach it. Thus the initial overlapping draw point estimates do not transfer to every draw scheme (Tables 7 and 8).

The registered local refit comparison resamples shared draw indices for each pair. V2.5, v3, TabICLv2, and TabFM are flatter than $\sqrt{n}$-NN by $+0.11$ to $+0.17$ (narrowest interval $[+0.09, +0.14]$) and flatter than XGBoost by $+0.22$ to $+0.31$ on the valid matched windows. V1 and Mitra are steeper than $\sqrt{n}$-NN by -0.16 and -0.17. V1 is similar to XGBoost (-0.02 $[-0.05, +0.02]$). Mitra is steeper over 1024–8192 (-0.05 $[-0.09, -0.01]$), the experiment's failed prediction X4. No checkpoint is flatter than 5-NN.

DH's failed prediction is v2.5's reproduction of the overlapping extension. All four original disjoint fits reproduce the initial overlapping slopes within 0.05. Extension differences also stay within 0.05

Table 7: Probability spread slopes on the original support draws. Query CI resamples queries. Support CI is the registered paired support resampling 95% interval. Dashes indicate unavailable intervals. R is draws per size. Corrected rows use the within arm κ_w or $\kappa = 1$ overlap sensitivity analysis (Table 14). Disjoint higgs results are in Table 8. p denotes the Linux environment.

Arm	data	slope	query CI	support CI	window	R
Mitra trained	higgs	-0.06	$[-0.09, -0.02]$	$[-0.11, +0.02]$	64–1024	24
Mitra shuffled	higgs	-0.55	$[-0.55, -0.54]$	$[-0.68, -0.44]$	64–1024	24
Mitra trained	magic	-0.30	—	$[-0.34, -0.26]$	64–1024	24
TabICLv2 trained	higgs	-0.28	$[-0.30, -0.25]$	$[-0.31, -0.24]$	64–1024	24
TabICLv2 trained	magic	-0.37	$[-0.45, -0.30]$	$[-0.41, -0.33]$	64–1024	24
TabICLv2 trunk shuffled	higgs	-0.34	$[-0.35, -0.32]$	$[-0.39, -0.28]$	64–1024	24
TabICLv2 full shuffled	higgs	-0.502	$[-0.509, -0.496]$	$[-0.60, -0.41]$	64–1024	24
Mitra shuffled[p]	magic	-0.52	$[-0.53, -0.51]$	$[-0.65, -0.41]$	64–1024	24
TabICLv2 full shuffled[p]	magic	-0.463	$[-0.48, -0.45]$	$[-0.56, -0.37]$	64–1024	24
TabPFN v2.5	higgs	-0.12	$[-0.15, -0.09]$	$[-0.16, -0.07]$	64–1024	48
TabPFN v2.5	magic	-0.34	$[-0.43, -0.26]$	$[-0.37, -0.30]$	64–1024	48
TabPFN v2.5 (ext.)	higgs	-0.12	—	$[-0.15, -0.10]$	1024–8192	48
TabPFN v2.5 (ext.)	magic	-0.51	—	$[-0.54, -0.47]$	1024–8192	48
TabPFN v2.5 (ext., corrected, κ_w)	magic	-0.38	—	$[-0.41, -0.34]$	1024–8192	48
TabPFN v2.5 (ext., corrected, $\kappa = 1$)	magic	-0.34	—	$[-0.38, -0.31]$	1024–8192	48
TabICLv2 (ext.)	higgs	-0.12	—	$[-0.18, -0.07]$	1024–4096	24
Mitra (ext.)	higgs	-0.41	$[-0.45, -0.37]$	$[-0.48, -0.34]$	1024–4096	24
TabPFN v1 (ext.)	higgs	-0.42	—	$[-0.45, -0.39]$	1024–8192	48
TabPFN v1 (ext.)	magic	-0.64	—	$[-0.67, -0.60]$	1024–8192	48
TabPFN v1 (ext., corrected, κ_w)	magic	-0.50	—	$[-0.53, -0.46]$	1024–8192	48
TabPFN v1 (ext., corrected, $\kappa = 1$)	magic	-0.475	—	$[-0.504, -0.437]$	1024–8192	48
TabPFN v3	higgs	-0.10	—	$[-0.15, -0.04]$	64–1024	24
TabPFN v3	magic	-0.38	$[-0.48, -0.29]$	$[-0.43, -0.33]$	64–1024	24
TabPFN v1	higgs	-0.07	—	$[-0.11, -0.02]$	64–1024	48
TabPFN v1	magic	-0.27	—	$[-0.30, -0.24]$	64–1024	48
TabFM trained	higgs	-0.12	$[-0.15, -0.09]$	$[-0.16, -0.08]$	64–1024	24
TabFM trained	magic	-0.43	$[-0.54, -0.35]$	$[-0.47, -0.40]$	64–1024	24
v2.5 ensemble (8 est.)	higgs	-0.145	—	$[-0.19, -0.09]$	64–1024	24
v2.5 ensemble (8 est.)	higgs	-0.157	—	$[-0.19, -0.13]$	64–4096	24
v2.5 ensemble (8 est.)	magic	-0.334	—	$[-0.38, -0.30]$	64–1024	24
v2.5 ensemble (8 est., $\kappa = 1$: -0.305)	magic	-0.337	—	$[-0.36, -0.31]$	64–4096	24
Mitra vendor stack, seeded	higgs	$+0.161$	—	$[+0.01, +0.34]$	64–256	24
Mitra vendor stack, seeded	magic	-0.310	—	$[-0.35, -0.27]$	64–1024	24
TabICLv2 vendor default (8 est.)[p]	higgs	-0.267	$[-0.30, -0.24]$	$[-0.31, -0.23]$	64–1024	24
TabICLv2 vendor default (8 est.)[p]	magic	-0.378	$[-0.46, -0.31]$	$[-0.42, -0.33]$	64–1024	24
TabFM vendor default (32 views, bf16)[p]	higgs	-0.141	$[-0.17, -0.11]$	$[-0.18, -0.10]$	64–1024	24
TabFM vendor default (32 views, bf16)[p]	magic	-0.440	$[-0.55, -0.36]$	$[-0.48, -0.40]$	64–1024	24
Mitra vendor stack, seeded, Linux[p]	higgs	-0.057	$[-0.09, -0.02]$	$[-0.12, +0.02]$	64–1024	24
logreg (refit)	higgs	-0.52	—	$[-0.54, -0.50]$	64–1024	48
logreg (refit)	magic	-0.45	—	$[-0.48, -0.42]$	64–1024	48
logreg (refit, ext.)	higgs	-0.53	—	$[-0.56, -0.50]$	1024–8192	48
logreg (refit, ext.)	magic	-0.62	—	$[-0.66, -0.59]$	1024–8192	48
logreg (refit, ext., corrected, $\kappa = 0.8$)	magic	-0.49	—	$[-0.53, -0.45]$	1024–8192	48
logreg (refit, ext., corrected, $\kappa = 1$)	magic	-0.46	—	$[-0.50, -0.42]$	1024–8192	48
logreg (refit, C=10^6)	higgs	-0.64	$[-0.65, -0.62]$	$[-0.65, -0.62]$	64–1024	48
logreg (refit, C=10^6, ext., uncorrected)	higgs	-0.53	$[-0.56, -0.51]$	$[-0.57, -0.50]$	1024–8192	48
logreg (refit, C=10^6)	magic	-0.568	—	$[-0.60, -0.53]$	64–1024	48
logreg (refit, C=10^6, corrected, $\kappa = 1$)	magic	-0.557	—	$[-0.59, -0.52]$	64–1024	48
logreg (refit, C=n)	magic	-0.556	—	$[-0.59, -0.52]$	64–1024	48
logreg (refit, C=n, corrected, $\kappa = 1$)	magic	-0.545	—	$[-0.58, -0.51]$	64–1024	48
XGBoost (refit)	higgs	-0.24	—	$[-0.26, -0.23]$	64–1024	48
XGBoost (refit)	magic	-0.29	—	$[-0.31, -0.27]$	64–1024	48
$\sqrt{n}$-NN (refit)	higgs	-0.27	—	$[-0.29, -0.25]$	64–1024	48
$\sqrt{n}$-NN (refit)	magic	-0.29	—	$[-0.32, -0.27]$	64–1024	48
5-NN (refit)	higgs	-0.02	—	$[-0.04, -0.00]$	64–1024	48
5-NN (refit)	magic	-0.09	—	$[-0.11, -0.07]$	64–1024	48
MLP (refit)	higgs	-0.05	—	$[-0.06, -0.04]$	64–1024	48
MLP (refit)	magic	-0.36	—	$[-0.39, -0.33]$	64–1024	48

Table 8: Probability spread slopes on disjoint higgs supports from the $3{,}440{,}379$-row pool (DH), with support resampling 95% intervals. Supports within a size share no rows. No overlap correction is applied. R is draws per size, with smaller draw cells resampled independently. Historical intervals condition on fixed queries. The post hoc joint sensitivity is reported separately. p denotes the Linux environment.

Arm	slope	support CI	window	R
TabPFN v2.5	-0.099	$[-0.13, -0.06]$	64–1024	48
TabPFN v2.5	-0.055	$[-0.08, -0.03]$	1024–8192	48
TabPFN v2.5	-0.082	$[-0.11, -0.07]$	1024–16,384	48/24
TabPFN v2.5	-0.120	$[-0.15, -0.11]$	1024–32,768	48/24/16
TabPFN v3$^{\mathrm{p}}$	-0.092	$[-0.14, -0.04]$	64–1024	24
TabPFN v3$^{\mathrm{p}}$	-0.155	$[-0.18, -0.13]$	1024–8192	24
TabPFN v3$^{\mathrm{p}}$	-0.141	$[-0.16, -0.12]$	1024–16,384	24
TabPFN v3$^{\mathrm{p}}$	-0.128	$[-0.15, -0.11]$	1024–32,768	24/16
TabICLv2	-0.256	$[-0.29, -0.23]$	64–1024	24
TabICLv2	-0.102	$[-0.13, -0.07]$	1024–8192	24
TabICLv2$^{\mathrm{p}}$	-0.100	$[-0.12, -0.08]$	1024–16,384	24
TabICLv2$^{\mathrm{p}}$	-0.109	$[-0.13, -0.10]$	1024–32,768	24/16
TabFM$^{\mathrm{p}}$	-0.127	$[-0.16, -0.10]$	64–1024	24
TabFM$^{\mathrm{p}}$	-0.154	$[-0.18, -0.13]$	1024–8192	24
TabFM$^{\mathrm{p}}$	-0.122	$[-0.14, -0.10]$	1024–16,384	24
TabFM$^{\mathrm{p}}$	-0.130	$[-0.15, -0.12]$	1024–32,768	24/16
TabICLv2 vendor default (8 est.)$^{\mathrm{p}}$	-0.243	$[-0.28, -0.21]$	64–1024	24
TabICLv2 vendor default (8 est.)$^{\mathrm{p}}$	-0.147	$[-0.18, -0.12]$	1024–8192	24
TabICLv2 vendor default (8 est.)$^{\mathrm{p}}$	-0.141	$[-0.16, -0.12]$	1024–16,384	24
TabICLv2 vendor default (8 est.)$^{\mathrm{p}}$	-0.133	$[-0.15, -0.12]$	1024–32,768	24/16
TabFM vendor default (32 views, bf16)$^{\mathrm{p}}$	-0.136	$[-0.17, -0.10]$	64–1024	24
TabFM vendor default (32 views, bf16)$^{\mathrm{p}}$	-0.172	$[-0.19, -0.15]$	1024–8192	24
TabFM vendor default (32 views, bf16)$^{\mathrm{p}}$	-0.144	$[-0.16, -0.13]$	1024–16,384	24
TabFM vendor default (32 views, bf16)$^{\mathrm{p}}$	-0.138	$[-0.16, -0.12]$	1024–32,768	24/16
Mitra	-0.040	$[-0.11, +0.05]$	64–1024	24
Mitra	-0.428	$[-0.46, -0.39]$	1024–8192	24
TabPFN v1$^{\mathrm{p}}$	-0.012	$[-0.06, +0.04]$	64–1024	48
TabPFN v1$^{\mathrm{p}}$	-0.384	$[-0.41, -0.36]$	1024–8192	48
TabPFN v1$^{\mathrm{p}}$	-0.409	$[-0.44, -0.39]$	1024–16,384	48/24
TabPFN v1$^{\mathrm{p}}$	-0.421	$[-0.45, -0.41]$	1024–32,768	48/24/16
Mitra shuffled$^{\mathrm{p}}$	-0.416	$[-0.55, -0.30]$	64–1024	24
Mitra shuffled$^{\mathrm{p}}$	-0.595	$[-0.81, -0.38]$	1024–8192	24
TabICLv2 full shuffled$^{\mathrm{p}}$	-0.397	$[-0.52, -0.29]$	64–1024	24
logreg (refit)	-0.504	$[-0.53, -0.48]$	64–1024	48
logreg (refit)	-0.506	$[-0.53, -0.48]$	1024–16,384	48
logreg (refit)	-0.518	$[-0.53, -0.51]$	1024–65,536	48/24
logreg (refit)	-0.517	$[-0.53, -0.51]$	64–65,536	48/24
logreg (refit, $C{=}10^6$)	-0.614	$[-0.64, -0.59]$	64–1024	48
logreg (refit, $C{=}10^6$)	-0.509	$[-0.53, -0.49]$	1024–16,384	48
logreg (refit, $C{=}10^6$)	-0.520	$[-0.54, -0.51]$	1024–65,536	48/24
logreg (refit, $C{=}10^6$)	-0.548	$[-0.56, -0.54]$	64–65,536	48/24
$\sqrt{n}$-NN (refit)	-0.247	$[-0.26, -0.23]$	64–1024	48
$\sqrt{n}$-NN (refit)	-0.251	$[-0.27, -0.23]$	1024–8192	48
$\sqrt{n}$-NN (refit)	-0.251	$[-0.27, -0.24]$	1024–16,384	48/24
$\sqrt{n}$-NN (refit)	-0.258	$[-0.28, -0.25]$	1024–32,768	48/24/16
XGBoost (refit)	-0.231	$[-0.24, -0.22]$	64–1024	48
XGBoost (refit)	-0.367	$[-0.38, -0.35]$	1024–8192	48
XGBoost (refit)	conf.	conf.	1024–16,384	48/24
XGBoost (refit)	conf.	conf.	1024–32,768	48/24/16
5-NN (refit)	-0.009	$[-0.02, +0.00]$	64–1024	48
5-NN (refit)	$+0.000$	$[-0.01, +0.02]$	1024–8192	48
5-NN (refit)	-0.016	$[-0.03, -0.01]$	1024–16,384	48/24
5-NN (refit)	-0.014	$[-0.04, -0.01]$	1024–32,768	48/24/16

Table 9: Separately registered XGBoost exact-tree, CPU, XGBoost 3.2.0: 100 depth-4 trees, learning rate 0.1, seed 0, one thread. Spread V and order floor F use 50 fixed queries; accuracy uses the first eight matching supports and 2000 fixed queries (95% support-level t intervals). A zero observed floor gives $V/F = \infty$, without establishing exact invariance beyond these permutations.

n	R	V	F	V/F	Accuracy (95% CI)
1,024	48	0.155550	0.000000	∞	0.6596 [0.6502, 0.6690]
2,048	48	0.119818	0.000000	∞	0.6807 [0.6756, 0.6858]
4,096	48	0.092581	0.000000	∞	0.6961 [0.6887, 0.7036]
8,192	48	0.069829	0.000000	∞	0.7007 [0.6944, 0.7071]
16,384	24	0.056701	0.000000	∞	0.7113 [0.7049, 0.7177]
32,768	16	0.044426	0.000000	∞	0.7162 [0.7118, 0.7206]

Table 10: Exact-tree finite-range spread slopes with 10^4 support-bootstrap percentile intervals. The last column ranges over leave-one-size-out point slopes; full omission intervals and draw counts are supplied in the reproduction outputs. Equal draw counts use paired indices across sizes; unequal counts use independent resampling at every size. None establishes an asymptotic rate.

Window	Slope	95% CI	One-size omission range
1024–8, 192	-0.384	$[-0.402, -0.365]$	$[-0.389, -0.374]$
1024–16, 384	-0.369	$[-0.390, -0.359]$	$[-0.384, -0.363]$
1024–32, 768	-0.362	$[-0.390, -0.355]$	$[-0.369, -0.357]$

for TabICLv2, Mitra, and logistic regression (Tables 7 and 8). V2.5's disjoint 1024–8192 slope of -0.055 is 0.065 flatter than the overlapping -0.12, in the direction predicted by overlap correction (DH5a). Pairings of Linux checkpoint runs with macOS refits match draws across machines. Appendix K gives their numerical check differences and registered outcomes.

Registered exact tree robustness comparison. A separate configuration, XGBoost exact tree (version 3.2.0, 100 trees, depth 4, learning rate 0.1, seed 0, one CPU thread), was registered and its file hashed and locally timestamped before measurement. It changes only the tree method from the historical default histogram arm. The original measurements, floor exclusions, and failed prediction X4 above remain unchanged. The six sizes are 1024 through 32,768, with 48 draws through 8192, 24 at 16,384, and 16 at 32,768. All support indices match the stored disjoint arms exactly. Spread and accuracy retain their 50 and 2000 fixed queries. Eight permutations of draw 0 use the original order seeds. Every cell clears the unchanged $V \geq 3F$ rule: all eight permutation predictions agree at each size, giving zero measured floor (Table 9). This establishes stability for the tested permutations, not universal row order invariance.

Tables 9–11 report every size, fit window, and paired checkpoint comparison, including v1 and Mitra. The 10^4-resample intervals condition on fixed queries and preserve the unequal draw convention. Larger window pairings are descriptive. No checkpoint versus XGBoost accuracy significance test was specified. The separate histogram configuration all-deletion comparison is in Table 19.

Accuracy and registered predictions (Table 12). The four accelerator checkpoints gain 0.040–0.053 accuracy from 1024 to 32,768, exceeding the registered 0.0205 threshold. Mitra gains 0.0195 by its 8192-row cap and misses that threshold (A9). Logistic regression is flat from 8192. Its gap to v2.5 does not close (SE of the difference, 0.005). At 16,384, across support accuracy SD exceeds logistic regression's 0.0052 only for v2.5 (0.0063), not v3 (0.0051), TabICLv2 (0.0045), or TabFM (0.0031). The registered prediction of larger accuracy variation therefore fails (A5), despite larger prediction spread.

V2.5, TabICLv2, and Mitra ran in Linux environment P1; v3 and TabFM in P2; and refits under macOS. Relative differences in the reference checks' mean probability are 1.5×10^{-4} and 8.4×10^{-5} on check A, and 2.2×10^{-4} and 7.4×10^{-5} on check B. Of sixteen registered units, A5 and A9 fail and the other fourteen pass (supplementary registration ledger). Run directories retain Brier scores and per-support accuracies.

Table 13. The $a = 1$ support interval comes from the support resampling rerun. Its query interval is $[-0.55, -0.54]$. The registered lowest differing pair rule selected the refinement at $a = 0.125$.

Table 11: Paired spread-slope differences, checkpoint minus XGBoost exact-tree, on the same first $\min(R_{\text{checkpoint}}, R_{\text{exact}})$ supports per size. Positive differences mean flatter checkpoint slopes. The same draw indices are resampled for both arms within each size, independently across sizes (10^4 resamples). Intervals are support-only and have no simultaneous-coverage claim; 32,768-endpoint comparisons are descriptive. Checkpoint hardware follows Table 8; the exact-tree refit uses the laptop CPU.

Checkpoint	Window	Difference	95% CI	Reading
Mitra	1024–8, 192	-0.034	$[-0.073, +0.004]$	unresolved
TabICLv2	1024–16, 384	$+0.272$	$[+0.249, +0.294]$	flatter
TabICLv2	1024–32, 768	$+0.255$	$[+0.234, +0.275]$	flatter
TabPFN v2.5	1024–16, 384	$+0.287$	$[+0.266, +0.308]$	flatter
TabPFN v2.5	1024–32, 768	$+0.243$	$[+0.219, +0.265]$	flatter
TabPFN v3	1024–16, 384	$+0.232$	$[+0.205, +0.257]$	flatter
TabPFN v3	1024–32, 768	$+0.236$	$[+0.212, +0.258]$	flatter
TabPFN v1	1024–16, 384	-0.040	$[-0.064, -0.015]$	steeper
TabPFN v1	1024–32, 768	-0.059	$[-0.083, -0.035]$	steeper
TabFM	1024–16, 384	$+0.250$	$[+0.223, +0.276]$	flatter
TabFM	1024–32, 768	$+0.234$	$[+0.210, +0.258]$	flatter

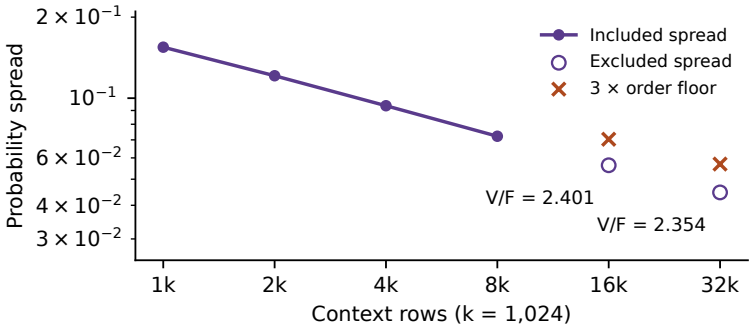


Figure 3: Historical histogram XGBoost exclusion evidence, retained separately from exact tree Figure 1. Hollow, disconnected points at 16,384 and 32,768 have spread/floor ratios 2.401 and 2.354, below 3. Crosses show $3F$. The four smaller cells have zero observed order floor. No fitted slope spans the excluded cells. Historical accuracy at every size remains in Table 12.

Mixtures reaching the rate have accuracies 0.54 and 0.51, respectively 3.1 and 3.7 unpaired binomial SE below the trained 0.69. The $a = 0.125$ accuracy, 0.72, is within one SE of trained. Three predictions fail: no mixture was expected both to retain label response and reach the rate, but two do; label response was expected to decline monotonically, but follows $1.10 \to 0.84 \to 1.11$; and the first rate reaching mixture was expected above $a = 0.5$, but occurs at 0.25.

Nagler's synthetic testbed. We use $p_0(1 \mid x) = 1/2 + \sin(\mathbf{1}^\top x)/2$ with $X \sim N(0, I_5)$ and TabPFN v1 (pip 0.1.8), the release studied by Nagler. There are 500 draws per initial window size and 200 per beyond cap size. Nagler's Figure 1 extends to about 4000 rows, as read from its linear axis. The text specifies no range. Our variance slopes are -1.03 over 1024–4096 and -1.16 over 1024–8192. The corresponding spread slopes, computed post hoc from the registered statistic, are -0.54 $[-0.55, -0.52]$ and -0.60 $[-0.61, -0.59]$, both with support intervals below $-1/2$. Squared bias plateaus near 0.002 from about 1000 rows, reproducing the qualitative bias pattern in that figure.

Over 64–1024, the variance slope is -0.48 (-0.49 in the amendment run), and the spread slope is -0.30, missing the registered ≤ -0.35 threshold. This range lies before the bias plateau. The spread statistic averages SDs over queries, whereas the variance curve averages variances, so query heterogeneity prevents treating their slopes as an exact factor-two pair.

On real data over 64–1024, v1's margin scale pair maximum exponents at $k = 32$ are unclassified: 0.42 ± 0.08 on higgs and 0.59 ± 0.04 on magic. Its two magic margin mean type exponents, 0.73 ± 0.04 and 0.69 ± 0.04, are above $1/2$. The magic probability pair maximum exponent is

Table 12: Accuracy on the first eight disjoint higgs supports at each size, with 2000 fixed held out queries. Each model has four rows: mean accuracy, 95% t interval over supports, paired gap to refit logistic regression ($C = 1$) with SE, and mean log loss. Checkpoints run in the Linux environment (p), Mitra on CPU through its 8192 cap, and refits under macOS.

		n					
Model		1024	2048	4096	8192	16,384	32,768
TabPFN v2.5^{p}	accuracy	0.688	0.704	0.712	0.716	0.720	0.728
	95% interval	[0.679, 0.697]	[0.696, 0.711]	[0.710, 0.715]	[0.709, 0.722]	[0.714, 0.725]	[0.724, 0.732]
	gap (SE)	+0.075 (0.004)	+0.077 (0.006)	+0.080 (0.003)	+0.078 (0.005)	+0.081 (0.003)	+0.090 (0.002)
	log loss	0.581	0.564	0.551	0.541	0.535	0.529
TabPFN v3^{p}	accuracy	0.693	0.712	0.715	0.727	0.732	0.739
	95% interval	[0.686, 0.700]	[0.706, 0.718]	[0.713, 0.718]	[0.724, 0.730]	[0.728, 0.736]	[0.732, 0.746]
	gap (SE)	+0.079 (0.002)	+0.085 (0.005)	+0.082 (0.003)	+0.089 (0.003)	+0.093 (0.002)	+0.101 (0.003)
	log loss	0.580	0.558	0.545	0.532	0.524	0.518
TabICLv2^{p}	accuracy	0.689	0.710	0.718	0.724	0.730	0.736
	95% interval	[0.680, 0.698]	[0.704, 0.716]	[0.714, 0.722]	[0.722, 0.727]	[0.726, 0.733]	[0.731, 0.741]
	gap (SE)	+0.075 (0.005)	+0.083 (0.005)	+0.085 (0.003)	+0.087 (0.002)	+0.091 (0.002)	+0.098 (0.002)
	log loss	0.579	0.559	0.545	0.534	0.525	0.516
TabFM^{p}	accuracy	0.695	0.715	0.727	0.735	0.740	0.748
	95% interval	[0.687, 0.702]	[0.707, 0.722]	[0.721, 0.732]	[0.731, 0.739]	[0.737, 0.742]	[0.743, 0.753]
	gap (SE)	+0.081 (0.004)	+0.088 (0.006)	+0.094 (0.004)	+0.097 (0.003)	+0.101 (0.002)	+0.110 (0.002)
	log loss	0.571	0.549	0.533	0.519	0.508	0.498
Mitra^{p} (CPU)	accuracy	0.684	0.699	0.703	0.704		
	95% interval	[0.674, 0.694]	[0.693, 0.705]	[0.697, 0.709]	[0.702, 0.706]		
	gap (SE)	+0.071 (0.005)	+0.072 (0.005)	+0.070 (0.003)	+0.066 (0.003)		
	log loss	0.584	0.569	0.566	0.563		
logreg (C=1)	accuracy	0.613	0.627	0.633	0.638	0.639	0.638
	95% interval	[0.606, 0.621]	[0.617, 0.637]	[0.627, 0.639]	[0.632, 0.643]	[0.635, 0.643]	[0.636, 0.640]
	gap (SE)	—	—	—	—	—	—
	log loss	0.657	0.648	0.644	0.641	0.640	0.640
logreg (C=10^6)	accuracy	0.614	0.627	0.633	0.638	0.639	0.638
	95% interval	[0.606, 0.622]	[0.618, 0.637]	[0.627, 0.639]	[0.632, 0.643]	[0.635, 0.643]	[0.636, 0.640]
	gap (SE)	+0.001 (0.000)	+0.000 (0.001)	+0.000 (0.000)	+0.000 (0.000)	+0.000 (0.000)	+0.000 (0.000)
	log loss	0.658	0.648	0.644	0.641	0.640	0.640
XGBoost (refit)	accuracy	0.661	0.681	0.693	0.704	0.714	0.716
	95% interval	[0.656, 0.667]	[0.674, 0.687]	[0.688, 0.698]	[0.697, 0.712]	[0.708, 0.721]	[0.713, 0.719]
	gap (SE)	+0.048 (0.003)	+0.054 (0.004)	+0.060 (0.004)	+0.067 (0.003)	+0.075 (0.002)	+0.078 (0.001)
	log loss	0.622	0.593	0.574	0.562	0.552	0.550
$\sqrt{n}$-NN (refit)	accuracy	0.581	0.594	0.607	0.613	0.624	0.627
	95% interval	[0.578, 0.584]	[0.589, 0.598]	[0.599, 0.614]	[0.606, 0.619]	[0.621, 0.628]	[0.623, 0.631]
	gap (SE)	−0.033 (0.003)	−0.033 (0.005)	−0.026 (0.005)	−0.025 (0.003)	−0.015 (0.002)	−0.011 (0.002)
	log loss	0.675	0.667	0.658	0.652	0.646	0.641

Table 13: Convex weight mixtures $\theta(a) = (1 - a)\theta_{\text{trained}} + a\theta_{\text{shuffled}}$ for Mitra on higgs. Slope: probability spread fit over 64–1024 with support interval. Flip $R(512)$: median absolute margin response at 512. The response criterion compares it with trained (1.184), recomputed in the same run. "Reads" marks the registered within-2× response criterion. Accuracy uses 200 fixed queries (binomial SE about 0.035).

a	slope	support CI	flip $R(512)$	reads	acc(512)
0 (trained)	−0.06	[−0.11, +0.02]	1.184	yes	0.69
0.125	−0.42	[−0.47, −0.38]	1.102	yes	0.72
0.25	−0.55	[−0.61, −0.49]	0.844	yes	0.54
0.5	−0.55	[−0.70, −0.42]	1.111	yes	0.51
0.75	−0.54	[−0.69, −0.43]	0.303	no	0.49
1 (shuffled)	−0.55	[−0.68, −0.44]	0.056	no	0.49

also above, 0.88 ± 0.04, the lineage's only such classification besides pol. Probability mean type exponents are 0.99 and 1.02 on magic, and Q_{mean} is marginally above on higgs at 0.64 ± 0.04.

The file dated dimension experiment (DS) appends noise coordinates without changing the five-dimensional target. For $d = 5, 10, 28$, post-bend slopes are -0.54, -0.53, and -0.59. Initial

window slopes are -0.30, -0.07, and $+0.16$. The prediction that $d = 28$ would not reach the rate fails.

Finite pool overlap. Supports are sampled without replacement from training pools of 15,216 rows for magic and 40,000 for the initial higgs experiments. Two independently sampled supports share an expected fraction n/pool. Stored indices give 0.5–6.8% overlap over magic's initial range, then 13.5%, 26.9%, and 53.8% at 2048, 4096, and 8192. Higgs overlap ranges from 2.6 to 20.5% over its extension. Finite pool sampling can change across support variance. The familiar finite population correction is exact for a sample mean, but is not established here for a nonlinear predictor.

The correction models variance attenuation as $(1 - f(n))^{\kappa}$. This κ is distinct from the aggregation quantity $\kappa(S)$ in Proposition 1. We use $\kappa = 1$, corresponding to the sample mean correction, as one sensitivity setting. The available argument does not establish a universal upper bound on κ for the measured predictors conditional on these realized data pools. The correction adds $-(\kappa/2)\log(1 - f(n))$ to log SD, increasing the extension slope by 0.164κ on magic and 0.048κ on higgs. In particular, we do not assume that this setting bounds the correction for order sensitive TabPFN. The disjoint measurements provide a comparison that does not require this correction.

The registered partition experiment compares across block spread from random disjoint partitions with the stored overlapping runs at matched n (Table 14). On magic, its registered pooled estimates are $\kappa = 1.40$ $[0.87, 2.14]$ for v2.5 and 1.28 $[0.99, 1.80]$ for v1, both point estimates above one. We retain these fitted estimates alongside the $\kappa = 1$ sensitivity results. The v1-magic rate prediction remains a failure under its registered calculation. At $\kappa = 1$, higgs slopes change from -0.42 to -0.37 for v1 and -0.12 to -0.08 for v2.5.

For v1-magic's corrected interval to lie wholly below $-1/2$ would require $\kappa \leq 0.62$. This is below the available point estimates: within arm per-size values of 0.73–0.87, the registered pooled estimate, and the post hoc logistic regression calibration. That calibration uses logistic regression's measured -0.62 $[-0.66, -0.59]$ over 1024–8192 and asks which correction returns it to -0.5, giving 0.76 $[0.52, 1.00]$. Across the examined correction range, v1-magic slopes run from -0.52 to -0.475, with resampling fractions $P(\text{attain})$ from 0.83 to 0.04. V2.5-magic runs from -0.39 to -0.34.

Table 14. The uncorrected magic extension slopes are -0.64 for v1 and -0.51 for v2.5, with supports sharing up to 54% of the pool. The registered pooled v1 estimate $\kappa = 1.28$ would give -0.43 $[-0.49, -0.33]$. Three other readings are reported: correction at $\kappa = 1$, post hoc within arm pooling, and a direct fit to disjoint partitions over 1024–4096. For v2.5 these give, respectively, -0.34 $[-0.38, -0.31]$, -0.38 $[-0.41, -0.34]$ at $\hat{\kappa}_w = 0.775$, and -0.31 $[-0.35, -0.27]$. All intervals lie above $-1/2$.

For v1, $\kappa = 1$ gives -0.475 $[-0.504, -0.437]$, with $P(\text{attain}) = 0.04$. Within arm pooling gives -0.50 $[-0.53, -0.46]$ at $\hat{\kappa}_w = 0.837$, with $P(\text{attain}) = 0.39$. Both intervals contain $-1/2$. The direct disjoint fit gives -0.45 $[-0.49, -0.41]$, above the boundary. Magic uses 4×29, 7×14, 14×7, and 32×3 partitions times blocks per size. Higgs uses 6×19, 11×9, and 12×4. For v2.5, the fingerprint also reduces spread at the two largest magic sizes (Appendix G).

Local slopes and a possible bend. Two post hoc analyses use the stored disjoint higgs series (DH), without new model calls or scored predictions. V2.5's consecutive octave slopes begin at $+0.03$ $[-0.18, +0.25]$ over 64–128, followed by -0.04, -0.20, -0.17, -0.09, -0.04, and -0.03. The last two are -0.21 $[-0.33, -0.13]$ over 8192–16,384 and -0.29 $[-0.47, -0.18]$ over 16,384–32,768. Intervals use 2000 paired draw index resamples. A continuous two-piece log-log fit places a free knot at 13,500 rows, but its interval spans the grid. The second slope is -0.29 $[-0.47, -0.13]$, above $-1/2$ in 98.7% of resamples. The fit improves AIC by only 3.2. Forcing the second slope to $-1/2$ places the knot at 21,000 (12,500–24,900 under resampling), beyond the registered 1024–16,384 fit window but inside the six-point extension series. This constrained fit is compatible with a transition near 2×10^4 rows. It neither locates nor establishes one.

The smoothness comparison simply relabels a slope magnitude $s < 1/2$ as $\beta = ds/(1 - 2s)$ under the rate $n^{-\beta/(2\beta+d)}$. Most initial higgs slopes imply β between 2 and 4.4, corresponding to SD exponents 0.06–0.12 at $d = 28$ (0.06–0.11 for $\beta \leq 4$), or 0.08–0.15 at $d = 21$. The exceptions are TabICLv2's initial window ($\beta = 18$, falling to 3.5 in the disjoint continuation), Mitra's disjoint

Table 14: Disjoint partition versus overlapping spread. $\bar{f}$: mean shared fraction of original supports; ratio: disjoint to overlapping SD with resampling 95% interval; lin.: $(1-\bar{f})^{-1/2}$. $\hat{\kappa} = \log(V_{\text{disj}}/V_{\text{ovl}})/[-\log(1-\bar{f})]$ uses query mean variances V; within: the secondary within arm estimate. Pooled rows give inverse variance weighted log-ratio fits over 512–4096 (magic) or 2048–8192 (higgs), with 95% intervals and corrected extension slopes (original slopes in parentheses).

Model	data	n	$\bar{f}$ (%)	ratio [95%]	lin.	$\hat{\kappa}$	within
v2.5	magic	512	3.4	1.093 [1.04, 1.18]	1.017	5.17	0.727
v2.5	magic	1024	6.8	1.009 [0.94, 1.10]	1.036	0.26	0.785
v2.5	magic	2048	13.5	1.118 [1.05, 1.22]	1.075	1.54	0.770
v2.5	magic	4096	26.9	1.228 [1.11, 1.38]	1.170	1.31	0.793
v1	magic	512	3.4	1.059 [1.02, 1.13]	1.017	3.31	0.735
v1	magic	1024	6.8	1.016 [0.96, 1.10]	1.036	0.46	0.819
v1	magic	2048	13.5	1.105 [1.05, 1.19]	1.075	1.38	0.835
v1	magic	4096	26.9	1.217 [1.15, 1.32]	1.170	1.25	0.868
v2.5	higgs	2048	5.1	1.063 [1.02, 1.13]	1.027	2.33	0.774
v2.5	higgs	4096	10.2	1.075 [1.04, 1.14]	1.055	1.34	0.786
v2.5	higgs	8192	20.5	1.103 [1.06, 1.17]	1.121	0.85	0.804
v1	higgs	2048	5.1	1.019 [0.98, 1.08]	1.027	0.74	0.914
v1	higgs	4096	10.2	1.052 [1.02, 1.11]	1.055	0.94	0.933
v1	higgs	8192	20.5	1.168 [1.11, 1.25]	1.121	1.35	0.973

Registered pooled $\hat{\kappa}$ [95%] and its corrected extension slope [95%] (measured slope in parentheses):

Model	data	Pooled $\hat{\kappa}$	Corrected slope
v2.5	magic	1.40 [0.87, 2.14]	−0.28 [−0.38, −0.14] (−0.51)
v1	magic	1.28 [0.99, 1.80]	−0.43 [−0.49, −0.33] (−0.64)
v2.5	higgs	0.99 [0.72, 1.51]	−0.08 [−0.10, −0.04] (−0.12)
v1	higgs	1.21 [0.87, 1.83]	−0.36 [−0.39, −0.32] (−0.42)

window (86), v1's beyond cap disjoint window (63), the v3 and TabFM disjoint windows (5.5 and 4.5), and Mitra's initial window (1.9). Magic's initial slopes imply β from 6 to 31, similar to the refit references. These values are not independent estimates of data smoothness. The underlying theorems concern risk minimizing transformers on specified function classes.

E Fit windows and multiplicity

Table 3 reports mean pair influence and maxima over all observed deletions. Table 15 retains the registered primary, the $k = 32$ sampled maximum, on both fit windows. All are finite window margin scale summaries, rather than tests of Theorem 5.2's asymptotic probability scale condition. The TabPFN runs use one estimator with the deletion scheme selected by the order probe. Fingerprint-off rows are labelled separately. Mitra uses the deterministic harness. On higgs, the low exponent foundation model intervals overlap each other's and XGBoost's, which does not establish equality. Logistic regression and the MLP span the range of mean influence exponents.

Table 3. Mean pair exponents change by at most 0.04 in the paired $k = 32$ and $k = n$ foundation model reruns, over 64–512 or 64–256 for Mitra on higgs. Table 3's choice of columns was made after registration. The 64 and 128 cells are reported under the registration's exception for full window fits. Tables 16 and 15 retain the registered windows. Table 17 gives the probability scale mean pair exponents and paired configuration differences.

The superscript h marks cuts surviving the mapped Holm correction detailed below. Unmarked cuts either fail correction or fall outside that family. The three mapped fingerprint-off $k = n$ maximum cuts (-5.0, -4.1, -2.5) fail correction. Four probability scale $k = n$ means survive: TabICLv2 on both datasets, fingerprint-off v3 on magic, and Mitra on magic. Configuration differences use seed paired unrounded exponents, so their t values need not follow from rounded cells. They remain conditional on each configuration's query selection.

Mitra's $k = n$ window is 64–256 on higgs and 64–512 on magic. Its magic 512 cell, initially omitted under a cost limit, was completed under macOS in a separately registered run. The vendor default TabPFN ensemble agrees within 1.1–1.5 pooled SE on higgs and in classification on magic. The magic full window point is 1.7 pooled SE above the single estimator's. Its registered higgs comparison, 0.55 ± 0.11 against 0.38 ± 0.11, misses one prediction. K11's magic predictions have four hits in seven and are recorded without independently verified prior timing. K12 assigned probability 55% to v3-magic falling below $1/2$: the sampled maximum is 2.2 SE below, whereas the mean pair exponent is 0.313 ± 0.058, 3.2 SE below. K8's prediction of v2.5 decay near 1 misses by 3.9–12.4 SE on the registered statistic. The v3 higgs cells marked p use file dated registrations.

Multiplicity. Influence classifications use $|t| \geq 3$ with five degrees of freedom (approximately $p = 0.03$, two-sided). The registered `mult_score.py` analysis also applies Holm correction at family-wise 0.05 to 154 mapped comparisons; 83 pass the uncorrected threshold and 27 survive Holm. The corresponding Bonferroni threshold is approximately $|t| \geq 8.7$. Holm uses stepwise thresholds. Spread intervals are per-arm 95% intervals without simultaneous coverage and are outside this family. Those closest to the reference are TabFM-magic, TabICLv2-jm1, and v2.5-wilt.

The complete comparison ledger, `scripts/exp/mult_score.json`, records every mapped statistic, scale, dataset, fit, t value, and family- and within table Holm decision. It includes original and later checkpoint runs, all-deletion comparisons, the twelve-dataset and fingerprint-off extensions, and specified vendor defaults. The 27 survivors include classical references, TabICLv2 means and sampled maxima, shipped TabPFN magic all-deletion maxima, and selected probability scale comparisons. Only two fingerprint-off comparisons survive the full family: v3-magic's all-deletion probability mean and the credit/GMSC mean. Three survive within their own table only: v2.5-magic's all-deletion probability mean, credit/GMSC's sampled maximum, and phoneme's mean. The shipped TabPFN main window mean and sampled maximum below classifications do not survive either correction. Neither do Mitra's and TabFM's magic sampled margin above classifications or pol's probability classification.

Unmapped analyses remain distinct: K13 sub-windows, fixed query and probability follow-ups, fingerprint-off v2.5-higgs at $k = n$ (four comparisons, none meeting 3 SE), the fingerprint-off $k = 32$ means, and most of Table 17. That table's exceptions are TabFM's two cells and shipped v3-higgs. The family also excludes 58 post hoc registered window fits (12 meeting 3 SE) and 40 RMS comparisons over two scales and 20 scored arms. The latter are retained in `scripts/exp/rms_score.json` and the separately expanded multiplicity output. The common query intervention and classical all-deletion completion each use a separate four-test family (Tables 2 and 19).

Above boundary shipped TabPFN exceptions depend on the scale, statistic, and window: pol's probability sampled maximum and v1's magic mean statistics are examples (Appendix D). The full ledgers preserve classifications that are not enumerated here. No historical decision is changed.

E.1 Sampled maxima: base and extended windows

Table 15. The vendor default magic runs give mean pair exponents of 0.915 (0.031) for TabICLv2's eight estimators and 0.781 (0.038) for TabFM's 32 views ($k = 32$, 64–512). Their differences from the single estimator or single view runs are -0.058 and $+0.018$, or 1.2 and 0.2 pooled SE. All four predictions pass the registered 0.15 and addendum's 1.5-pooled-SE bands. The exponents exceed $1/2$ by 13 and 7 SE. TabICLv2 survives the full Holm correction, TabFM only the within table correction.

Window changes affect both estimates and precision (Tables 15 and 16). Full window classifications retain direction on the two-point registered window but usually lose classification. Only shipped v3-higgs remains classified among checkpoint sampled maxima ($t = -3.35$). Mitra-higgs, the vendor ensemble on higgs, and the MLP on magic cross $1/2$ at the point estimate. The MLP's registered higgs exponent is negative. V2.5-magic classifies below on the four-point extended registered window (5.4 SE). A post hoc mean fit classifies the vendor ensemble on magic below (0.172 (0.090)). Ten of twenty-six all-deletion registered window fits classify. These fits are not interchangeable estimates of one power law.

Table 15: Margin scale sampled pair maximum exponents ($k = 32$), mean (SE) over six seeds. Full fits use 64–512. Registered fits use the two points at 256 and 512. TabPFN uses shipped preprocessing. Its vendor ensemble row uses eight seeded estimators. AP1 vendor default rows use eight TabICLv2 estimators or 32 TabFM views on magic only. p denotes the Linux environment. V3-magic's original full window value, 0.354 (0.065), $t = -2.2$, becomes 0.348 (0.048), $t = -3.2$, on that platform, the sole platform dependent full window classification on the registered statistics.

	higgs		magic	
Model	full	registered	full	registered
logreg (refit)	1.335 (0.010)	1.230 (0.144)	0.983 (0.079)	1.112 (0.208)
TabICLv2	0.952 (0.052)	0.798 (0.128)	0.827 (0.037)	0.726 (0.225)
TabFM	0.610 (0.084)	0.588 (0.066)	0.736 (0.057)	0.853 (0.246)
TabPFN v2.5	0.381 (0.112)	0.430 (0.122)	0.301 (0.040)	0.255 (0.093)
v2.5, vendor ens.	0.553 (0.113)	0.454 (0.115)	0.385 (0.031)	0.190 (0.118)
TabICLv2, vendor default (8 est.)	—	—	0.783^{p} (0.032)	0.671^{p} (0.251)
TabFM, vendor default (32 views)	—	—	0.770^{p} (0.044)	0.775^{p} (0.178)
TabPFN v3	0.337^{p} (0.078)	0.050^{p} (0.134)	0.354 (0.065)	0.364 (0.176)
Mitra	0.292 (0.140)	0.594 (0.161)	0.703 (0.048)	0.659 (0.224)
XGBoost (refit)	0.193 (0.030)	0.130 (0.022)	0.189 (0.029)	−0.037 (0.101)
MLP (refit)	0.130 (0.030)	−0.208 (0.059)	0.561 (0.068)	0.236 (0.222)

Mitra's trained and randomized higgs runs extend to 1024 and 2048, making the registered $n \geq 256$ window a four-point fit. The trained mean influence exponent 0.867 ± 0.077 misses the registered $[0.4, 0.8]$ band, and its paired difference from the control reaches only $t = 1.6$, below 3. Under the registered fallback, no slope conclusion is drawn. Trained to control level ratios remain 18.7, 20.0, 49.0, 62.8, 57.2, and 67.0 from 64 to 2048 rows, but the control's weak label response prevents attribution. The same registration's fourth prediction concerns TabICLv2's trained minus trunk control slope over $n \geq 256$. It fails on magic ($+0.109 \pm 0.032$, $t = 3.4$). The higgs difference is consistent with zero (-0.010 ± 0.018).

Registration and query sensitivity. The 2026-07-31 registration designated $n \geq 256$ as primary and required full window fits alongside it. The 2026-08-01 replication registration allowed both windows if labelled. We report full window classifications as such when they do not survive the stricter window. Disagreement beyond the intervals is itself a registered result. Table 15 applies these rules to every shipped preprocessing arm in Table 3.

Freezing the queries to the clean-correct selection from $(n = 64, \text{seed } 0)$ changes TabPFN-magic's exponent from 0.301 ± 0.040 to 0.292 ± 0.037 and Mitra-higgs's from 0.292 ± 0.140 to 0.271 ± 0.135. Both changes are within the registered two-pooled-SE bands. This within configuration check does not resolve the different query populations in the fingerprint-on/off comparison. Mitra's local signed slopes from 64 to 2048 are -1.044, $+0.259$, -0.645, -1.103, and -0.757, a range of 1.36. The increasing segment and changing decay rates explain its sensitivity to the fit window.

Table 16. Extended arms use 64–2048, rather than Table 3's 64–512 window. TabFM's Linux reruns store both scales and reproduce the margin fits to 2×10^{-5}. V3's higgs row uses the Linux run. Its magic row combines the original margin fit with the probability fit from its Linux replication. Above boundary entries describe sampled maxima ($k = 32$). Every foundation model and logistic regression was also measured at $k = n$ over 64–512 (64–256 for Mitra on higgs). No foundation model margin scale above classification survives on that window (Table 18). The longer windows were not rerun at $k = n$. A separately registered completion measures histogram XGBoost and the MLP over 64–512 (Table 19).

On the extended registered window, v2.5's margin exponents are 4.1 SE (higgs) and 5.4 SE (magic) below $1/2$. Its probability exponent classifies below on the full higgs window (3.5 SE) and registered magic window (3.1 SE). The other two comparisons are below at 1.4 SE (full magic) and 2.5 SE (registered higgs). The predicted full window magic classification fails. TabICLv2 remains above on both scales and windows, whereas the MLP's maximum influence exponent is negative on both datasets and scales over the extended registered window.

Table 16: Sampled pair maximum exponents ($k = 32$), mean (SE) over six seeds. Full/registered windows are 64–2048/256–2048 for v2.5, TabICLv2 and the three refits. Mitra, TabFM and v3 use 64–512/256–512. No fingerprint-off maximum is included here. Bold marks $|t| \geq 3$ against $1/2$. p denotes Linux. Main window probability means are in Table 17.

		Margin		Probability	
Model	data	full	registered	full	registered
TabPFN v2.5	higgs	**0.261** (0.060)	**0.175** (0.079)	**0.339** (0.046)	0.303 (0.079)
TabPFN v2.5	magic	**0.293** (0.021)	**0.270** (0.043)	0.471 (0.021)	**0.347** (0.049)
TabICLv2	higgs	**0.835** (0.030)	**0.711** (0.032)	**0.882** (0.022)	**0.793** (0.033)
TabICLv2	magic	**0.785** (0.040)	**0.738** (0.064)	**0.931** (0.035)	**0.798** (0.061)
Mitra (no ext.)	higgs	0.292 (0.140)	0.594 (0.161)	0.363 (0.132)	0.710 (0.163)
Mitra (no ext.)	magic	**0.703** (0.048)	0.659 (0.224)	**0.925** (0.033)	0.698 (0.255)
TabFM^{p}	higgs	0.610 (0.084)	0.588 (0.066)	0.649 (0.064)	0.675 (0.071)
TabFM^{p}	magic	**0.736** (0.057)	0.853 (0.246)	**1.017** (0.036)	0.892 (0.238)
TabPFN v3^{p}	higgs	0.337 (0.078)	**0.050** (0.134)	0.374 (0.049)	0.136 (0.144)
TabPFN v3	magic	0.354 (0.065)	0.364 (0.176)	0.598^{p} (0.064)	0.294^{p} (0.074)
logreg	higgs	**1.210** (0.019)	**1.099** (0.044)	**1.087** (0.016)	**1.086** (0.036)
logreg	magic	**0.965** (0.060)	**0.966** (0.122)	**0.982** (0.054)	**0.912** (0.121)
XGBoost	higgs	**0.241** (0.012)	**0.277** (0.011)	**0.169** (0.020)	**0.202** (0.021)
XGBoost	magic	**0.165** (0.011)	**0.132** (0.023)	**0.161** (0.031)	**−0.009** (0.077)
MLP	higgs	**−0.079** (0.010)	**−0.309** (0.016)	**−0.030** (0.013)	**−0.338** (0.035)
MLP	magic	**0.257** (0.045)	**−0.052** (0.040)	**0.298** (0.044)	**−0.091** (0.097)

Table 17: Probability scale mean pair influence, $k = 32$, 64–512, six-seed mean (SE). These share the main margin mean table's draws. Fingerprint-off differences pair seeds but condition on each configuration's own correctly classified queries. These comparisons are outside the historical mapped family except TabFM's two cells and shipped v3-higgs. No new significance annotations are added.

Model/configuration	data	$\alpha_p(Q_{\text{pair}})$ or paired conditional Δ
TabPFN v2.5	higgs	0.406 (0.100)
TabPFN v2.5	magic	0.534 (0.053)
TabPFN v2.5, fp off	higgs	0.589 (0.108)
TabPFN v2.5, fp off	magic	1.116 (0.091)
v2.5, fp off minus shipped	higgs	+0.18 (0.02)
v2.5, fp off minus shipped	magic	+0.58 (0.05)
TabICLv2	higgs	1.034 (0.049)
TabICLv2	magic	1.192 (0.064)
Mitra (no ext.)	higgs	0.415 (0.131)
Mitra (no ext.)	magic	1.006 (0.036)
TabFM^{p}	higgs	0.689^{p} (0.065)
TabFM^{p}	magic	1.070^{p} (0.060)
TabPFN v3^{p}	higgs	0.314^{p} (0.072)
TabPFN v3	magic	0.576^{p} (0.058)
TabPFN v3, fp off	higgs	0.611^{p} (0.105)
TabPFN v3, fp off	magic	1.085 (0.053)
v3, fp off minus shipped	higgs	+0.30^{p} (0.09)
v3, fp off minus shipped	magic	—
logreg	higgs	1.009 (0.028)
logreg	magic	1.057 (0.081)
XGBoost	higgs	0.065 (0.049)
XGBoost	magic	0.137 (0.050)
MLP	higgs	0.159 (0.060)
MLP	magic	0.524 (0.123)

E.2 All-deletion maxima, means and historical RMS

Table 18. Among the checkpoint configurations, only shipped v2.5 classifies below $1/2$ on magic's probability scale all-deletion maximum (0.29, $t = -4.6$). Fingerprint-off v2.5 and v3,

Table 18: Sampled ($k = 32$) and all-deletion ($k = n$) margin scale exponents on matched windows through 512. Mitra-higgs and Mitra-magic RMS fits end at 256. The latter's 512 matrices are on a different machine. t compares the all-deletion fit with $1/2$ over six seeds. h marks Holm survivors and p Linux measurements. RMS values are historical scored summaries. The 48 original TabICLv2 matrices are included; other foundation model deletion matrices are outside this supplement. Original shipped-v2.5 rows remain dashed because the matrix storing reruns differ and appear separately. Cross-platform rows (v3-magic and TabFM-higgs) use same machine counterparts for paired sampling comparisons. The notes give completion and replication details.

		$Q_{\text{pair}}^{\max}$		Q_{pair}		$Q_{\text{pair}}^{\text{rms}}$
Arm	data	k=32	k=n (t)	k=32	k=n (t)	k=n (t)
TabPFN v2.5	higgs	0.38	0.06 (-4.4)	0.37	0.37 (-1.0)	—
TabPFN v2.5$^{\text{p}}$	higgs	0.33	0.08 (-3.3)	0.35	0.36 (-1.0)	0.36 (-1.0)
TabPFN v2.5	magic	0.30	0.03 (-21.2^{h})	0.28	0.27 (-5.5)	—
TabPFN v2.5$^{\text{p}}$	magic	0.30	0.04 (-22.0^{h})	0.29	0.27 (-5.4)	0.27 (-6.5)
v2.5 defing.	higgs	0.53	0.22 (-2.0)	0.56	0.56 ($+0.4$)	0.54$^{\text{p}}$ ($+0.3$)
v2.5 defing.	magic	0.73	0.22 (-5.0)	0.83	0.80 ($+4.7$)	0.67$^{\text{p}}$ ($+2.9$)
TabPFN v3$^{\text{p}}$	higgs	0.34	0.10 (-5.0)	0.25	0.24 (-2.6)	0.28 (-2.4)
v3 defing.$^{\text{p}}$	higgs	0.53	0.17 (-2.5)	0.54	0.53 ($+0.3$)	0.51 ($+0.1$)
TabPFN v3	magic	0.35	0.09$^{\text{p}}$ (-8.9^{h})	0.31	0.29$^{\text{p}}$ (-3.7)	0.30$^{\text{p}}$ (-3.8)
v3 defing.	magic	0.74	0.25$^{\text{p}}$ (-4.1)	0.81	0.80$^{\text{p}}$ ($+5.8$)	0.70$^{\text{p}}$ ($+4.0$)
TabICLv2	higgs	0.95	0.55 ($+1.2$)	1.02	1.01 ($+8.3$)	0.96 ($+8.2$)
TabICLv2	magic	0.83	0.33 (-4.4)	0.97	0.98 ($+9.8^{\text{h}}$)	0.83 ($+7.4$)
Mitra	magic	0.70	0.17 (-6.6)	0.78	0.76 ($+5.4$)	0.60$^{\text{p}}$ ($+1.8$)
Mitra	higgs	0.25	-0.05 (-2.8)	0.31	0.30 (-1.3)	0.27 (-1.4)
TabFM	higgs	0.61	0.31$^{\text{p}}$ (-2.3)	0.65	0.65$^{\text{p}}$ ($+1.8$)	0.63$^{\text{p}}$ ($+1.5$)
TabFM	magic	0.74	0.31 (-3.4)	0.76	0.75 ($+3.2$)	0.69$^{\text{p}}$ ($+2.9$)
logreg (refit)	higgs	1.34	1.09 ($+27.0$)	1.25	1.23 ($+42.9$)	1.26 ($+53.5$)
logreg (refit)	magic	0.98	0.65 ($+1.9$)	0.99	0.95 ($+16.1$)	0.92 ($+9.9$)

TabICLv2, and Mitra give 0.51, 0.53, 0.54, and 0.40, all unclassified. TabFM gives 0.61 ($t = +5.0$), above on the full window but outside the mapped family. Mitra's unrounded maximum $t = -2.97$ remains unclassified despite rounding to -3.0. Its 64–512 probability mean exponent is 0.98 ($t = +16.8$), and its 512-row RMS level is 0.057 ± 0.003.

Increasing k lowers every maximum exponent and removes every foundation model margin scale above classification in these comparisons. Logistic regression on higgs remains above ($1.34 \to 1.09$, $t = +27.0$). Mean pair exponents change little and retain their classifications for the foundation models and logistic regression, whose paired changes are -0.02 and -0.04. The 5-NN mean changes more because a 32-row deletion sample often misses influential neighbours (Appendix F).

The recorded RMS analysis used TabICLv2 and Linux deletion matrices. The 48 original TabICLv2 matrices now included reproduce every recorded cell and fit. The other RMS summaries remain in the release, but their underlying matrices are not included. Table 18 gives the margin results. `scripts/exp/rms_score.json` retains both scales and every comparison. The registered probability scale RMS limits, 0.78 on higgs and 0.87 on magic, are exceeded (R4, miss).

A rerun fills an existing row only if both $k = n$ exponents round to the printed values and preserve the paired difference's sign. TabFM-magic, Mitra-magic, and both fingerprint-off v2.5 reruns reproduce the two exponents within 2×10^{-6}. Shipped v2.5 reruns have relative cell differences up to 0.27 and 0.06 and appear separately. Mitra-magic's maximum and mean fits include the macOS completion at 512, with sampled values 0.703 and 0.777. Its RMS fit retains the Linux environment's 64–256 range, because the 512 matrices are under macOS and these derived statistics do not mix machines.

Registered classical all-deletion completion. A separate protocol, hashed and locally timestamped before measurement on 21 September 2026, fills the four previously unmeasured XGBoost and MLP all-deletion entries in Table 3. The historical constructors are unchanged: CPU histogram XGBoost 3.2.0 with 100 depth-4 trees and learning rate 0.1, and a standardized 64-hidden-unit

Table 19: Registered completion of all-deletion classical influence on 64–512 rows, six seeds and up to 50 clean-correct queries per cell. Entries are mean exponent (SE), with margin-scale maximum t against $1/2$ and Holm-adjusted p over these four new primary tests. This family is separate from the historical mapped correction. All statistics are finite-window estimates.

		margin		probability			
Arm	data	mean	maximum	mean	maximum	$t_{\max}$	p_{Holm}
XGBoost hist.	higgs	0.147 (0.032)	0.087 (0.024)	0.065 (0.046)	0.017 (0.035)	−17.35	$< 10^{-4}$
XGBoost hist.	magic	0.050 (0.023)	0.057 (0.020)	0.118 (0.043)	0.152 (0.029)	−22.03	$< 10^{-4}$
MLP	higgs	−0.024 (0.031)	−0.029 (0.034)	0.177 (0.061)	0.155 (0.059)	−15.71	$< 10^{-4}$
MLP	magic	0.411 (0.040)	0.295 (0.063)	0.536 (0.121)	0.507 (0.099)	−3.28	0.0220

MLP with scikit-learn 1.8.0 defaults and a 600-iteration cap. Each refit uses its support seed. This is not the exact tree spread configuration. Both datasets use the original small context split, $n \in \{64, 128, 256, 512\}$, six seeds, and the original first 50 clean-correct queries from each seeded candidate pool. Deleting every row requires 23,040 deletion refits plus 96 full support fits, executed locally on CPU. Full probability and margin matrices, support/query/deletion indices, constructor settings and warnings are archived.

For scalar outputs, the query-wise pair maximum equals the range of all deletion predictions. The reported statistic averages these ranges across queries. Negated log-log slopes are fitted separately for each seed over all four sizes. Table 19 reports the six-seed mean and SE. The four margin maximum comparisons use a predeclared, separate Holm family of two-sided t_5 tests. These new tests do not alter the historical 154-test mapped family or receive its h marker. Boldface in Table 3 continues to denote $|t| \geq 3$. All four margin maximum exponents are below $1/2$ after this separate correction. The recorded completion retains all cells and its prespecified fit window and model settings. The finite window exponents do not establish an asymptotic rate.

The MLP records 10,221 convergence warnings across 11,568 fits (including bases); 10,250 fits have `n_iter_=600`. These are different quantities. All recorded warnings are retained, and the iteration limit was not tuned. Its exponents therefore describe this fixed budget optimization procedure, including optimization sensitivity. The original $k = 32$ deletion subsets reconstructed from the full matrices reproduce all historical three-decimal exponent means and SEs on both scales. The maximum cell statistic discrepancy is 2.2×10^{-16}. This checks the matched construction without changing the historical sampled entries. The original records lack stored support/query indices. The new indices are checked against a separate reconstruction of the historical RNG stream, rather than against unavailable old index arrays.

F Classical refit references

Three classical models were refit on every support and deletion using the same protocol. Logistic regression standardizes features and uses sklearn's default L_2 penalty. The MLP has one hidden layer of 64 units. XGBoost uses 100 depth-4 trees with learning rate 0.1. Each refit uses the cell's seed. For the $k = 32$ sampled pair maximum over 64–512, logistic regression gives $\alpha = 1.335 \pm 0.010$ on higgs and 0.983 ± 0.079 on magic, missing the registered higgs validation prediction of an exponent near 1. Higgs exponents are 0.193 ± 0.030 for XGBoost and 0.130 ± 0.030 for the MLP. Slow influence decay therefore also occurs in refitted classical models.

Both additional anchors in the 2026-08-04 registration miss their predictions. With $k = 32$ sampled deletions, 5-NN gives mean exponents of 1.22 ± 0.26 (higgs) and 1.49 ± 0.25 (magic), and pair maximum exponents of 0.98 ± 0.24 and 1.19 ± 0.29, against a predicted near-zero band. Sparse influential neighbours are often absent from a fixed deletion sample. The registered all-deletion rerun gives pair maximum exponents of -0.17 ± 0.06 and 0.15 ± 0.10, below $1/2$ by 10.5 and 3.6 SE. Mean pair exponents are 0.86 ± 0.15 and 1.21 ± 0.09. Logistic regression's corresponding all-deletion maxima are 1.09 ± 0.02 and 0.65 ± 0.08, and its means 1.23 and 0.95. The 100-tree random forest was predicted to exceed XGBoost with probability 55% but falls below it by 6–8 SE: $\alpha(Q_{\text{pair}}^{\max}) = -0.01 \pm 0.01$ on higgs and -0.53 ± 0.09 on magic. Deleting a row changes the bootstrap samples. How this statistic relates to the bagging stability results of Soloff et al. (2024) remains open.

The 5-NN comparisons do not establish a nearest-neighbour explanation for TabPFN. At $k = 32$, its pair maximum exponent differs from TabPFN's 0.30–0.38 by 2.3–3.0 pooled SE (3.2–4.2 for Q_{mean}). At $k = n$, the maximum gap falls to 0.6–2.6 pooled SE against shipped and fingerprint-off v2.5, while the mean pair gap is 3.8–9.6 SE on magic and 1.5–2.5 on higgs. Appendix I reports the separate prior mismatch and attack cost tests.

Regularization on magic (MA). On the stored magic draws, logistic regression with $C = 10^6$ and `max_iter=5000` gives spread slope -0.568 $[-0.60, -0.53]$ over 64–1024, or -0.557 $[-0.59, -0.52]$ under the $\kappa = 1$ correction. An n-scaled penalty ($C = n$) gives -0.556 $[-0.59, -0.52]$, or -0.545 after correction. Both intervals lie below $-1/2$. Descriptive 1024–4096 fits, whose draws share up to 27% of the pool, are -0.66 for both configurations, -0.58 under the $\kappa = 1$ correction, and -0.59 at the pooled within arm $\kappa = 0.81$. The default penalty slope is -0.45 $[-0.48, -0.42]$, so its failure to reach the reference rate depends on regularization.

G Inference stack effects

The initial Mitra run through AutoGluon 1.5.0 gave a positive Q_{mean} slope of $+0.58$ over $n \in \{64, 128, 256\}$ on two completed seeds. This result was excluded after the noise diagnosis below. The arm had scheduled three seeds and $k = 128$, capped at n. The relevant inference stages are unchanged through AutoGluon 1.6.1.

Three stages add variation between the input arrays and the transformer: `random_mirror_x` draws feature sign flips from the unseeded global RNG at each fit, even with the default single estimator; each prediction permutes the support; and autocast uses bfloat16 (unit roundoff 3.9×10^{-3}). Repeated fit and predict calls on identical arrays change margins by up to 0.19 at $n = 64$ and 0.56 at $n = 256$, compared with a median margin of 0.63 at the latter size. Mean pairwise margin differences across refits are 0.050 in bfloat16 and 0.040 in fp32 at 64, rising to 0.128 and 0.109 at 256. At 256, this floor lies between the two excluded influence measurements, 0.11 and 0.15. Its fitted slope is about $+0.68$, accounting for the apparent increasing influence (Figure 4). Fp32 removes about a fifth of the noise. Unseeded augmentation is the larger source.

The deterministic harness disables augmentation, fixes caller controlled row order, and uses fp32, with unchanged weights. It passes the registered margin scale checks of repeat variation below 10^{-4} and order variation below 3×10^{-3}: repeated calls agree exactly and order variation is 4.7×10^{-6}, compared with TabICLv2's 1.5×10^{-5}. Mitra has no positional encoding over support rows. The residual permutation effect is numerical.

Historical fingerprint controls. The shipped/off influence comparisons here use the configuration specific correctness filters defined in Section 2. The frozen stack contrast below is the same query exception. TabPFN v2.5 and v3 append a SHA-256 fingerprint of each row's preprocessed features by default, to distinguish duplicate rows. From package 6.4.0, the hash is salted by the cell count $n \times d$. Packages 2.0.5–6.3.2 used a seed keyed salt and 2.0.0–2.0.4 a process salted builtin hash. V1 has no fingerprint step. The vendor reports do not describe the retained row deletion mechanism (Grinsztajn et al., 2025; 2026). The changelog records deterministic hashing at the TabPFN-3 transition (Prior Labs, 2026). The package supplies no pretraining code, so we cannot determine whether pretraining used the feature.

At the $n = 256$ diagnostic, both single deletions and two calls deleting different rows retain 0.0% of fingerprints, including with fixed salt. Same size deletion pairs therefore do not cancel the retained row perturbation. The step is byte-identical and enabled by default in packages 8.2.0 and 8.5.0 (released 2026-08-27). A diagnostic under package version 8.5.0 repeats the fingerprint survival test and reproduces the magic influence cells bit-for-bit. It scores no registered unit.

The historical study compares five configurations on both datasets and scales (with registration timing limits in Appendix K): shipped preprocessing (0), fingerprint off (A), fixed hash salt (B), SVD removed (C), and all fitted preprocessing frozen at the base call (D). Repeat variation is exactly zero throughout. The CPU baseline reproduces every production cell at $z \leq 0.01$ (Appendix L). At $n = 512, 1024, 2048$, shipped Q_{mean} levels on higgs are 0.060, 0.061, 0.060, versus 0.026, 0.021, 0.011 with the fingerprint off. Magic levels are 0.17, 0.14, 0.13, versus 0.020, 0.019, 0.008. The slower decay of the shipped levels flattens the fitted exponent.

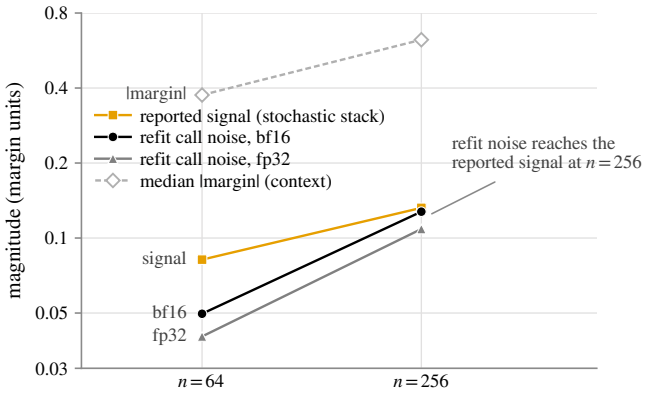


Figure 4: Repeat call variation in the shipped Mitra stack versus its initially reported influence. At 256 rows, the maximum repeat call margin change (0.56) approaches the median margin (0.63). Fp32 reduces rounding error. Unseeded augmentation accounts for most of the floor.

Table 20: Historical TabPFN v2.5 fingerprint comparison, one estimator. Magic influence exponents over 64–512 rows are means (SE) over six seeds; larger α means faster decay. Δ is the seed-paired fingerprint-off minus shipped conditional contrast: each configuration selects its own correctly classified queries. It is not an effect on common queries. Mean pair influence is the principal descriptive endpoint here; the historical registered primary was a sampled maximum. The all-deletion maximum remains distinct on both scales. Accuracy pools paired changes over fourteen datasets (336 cells, 200 fixed queries per dataset); its SE and 95% t interval use dataset means. No equivalence test was performed; the vendor-default ensemble was not tested fingerprint-off.

Influence statistic	Scale	Shipped	Fingerprint off	Δ (SE)
Mean pair, $k = 32$	log-odds	0.285 (0.046)	0.830 (0.053)	+0.545 (0.056)
Mean pair, $k = 32$	probability	0.534 (0.053)	1.116 (0.091)	+0.583 (0.052)
Maximum pair, $k = n$	log-odds	0.034 (0.022)	0.216 (0.056)	+0.182 (0.048)
Maximum pair, $k = n$	probability	0.289 (0.045)	0.509 (0.058)	+0.220 (0.035)
Paired accuracy change, fourteen datasets: +0.13 (0.10) percentage points. Dataset-level 95% interval: $[-0.09, +0.34]$ percentage points.				

Over 64–512, `FINGERPRINT_FEATURE=False` changes $\alpha(Q_{\text{mean}})$ from 0.31 to 0.58 on higgs and 0.36 to 0.87 on magic. Sampled pair maximum exponents change from 0.38 to 0.53 and 0.30 to 0.73. Geometric means of paired cell level influence ratios are 0.62 on higgs and 0.21 on magic. The same query frozen stack arm D has magic mean pair margin exponent 0.803868, versus 0.284666 for arm 0, 0.829856 for arm A, and 0.289783 for fixed salt arm B. It uses the shipped base fit's queries and passes an unchanged support identity check, but recaptures preprocessing separately for each reordered full support base under position-clean deletion. Pair statistics can retain order effects. The post hoc seed paired D-minus-0 margin contrast is +0.519 (SE 0.051, ordinary t_5 interval $[0.388, 0.650]$) on magic and +0.145 (SE 0.030) on higgs. The probability scale contrasts are +0.515 (SE 0.056) and +0.138 (SE 0.033), respectively. The supplementary audit reports all twelve arm by dataset by scale contrasts with a separate exploratory Holm correction. These do not alter the historical family. This is corroboration for the whole refitted stack on shipped-correct queries, not isolation of the fingerprint feature. Arm B also reselects queries. Numerical agreement alone does not establish common identities.

The maximum and mean results differ at $k = n$. Fingerprint-off magic gives a sampled maximum exponent of 0.73 ± 0.04 over 64–512 and a probability scale exponent of 0.82 ± 0.02 on the extended window. Higgs gives 0.53 ± 0.14, or 0.53 ± 0.08 on the extension. At $k = n$, however, the magic maximum exponent is 0.22 ± 0.06, below $1/2$, contrary to the registered prediction. The mean pair and mean base to deletion exponents remain above, at 0.80 and 0.84 ($t \approx +5$). If pretraining used fingerprints, disabling them also changes the input distribution seen by the model. That history is undisclosed. Vendor ensemble comparisons appear in Appendix E.

Spread shows little fingerprint effect over 64–1024: fingerprint-off to shipped level ratios are 0.98–1.04 on both datasets, with no trend in n, and paired slope intervals contain zero. Registered extensions use the same 48 draws as the shipped runs. Higgs ratios remain 0.96–1.04 through 8192, with slope change -0.01 $[-0.03, +0.02]$. Magic ratios remain 0.98–1.05 through 2048, then rise to 1.17 and 1.15 at 4096 and 8192. The fingerprint therefore reduces spread at those large magic cells: the shipped extension slope is -0.51, versus -0.43 with the feature off, a paired difference of $+0.08$ $[+0.05, +0.11]$. The fingerprint-off fit is slower than the reference before overlap correction. The reason this effect appears only above 2048 on magic is unmeasured, and the disjoint 16,384 and 32,768 cells have no fingerprint-off counterparts.

Order sensitivity also depends on the interaction between hashing and floating-point preprocessing. A same size permutation with fixed salt changes 6.25% of higgs fingerprints. Removing either the fingerprint or SVD reduces its order floor from 2.2×10^{-2} to at most 2.8×10^{-6}. Fixing the salt does not. Order dependent numerical changes in magic's two-component SVD remain below the hash's 10^{-12} rounding threshold at the probe and variance cells through 256. Those in higgs's seven near-degenerate components do not. At larger magic cells the floor exceeds floating-point noise but remains at least 208 times below the spread signal. The registered prediction that shared salt would protect $Q_{\text{pair}}^{\max}$ fails: its movements are 5.2 and 6.4 SE. The registration records this amendment.

The v3 magic rerun has four hits and no misses. Disabling the fingerprint changes $\alpha(Q_{\text{pair}}^{\max})$ from 0.354 ± 0.065 to 0.743 ± 0.021 (paired difference $+0.389 \pm 0.049$, 8.0 SE), with the all-deletion qualification in Table 18. Mean exponents become 0.848 ± 0.045 for Q_{mean} and 0.813 ± 0.037 for Q_{pair}. Influence levels fall to 0.454, 0.303, 0.234, and 0.133 of shipped levels at 64, 128, 256, and 512 rows. Appendix H reports the twelve-dataset fingerprint-off comparison.

G.1 Controlled common query experiment

The experiment was designed using historical results and their query selection limitation. Its protocol, source hashes, and resource amendment precede the corresponding calls in the local record. These timestamps are not independently witnessed preregistration. All 192 planned CUDA cells completed, preserving 12,672 fit/predict calls, including 6,144 deletion calls. The protocol was amended from CPU to CUDA execution before any GPU call. Seven $n = 64$ CPU pilot cells run before the amendment are archived and not analyzed.

Inputs and matched execution. The v2.5 default checkpoint and all eight split array hashes match the historical inputs. Higgs uses the first 50,000 complete rows of the 98,050-row OpenML 23512 subset. Missing value removal precedes truncation. Magic uses OpenML 1120's ten features, excluding its ID field. Both use stratified 80/20 splits with random state 0, with the first 50 test rows fixed before evaluating any configuration. Archived features, labels, split indices, and source row ordinals identify these queries. The support/deletion RNG preserves the historical schedule, including its unused 240-candidate draw, for $n = 64, 128, 256, 512$, seeds 0–5, and $k = 32$.

TabPFN 8.2.0 uses one estimator, constructor random state 0, and float32. Four H100 80 GB GPUs run Python 3.12.3, torch 2.9.1+cu128, CUDA 12.8, and driver 580.126.09. The archived launcher keeps each seed's sizes and paired arms on one physical GPU. Per-call telemetry records logical `cuda:0`, without a physical UUID. Deterministic algorithms are enabled, TF32 and mixed precision disabled, and RNGs reset before every fit/predict. All 192 base repeat checks agree exactly. This CUDA grid does not mix in CPU inference or claim bitwise reproduction of historical MPS measurements.

Interventions and diagnostics. Arm0 is shipped preprocessing; A changes only `FINGERPRINT_FEATURE=False`. D captures the fitted CPU preprocessing pipeline and torch soft clip stage once on the original full support and reuses them for every reordered base and deletion call. It differs from historical D's per-reordered-base recapture. Original base predictions equal arm0 exactly. Retained fingerprints and query transforms remain exact for all D deletions. Duplicate row transform semantics could have invalidated this control. The specified identity checks passed on these inputs. Arm B pins the original full support cell count salt, rather than the historical 2^{20} salt, while refitting the remaining preprocessing. Constant salt is verified but does not imply retained feature stability. The $n = 256$, seed0 traces preserve transformed arrays for independent checking. No support

Table 21: Exploratory mean-pair influence exponents on the common-query six-seed, four-size grid. D freezes the original-support CPU preprocessing pipeline and torch soft-clip state once; B pins the original full-support salt while refitting other preprocessing. These definitions differ from the historical controls. Seed-paired intervals are ordinary, unadjusted t_5 intervals and are outside the primary four-test family. D preserves retained fingerprints; B does not.

Dataset	Scale	Control	α (SE)	Control minus shipped	95% CI
higgs	probability	D	0.649 (0.079)	+0.266	[+0.206,+0.326]
higgs	probability	B	0.442 (0.068)	+0.059	[-0.045,+0.164]
higgs	margin	D	0.635 (0.109)	+0.266	[+0.210,+0.322]
higgs	margin	B	0.429 (0.093)	+0.061	[-0.049,+0.171]
magic	probability	D	1.086 (0.055)	+0.512	[+0.372,+0.653]
magic	probability	B	0.533 (0.027)	-0.041	[-0.109,+0.027]
magic	margin	D	0.805 (0.036)	+0.533	[+0.392,+0.674]
magic	margin	B	0.245 (0.061)	-0.027	[-0.095,+0.040]

Table 22: Exploratory original-base prediction quality on the same 50 unfiltered queries. Average the four sizes within each seed before paired inference over six seeds. Accuracy is a fraction; lower log loss and Brier score are better. Intervals are ordinary unadjusted t_5 intervals and do not establish equivalence. D and B equal shipped on the original base by construction. All per-cell and per-seed scores are archived.

Dataset	Metric	Shipped	Off minus shipped	95% CI
higgs	accuracy	0.5933	-0.0067	[-0.0256,+0.0123]
higgs	log loss	0.6546	+0.0020	[-0.0051,+0.0091]
higgs	brier	0.2310	+0.0012	[-0.0016,+0.0040]
magic	accuracy	0.8550	+0.0025	[-0.0119,+0.0169]
magic	log loss	0.3486	+0.0051	[-0.0007,+0.0109]
magic	brier	0.1090	+0.0008	[-0.0006,+0.0022]

independent hash method is introduced: maintaining duplicate occurrence semantics would require a separate design.

Endpoints and inference. The primary cell statistic averages absolute differences over all 496 distinct pairs of the 32 deletion predictions, then over queries. Probability and float64 log-odds are separate scales. Only the log-odds calculation clips both class columns using the same $[10^{-12}, 1 - 10^{-12}]$ clipping rule, with no clipped values recorded. Negated OLS slopes use all four sizes within each seed, with no omitted cells or pseudocounts. All six seed exponents and all per-size levels are provided as CSVs. Table 2 gives all four primary paired A-minus-0 contrasts, their ordinary t_5 intervals, and the single four-test Holm correction. These intervals are not simultaneous and condition on the fixed query sets. The sampled $k = 32$ maxima remain exploratory, not exact all-deletion maxima.

D's broader intervention supports a refitting contribution without identifying a unique internal cause. B-minus-0 exponent intervals span zero on both scales and datasets (Table 21). The quality results (Table 22) establish neither equivalence nor improved prediction. D and B match shipped original base predictions by construction, so their zero base quality differences are not a performance test. Differences in devices, query populations, and control definitions prevent attributing a fraction of the historical contrasts to selection. All signed exploratory contrasts, quality scores, sampled maxima, diagnostics, attempts, and raw predictions remain in the supplement.

H TWELVE-DATASET EXTENSION

Scale dependence on pol. Pol is the extension's only probability scale above classification for the shipped TabPFN sampled pair maximum. The mean influence link diagnostic of Appendix C, computed from stored cells, gives margin and probability exponents of 0.238 ± 0.052 and 0.902 ± 0.145. Their gap is 0.665, versus a link term of 0.639, which accounts for 96% of the gap. The residual

is 0.026 ± 0.017, within 2 SE. This diagnostic concerns mean influence, not the pair maximum or spread, and the registered aggregate deletion claims remain on the margin scale.

The twelve additional datasets follow a registered selection rule applied to OpenML-CC18 and the tabular-benchmark numeric classification suite: binary targets, no previously used datasets or same source duplicates, at least 3000 rows after dropping missing values, at least 80% numeric features, and at most 100 features. The first twelve by ascending dataset ID qualify. All 71 exclusions are documented. Each dataset has a 20-permutation order probe at $n \in \{256, 1024\}$, followed by deletion measurements over 64–512, with six seeds, $k = 32$, both scales, CPU, and shipped preprocessing. Dataset level predictions and aggregate criteria were fixed before measurement. Electricity follows the earlier pilot registration. All twelve exceed the 10^{-4} order threshold, with maxima from 6.4×10^{-4} to 0.15. Wilt and covertype exceed it only at 1024. All therefore use scheme B. Using the matched size comparison of Appendix B, the probe floor is at most 13% of Q_{pair} on ten datasets and 54% on numerai28.6 and MiniBooNE. For the four below classified rows, it is at most 3%.

Table 23: V2.5 on twelve additional datasets with shipped preprocessing, 64–512 rows. Deletion columns: sampled pair maximum exponents ($k = 32$), mean (SE) over six seeds. t and classification use unrounded margin fits and the 3-SE threshold. Table 24 contains all spread slopes, intervals, overlap corrections and levels, avoiding repeated columns. V2.5 was fine-tuned on eye_movements' source upload (OpenML 1044, from which 44130 derives).

	deletion exponent			
Dataset	$\alpha(Q_{\text{pair}}^{\max})$	prob. scale	t	class
electricity	0.300 (0.061)	0.419 (0.047)	-3.3	below
jm1	0.530 (0.052)	0.663 (0.055)	$+0.6$	point
phoneme	0.185 (0.062)	0.379 (0.043)	-5.1	below
numerai28.6	0.530 (0.132)	0.499 (0.128)	$+0.2$	point
churn	0.247 (0.102)	0.426 (0.045)	-2.5	point
wilt	0.216 (0.035)	0.645 (0.085)	-8.2	below
credit/GMSC	0.439 (0.111)	0.424 (0.109)	-0.6	point
covertype (bin.)	0.501 (0.036)	0.567 (0.025)	$+0.0$	point
pol	0.367 (0.055)	1.005 (0.119)	-2.4	point
house_16H	0.374 (0.032)	0.529 (0.062)	-3.9	below
MiniBooNE	0.379 (0.058)	0.655 (0.058)	-2.1	point
eye_movements	0.406 (0.111)	0.392 (0.108)	-0.8	point

Nine of twelve margin scale point estimates fall below $1/2$, four classify below at 3 SE, and none classify above, meeting the three aggregate predictions. Probability exponents are larger on nine datasets. The other three differences are smaller than their SEs. Pol classifies above on the probability scale (1.005 ± 0.119, $t = +4.2$). These are uncorrected comparisons. Appendix E gives the multiplicity analysis. Excluding eye_movements, whose source upload was used to fine-tune the checkpoint, leaves eight of eleven point below, four below at 3 SE, and none above.

Tables 24 and 25. The two checkpoints' dataset rankings by spread slope have Spearman $\rho = 0.52$ (average ranks, point estimate). At 512, TabICLv2's levels on pol (0.025) and wilt (0.014), and v2.5's on wilt (0.012), fall below the registered $[0.03, 0.20]$ band. At 64, levels range from 0.036 (wilt) to 0.166 (house_16h) for v2.5 and 0.057 (wilt) to 0.212 (eye_movements) for TabICLv2. Pol and wilt also have the highest shipped accuracy at 512, 0.976 and 0.967 (Table 26). Probability saturation on nearly separable tasks remains a possible explanation for their fast probability spread decay. V2.5's margin spread on the same draws is nearly flat on pol, -0.037 $[-0.09, +0.02]$, and decays slowly on wilt, -0.197 $[-0.29, -0.09]$, with neither affected by probability clipping (registered post hoc diagnostic AP3(b)). TabICLv2's margin spread was not measured.

Several classical comparison predictions fail. Across fourteen datasets, checkpoint slope rankings correlate with $\sqrt{n}$-NN at $+0.35$ and $+0.34$, below the registered $+0.5$ threshold (RS1a, two misses). The prediction that local refits reach the rate on pol or wilt also fails, although XGBoost reaches it on pol (-0.64), the only resolved classical crossing there. Default penalty logistic regression reaches it at the point on two of twelve datasets, rather than the predicted nine: churn is resolved and pol lies at the interval edge. XGBoost is slower on eleven of twelve, as predicted. The supplementary registration ledger lists SB1–SB6 and the remaining outcomes.

Table 24: Probability spread for TabICLv2 (SB) and v2.5 (SA) on identical support draws, 64–512, $R = 24$. Columns give signed slopes with support resampling 95% intervals, $\kappa = 1$ corrections where 512 rows exceed 5% of the pool, and query mean SD at 512. Bold marks intervals wholly below $-1/2$. Table 23 contains deletion statistics only.

		TabICLv2			TabPFN v2.5 (reference)		
Dataset	d	slope [support CI]	$\kappa = 1$	level 512	slope [support CI]	$\kappa = 1$	level 512
pol	26	**−0.86** [−0.94, −0.78]	−0.85	0.025	**−0.71** [−0.83, −0.60]	−0.70	0.032
wilt	5	**−0.66** [−0.73, −0.58]	−0.63	0.014	−0.51 [−0.63, −0.40]	−0.49	0.012
jm1	21	−0.44 [−0.50, −0.37]	−0.43	0.062	−0.36 [−0.44, −0.28]	−0.35	0.060
credit_gmsc	10	−0.42 [−0.47, −0.36]	–	0.067	−0.32 [−0.38, −0.26]	–	0.064
house_16h	16	−0.41 [−0.46, −0.36]	–	0.074	−0.32 [−0.37, −0.28]	–	0.084
electricity	8	−0.36 [−0.41, −0.31]	–	0.091	−0.22 [−0.28, −0.17]	–	0.085
churn	20	−0.36 [−0.42, −0.28]	−0.33	0.064	−0.29 [−0.35, −0.22]	−0.26	0.057
eye_movements	20	−0.36 [−0.41, −0.31]	−0.34	0.099	−0.22 [−0.31, −0.13]	−0.20	0.069
covertype_b	10	−0.35 [−0.41, −0.28]	–	0.097	−0.25 [−0.32, −0.19]	–	0.086
miniboone	50	−0.33 [−0.39, −0.26]	–	0.071	−0.37 [−0.45, −0.29]	–	0.071
numerai	21	−0.28 [−0.37, −0.19]	–	0.090	−0.37 [−0.49, −0.26]	–	0.047
phoneme	5	−0.20 [−0.25, −0.16]	−0.17	0.103	−0.20 [−0.24, −0.15]	−0.17	0.099

Table 25: Classical refits on the same twelve-dataset supports, 64–512, $R = 24$ (RS). Entries give spread slopes with support resampling 95% intervals. Parentheses give $\kappa = 1$ corrections where 512 rows exceed 5% of the pool.

Dataset (d)	logistic regression	$\sqrt{n}$-NN	XGBoost
electricity (8)	−0.46 [−0.51, −0.40]	−0.26 [−0.30, −0.22]	−0.23 [−0.27, −0.18]
jm1 (21)	−0.40 [−0.46, −0.35] (−0.39)	−0.30 [−0.34, −0.25] (−0.28)	−0.21 [−0.27, −0.15] (−0.20)
phoneme (5)	−0.47 [−0.56, −0.39] (−0.45)	−0.33 [−0.37, −0.29] (−0.31)	−0.22 [−0.27, −0.18] (−0.20)
numerai28.6 (21)	−0.40 [−0.45, −0.36]	−0.26 [−0.30, −0.22]	−0.19 [−0.22, −0.16]
churn (20)	−0.60 [−0.66, −0.52] (−0.57)	−0.33 [−0.38, −0.27] (−0.30)	−0.35 [−0.41, −0.27] (−0.33)
wilt (5)	−0.33 [−0.41, −0.21] (−0.30)	−0.34 [−0.41, −0.24] (−0.31)	−0.23 [−0.34, −0.09] (−0.20)
credit_gmsc (10)	−0.23 [−0.30, −0.16]	−0.14 [−0.19, −0.08]	−0.25 [−0.27, −0.21]
covertype (10)	−0.47 [−0.53, −0.42]	−0.25 [−0.29, −0.22]	−0.24 [−0.27, −0.22]
pol (26)	−0.52 [−0.57, −0.47] (−0.50)	−0.41 [−0.45, −0.37] (−0.39)	−0.64 [−0.73, −0.54] (−0.62)
house_16h (16)	−0.45 [−0.51, −0.39]	−0.32 [−0.36, −0.27]	−0.34 [−0.39, −0.30]
miniboone (50)	−0.36 [−0.49, −0.27]	−0.29 [−0.34, −0.24]	−0.38 [−0.43, −0.33]
eye_movements (20)	−0.46 [−0.51, −0.41] (−0.44)	−0.28 [−0.32, −0.25] (−0.27)	−0.23 [−0.26, −0.20] (−0.21)

Fingerprint-off comparison. The registered 2026-08-11 rerun changes only `FINGERPRINT_FEATURE=False`, retaining each dataset's protocol and scheme B. Its scorer was committed before measurement. Repeats are exact and the intervention changes predictions on all twelve datasets (Table 27). On the margin scale sampled pair maximum, one point estimate is below $1/2$ (pol), none classify below, and seven classify above. The prediction of at most four below classifications passes, but the prediction of at least six point below estimates fails. There are no $k = n$ reruns on these datasets. The higgs and magic all-deletion results in Table 18 caution against interpreting sampled maxima as exact maxima.

For Q_{mean}, nine datasets classify above (all except churn, pol, and MiniBooNE), contrary to the prediction of none. Only pol has a point estimate below. On the probability scale, all twelve pair maximum exponents and eleven mean exponents classify above. Churn's mean is 2.9 SE above. Every seed paired conditional margin scale change is positive: pair maximum changes range from $+0.05$ (pol) to $+0.58$ (numerai28.6), median $+0.33$, and mean changes from $+0.20$ to $+0.76$, median $+0.48$. These results compare configuration specific correctly classified populations within v2.5. They do not isolate an effect on common queries. Shipped preprocessing remains the primary twelve-dataset analysis.

Paired accuracy. The registered accuracy comparison uses the same supports and first 200 unfiltered test rows, with six seeds, four context sizes, and CPU inference. Across the twelve datasets, fingerprint-off minus shipped accuracy is $+0.0016 \pm 0.0011$, with a dataset level 95% t interval of $[-0.0008, +0.0040]$. Dataset changes range from -0.0044 (covertype) to $+0.0102$ (numerai28.6, at

chance in both configurations). None exceeds 1.8 SE. Changes by context size range from -0.003 at 64 to $+0.004$ at 256, with $+0.003 \pm 0.001$ at 512. Table 26 gives the dataset level results.

The same protocol on higgs and magic gives -0.25 ± 0.33 and $+0.10 \pm 0.36$ percentage points, with both intervals containing zero. Across all fourteen datasets (336 cells), the mean change is $+0.13 \pm 0.10$ percentage points. The datasets include exact feature duplicates, the stated reason for the fingerprint. After dropping missing values, 25.1% of jm1's rows are duplicates and 8.2% belong to groups with conflicting labels. The other eleven datasets have at most 0.83% (wilt), and five have none. Jm1's paired accuracy change is $+0.0021 \pm 0.0020$.

Mitra's Linux CPU comparison tests three inference configurations. Its deterministic harness reproduces the macOS cells exactly. The seeded vendor stack without fine-tuning differs from the harness by $+0.35 \pm 0.30$ percentage points on higgs and -0.33 ± 0.19 on magic across 30 cells, with no per-size mean beyond 3 SE. Default fine-tuning (50 steps) gives $+0.54 \pm 0.98$ over 24 complete higgs cells (64–512) and -0.23 ± 0.50 over 30 magic cells. The higgs run reached only two of six seeds at 1024 before the 7200 s cap, so that cell was excluded under the registered rule. The supplementary ledger records the eight prediction outcomes.

Table 26: V2.5 paired accuracy change, fingerprint off minus shipped, on fourteen datasets. Each difference is a mean (SE) over 24 cells: four sizes (64, 128, 256, 512), six seeds, and 200 fixed queries. The last columns give both configurations' accuracy at 512.

Dataset	Δacc	SE	acc_{512} shipped	acc_{512} defp.
electricity	+0.0035	0.0025	0.780	0.786
jm1	+0.0021	0.0020	0.802	0.803
phoneme	+0.0035	0.0038	0.833	0.832
churn	−0.0006	0.0025	0.895	0.904
wilt	−0.0002	0.0006	0.967	0.967
credit/GMSC	+0.0040	0.0032	0.790	0.794
covertype	−0.0044	0.0042	0.774	0.779
pol	+0.0025	0.0020	0.976	0.980
eye_movements	+0.0023	0.0044	0.582	0.590
house_16H	−0.0010	0.0036	0.868	0.863
MiniBooNE	−0.0029	0.0025	0.912	0.912
numerai28.6	+0.0102	0.0058	0.499	0.506
higgs (production)	−0.0025	0.0033	0.667	0.674
magic (production)	+0.0010	0.0036	0.876	0.873
pooled (twelve additional datasets)	+0.0016	0.0011		
pooled (over all fourteen)	+0.0013	0.0010		

Table 27: V2.5 with fingerprints off, otherwise matching Table 23: 64–512, $k = 32$, six seeds, scheme B, both scales. Entries give exponents (SE) and seed paired conditional changes Δ = fingerprint off minus shipped (SE). The two configurations select correctly classified queries separately. Bold marks exponents above $1/2$ by at least 3 SE. None classifies below. Maximum columns remain sampled maxima. There are no $k = n$ reruns on these datasets.

	$\alpha(Q^{\max}_{\text{pair}})$, margin		$\alpha(Q_{\text{mean}})$, margin		$\alpha(Q^{\max}_{\text{pair}})$, prob.		$\alpha(Q_{\text{mean}})$, prob.	
Dataset	off	Δ	off	Δ	off	Δ	off	Δ
electricity	**0.69** (0.06)	+0.39 (0.05)	**0.71** (0.05)	+0.50 (0.05)	**0.79** (0.06)	+0.37 (0.04)	**0.81** (0.06)	+0.49 (0.04)
jm1	**0.73** (0.05)	+0.20 (0.08)	**0.80** (0.06)	+0.30 (0.14)	**0.85** (0.06)	+0.19 (0.08)	**0.92** (0.08)	+0.31 (0.14)
phoneme	**0.75** (0.06)	+0.56 (0.07)	**0.86** (0.05)	+0.76 (0.05)	**0.90** (0.07)	+0.52 (0.09)	**1.04** (0.06)	+0.72 (0.06)
numerai28.6	**1.11** (0.13)	+0.58 (0.08)	**1.01** (0.14)	+0.47 (0.07)	**1.08** (0.13)	+0.58 (0.07)	**0.99** (0.13)	+0.47 (0.06)
churn	0.56 (0.14)	+0.31 (0.06)	0.51 (0.12)	+0.51 (0.06)	**0.78** (0.07)	+0.35 (0.05)	0.75 (0.09)	+0.54 (0.06)
wilt	0.71 (0.14)	+0.49 (0.14)	**0.79** (0.07)	+0.60 (0.10)	**1.12** (0.19)	+0.48 (0.22)	**1.15** (0.08)	+0.64 (0.12)
credit/GMSC	**0.90** (0.06)	+0.46 (0.09)	**0.94** (0.06)	+0.56 (0.05)	**0.89** (0.06)	+0.47 (0.09)	**0.95** (0.05)	+0.57 (0.05)
covertype (bin.)	**0.81** (0.09)	+0.31 (0.08)	**0.87** (0.10)	+0.46 (0.06)	**0.90** (0.07)	+0.34 (0.08)	**0.97** (0.08)	+0.46 (0.07)
pol	0.42 (0.11)	+0.05 (0.07)	0.44 (0.12)	+0.20 (0.07)	**1.11** (0.13)	+0.11 (0.09)	**1.09** (0.15)	+0.18 (0.08)
house_16H	0.66 (0.06)	+0.29 (0.06)	**0.84** (0.07)	+0.57 (0.08)	**0.83** (0.08)	+0.31 (0.06)	**1.00** (0.08)	+0.59 (0.09)
MiniBooNE	0.59 (0.09)	+0.21 (0.08)	0.65 (0.07)	+0.31 (0.10)	**0.96** (0.12)	+0.30 (0.08)	**1.01** (0.09)	+0.37 (0.09)
eye_movements	**0.76** (0.07)	+0.35 (0.10)	**0.75** (0.07)	+0.40 (0.13)	**0.74** (0.07)	+0.35 (0.10)	**0.74** (0.07)	+0.40 (0.13)

I MECHANISM TESTS AND ATTACK COST

The following tests did not support the proposed explanations for the influence exponents. The first three had predictions and rejection criteria recorded before measurement. The fourth is a pilot with file dated registration whose initial timing cannot be independently verified (Appendix K). These results reject the stated predictions within their measured settings, rather than all possible versions of the mechanisms.

Readout family. Biloš et al. (2026) describe Mitra as an attention weighted vote at a late layer: L9 in their v1 model, corresponding to L12 in the v1.1 checkpoint tested here. Our checkpoint lacks all three tested vote signatures. Query to support attention is nearly uniform throughout (L12 mean maximum weight 0.0105, versus uniform 0.0039). Forcing it uniform at L12 or L6 has no accuracy cost, whereas the same intervention at L1 reduces accuracy to chance. An L12 label vote surrogate agrees on only 64% of queries the model classifies perfectly. The proposed readout description therefore does not transfer to this checkpoint.

Layer criticality. We tested whether dependence on a single attention layer predicts slow influence decay. The intervention forces query to support attention uniform at one block at a time. Criticality is the largest accuracy drop divided by the model's accuracy above chance. Mitra falls to chance when L1 is altered (criticality 1.02), while TabICLv2 and TabFM have criticalities 0.21 and 0.20. No TabFM layer exceeds the noise threshold. One TabICLv2 layer has a marginally significant three-percentage-point effect. The originally registered transfer depth centroid is undefined for TabFM. Criticality replaced it after the two calibration profiles were seen and was tested prospectively only on TabFM. The resulting registered prediction, $\alpha(Q_{\text{mean}}) = 1.06 \pm 0.25$, misses TabFM's 0.745 ± 0.038 on the full window and 0.786 ± 0.049 for $n \geq 256$. The secondary n_{eff} slope prediction passes ($+0.962$ in $[0.90, 1.00]$), but that statistic does not separate the trained models.

Trunk depth. The prediction that deeper row mixing trunks produce slower decay fails on shuffled TabFM trunks with $L \in \{3, 6, 12, 18, 24\}$ blocks. The sequence is non-monotone: the shallowest has exponent 0.687, versus a plateau near 0.82 from $L = 12$. At 12 blocks, TabFM gives 0.828 over $n \leq 2048$, versus 0.978 and 0.998 for the different 12-block TabICLv2 and Mitra architectures over $n \leq 512$. The untruncated control reproduces the trunk shuffled exponent 0.825 exactly, passing the registered ± 0.08 check (Table 28).

Prior mismatch (pilot). Müller et al. (2024) report a nearest-neighbour fallback when data are unlikely under a PFN's prior. We tested the prediction that influence would decay faster on data from the checkpoint's prior family. Three datasets were drawn from TabPFN v1's public `prior_bag`, the closest public ancestor of v2.5's undisclosed prior. TabICLv2 on its own `mix_scm` prior served as a positive control, and refit logistic regression as a pipeline check. The run record contains the generator configuration, commit, SHA-256, predictions, amendments, and scores. TabPFN uses scheme B.

TabICLv2 gives exponents 1.04 ± 0.06, 1.16 ± 0.08, and 0.92 ± 0.06, with the middle point 0.012 above the registered $[0.80, 1.15]$ band. TabPFN v2.5 gives $\alpha(Q_{\max}) = 0.51 \pm 0.16$, 0.18 ± 0.04, and 0.30 ± 0.04. The order sensitivity amendment makes $Q_{\max}$ primary. Pair maximum exponents are 0.47, 0.21, and 0.32. The mean 0.33 fails the registered 0.5 threshold, no draw exceeds the real data exponent by two pooled SE, and one falls below it. On two datasets from the TabICL prior, the same v2.5 checkpoint gives 0.44 ± 0.09 and 0.61 ± 0.28. The second stops at $n \leq 256$ with 15 of 24 cells complete. The first exceeds $1/2$ on the mean statistics, but the two prior families do not separate at this precision. The result concerns an ancestor prior, six-seed pilot measurements at $n \leq 512$, and a pipeline with one failed validation among six draws (0.79 against a required $\alpha \geq 0.9$). It does not test v2.5 on its actual pretraining prior.

Attack cost. An additive approximation suggests that influence decaying as $n^{-\alpha}$ would require $b^*(n) \sim n^{\alpha}$ label flips for a fixed margin shift. This prediction, assigned probability 50% per model, fails on every evaluable model. Rows are ranked by single flip influence, then the top b labels are flipped. For a unit absolute margin shift, budget exponents are 1.6 ± 0.5 for v2.5, 1.6 ± 0.3 for TabICLv2, and 1.3 ± 0.4 for TabFM. Matched window mean deletion exponents are 0.31, 1.06, and 0.69, refit from stored artifacts over 64–512. The budget grows faster than the additive prediction:

$|\beta - \alpha| = 0.58$–1.34, versus the registered tolerance 0.20. Budget exponents are 3.4–4.9 SE above zero.

The test uses higgs, three seeds, $n \leq 512$, and log-linear interpolation on the registered budget grid. Cells are right-censored when flipping all labels fails to reach the threshold. V2.5 has two evaluable seeds, with only two uncensored cells in seed 0. Mitra's exponent is unevaluable: four of nine cells are censored, and a same day registration revision drops its 512 row. Ten cells are censored overall, exceeding the predicted maximum of one. The attack itself is standard influence ranked label flipping (Koh & Liang, 2017; Fang et al., 2020; Nguyen & Wong, 2023; Zhang et al., 2026; Biloš et al., 2026). The negative result is that the measured deletion exponent does not predict its cost over this range.

J WEIGHT RANDOMIZED CONTROLS

Table 28: Signed Q_{mean} slopes for trained models and randomized copies by scope, mean (SE). TabFM fits use 64–2048. Others use 64–512. The final column gives trained to randomized label flip response ratios for one scope per model: TabICLv2 and Mitra trunk, TabFM and v2.5 full. p denotes the Linux environment.

Model	data	trained	trunk scope	full scope	control reads context?
TabICLv2	higgs	-1.057 (0.064)	-0.978 (0.045)	-0.993^{p} (0.017)	4–14× below
TabICLv2	magic	-1.017 (0.043)	-1.033 (0.051)	-1.041^{p} (0.012)	(same control)
TabFM	higgs	-0.745 (0.038)	-0.825 (0.042)	-1.064 (0.013)	14–34× below (full scope)
TabFM	magic	-0.760 (0.075)	-0.849 (0.059)	-1.036 (0.019)	(same, trunk not measured)
Mitra	higgs	-0.344 (0.141)	-0.998 (0.009)	scope ≈ full	21–37× below
Mitra	magic	-0.749 (0.049)	-1.030 (0.014)	scope ≈ full	(same control)
TabPFN v2.5	higgs	-0.305 (0.127)	not run	-0.219 (0.057)	74–166× below

Table 29: TabICLv2 signed influence slopes, mean (SE), on six datasets. Higgs, magic, and spambase use $k = 32$, six seeds, and 64–512; adult and bank use $k = 128$, three seeds, and 64–2048; credit-g uses the latter protocol through 512. Fits include sizes below 256 under the registered exception and cap k at n for $n = 64$. Spambase fails the pipeline band for trained and shuffled models and is excluded from the main comparisons.

Dataset	arm	slope Q_{mean}	slope $Q_{\max}$	slope n_{eff}
higgs	trained	-1.057 (0.064)	-0.931 (0.044)	$+0.948$ (0.012)
higgs	shuffled	-0.978 (0.045)	-0.959 (0.052)	$+0.989$ (0.007)
magic	trained	-1.017 (0.043)	-0.770 (0.042)	$+0.852$ (0.031)
magic	shuffled	-1.033 (0.051)	-0.950 (0.101)	$+0.936$ (0.031)
spambase†	trained	-0.803 (0.051)	-0.681 (0.084)	$+0.940$ (0.029)
spambase†	shuffled	-0.840 (0.054)	-0.808 (0.081)	$+0.967$ (0.039)
adult	trained	-0.963 (0.046)	-0.702 (0.073)	$+0.851$ (0.034)
adult	shuffled	-0.903 (0.025)	-0.770 (0.060)	$+0.953$ (0.034)
bank	trained	-0.846 (0.021)	-0.624 (0.019)	$+0.857$ (0.014)
credit-g	trained	-1.019 (0.030)	-0.704 (0.076)	$+0.861$ (0.037)

TabICLv2 across six datasets. Table 29 reports trained and trunk shuffled fits under the 2026-07-31 registration. Four of six trained Q_{mean} slopes lie within 0.06 of -1 on their respective windows. Refitting adult and bank over $n \leq 512$ changes -0.963 and -0.846 to -1.084 and -0.991, leaving five of six within 0.09 of -1. Bank misses the registered full window pipeline band by 0.10 but passes on the matched window. Spambase (-0.803) misses for both trained and shuffled models. At $k = 128$, the sampled maximum decays more slowly than the mean (adult: -0.702 versus -0.963). All three $k = 128$ datasets miss the registered $\alpha(Q_{\max}) \in [0.8, 1.1]$ prediction, giving 0.62–0.70. Restricting to $n \geq 128$ to hold k constant gives 0.68–0.76, so the $k = n$ cell at 64 does not explain the failure.

Trained minus shuffled n_{eff} slopes differ by 0.041 on higgs (2.93 pooled SE) and 0.102 on adult (2.1 SE). Mitra's slopes are $+0.954$ (0.010) and $+0.995$ (0.005), separated by 3.6 pooled SE from

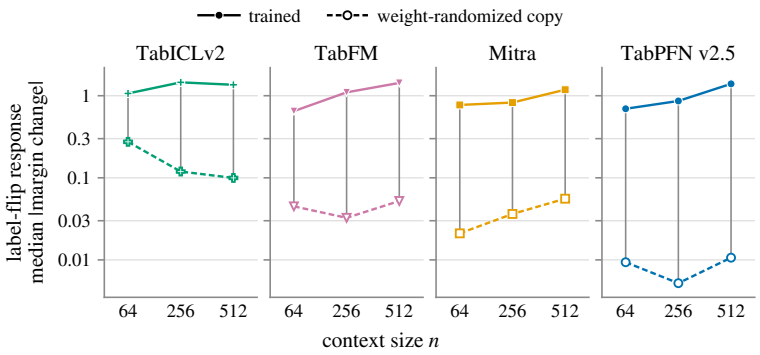


Figure 5: Median absolute margin change after flipping all support labels, trained and randomized copies (log scale). The four displayed controls respond 3.9–166 times less than trained. The separately measured full scope TabICLv2 copy responds 136–252 times less (Section 4.4).

unrounded fits. Across trained models, n_{eff} does not distinguish the influence decay patterns. No conclusion relies on it.

Higgs comparisons. TabICLv2 and Mitra use the matched deterministic fp32 protocol, $k = 32$ and six seeds, with exact repeats and order variation 4.7×10^{-6}. Their trained mean influence exponents differ by 0.71 (4.6 pooled SE). Their maximum influence exponents differ by 0.67 (4.3 SE). Signed $Q_{\max}$ slopes are -0.931 and -0.959 for trained and shuffled TabICLv2, and -0.266 and -0.997 for Mitra. The registered -0.5 threshold lies 1.1 SE below trained Mitra's point estimate. Its two-SE interval edge, -0.63, is 2.6 pooled SE above its control. Mitra has the largest per-seed SE, 0.141. The shuffled copies have exponents near 1, but their weak label response prevents a causal interpretation of this contrast.

The trained n_{eff} slopes are similar ($+0.948$ and $+0.954$). Mitra's estimated total influence, nQ_{mean}, increases 3.3 times between 64 and 2048 rows. Endpoint ratio exponents are 0.34 over the full range, 0.52 through 512, and 0.17 over the registered window. OLS fits to the same mean levels give 0.38, 0.60, and 0.14. These replace the earlier 0.66 obtained indirectly from a fitted influence exponent (Appendix K).

Influence levels. Paired trained to shuffled Q_{mean} ratios for Mitra on higgs are 18.7, 20.0, 49.0, and 62.8 at 64, 128, 256, and 512 rows, with per-size log-ratio SEs of 0.06–0.22. Magic ratios are 33.1, 42.9, 47.4, and 61.3. TabICLv2's ratio instead declines from 2.70 to 2.31. Appendix E extends Mitra's higgs series to 2048. At 512 on higgs, trained Mitra's mean sampled row margin change is 0.022 and its largest sampled change is 0.095. For scale, the deterministic harness gives a median margin of 0.394 at 256. Because these values use different context sizes, their ratio is not a measured relative influence at 512. The stochastic stack median 0.63 in Appendix G is a separate configuration.

Replication datasets. On magic, signed mean influence slopes are -1.017 (0.043) and -1.033 (0.051) for trained and shuffled TabICLv2, and -0.749 (0.049) and -1.030 (0.014) for Mitra. Mitra's slope difference is 0.282 (5.5 SE), versus 0.654 on higgs. The difference between datasets is 2.5 pooled SE, below the comparison threshold. Its predicted slope above -0.5 fails on magic and spambase (-0.749 and -0.722). On the datasets passing the pipeline band, trained Mitra decays more slowly than its approximately exponent-1 randomized copy. Spambase fails the band: both controls also decay slowly (TabICLv2 -0.840, Mitra -0.883), so the deviation cannot be attributed specifically to training. It remains excluded from the main comparisons.

TabFM randomization scope. Trunk randomization shuffles `icl_predictor` (1.62B parameters), leaving 19.5M parameters in two column embedders and two row interactors trained. The 13.2M-parameter embedder pair contains 256 learned vectors that pool support rows before the trunk, outside the layer interventions of Appendix I. The trunk-plus-columns copy shuffles those stages too. The full copy shuffles all 909 parameter tensors. The trunk copy misses the predicted

signed slope band $[-1.1, -0.9]$, giving -0.825 ± 0.042, within 1.4 SE of trained. The full copy gives $\alpha(Q_{\text{mean}}) = 1.064 \pm 0.013$ over the full window and 1.025 ± 0.035 over $n \geq 256$, separated from the trunk copy by 5.4 pooled SE. The trunk-plus-columns slope is -1.075 ± 0.029.

The higgs sequence (trained -0.745, trunk -0.825, trunk-plus-columns -1.075, full -1.064) is similar on magic (-0.760, -0.849, -1.145 ± 0.029, -1.036). The magic trunk-plus-columns copy misses its prediction of lying within 0.1 of the full copy and inside $[-1.1, -0.9]$. It is 3.2 pooled SE steeper (K6b arm 3). These comparisons show where randomization changes the exponent but do not localize a learned mechanism. All scopes classify at chance without probability or margin saturation. The full copy's label flip response is 14–34 times below trained. Label flip responses of the trunk and intermediate copies were not measured, so no response trend across scopes is established. These measurements do not identify the component responsible for the trained model's decay. The exact run time source for the higgs full copy is unrecoverable (Appendix K). Later TabPFN-magic, v3, and K13 extension runs have no newly constructed randomized counterparts. The existing v2.5 copy is full scope.

Scope and label flip checks. TabICLv2 has trunk and full copies. TabFM has trunk, trunk-plus-columns, and full copies. Mitra and v2.5 have one each. Only Mitra and TabICLv2 copies were tested for spread. On magic their spread slopes are -0.52 and -0.46, both faster than trained and with intervals containing $-1/2$. Tables 7 and 8 report the overlapping and disjoint results, including the inconclusive long window TabICLv2 fits.

Over Table 28's windows, trained control influence slope differences are 4.6 and 5.5 SE for Mitra and 8 and 3.6 SE for full scope TabFM on higgs and magic. TabFM's trunk differences are 1.4 and 0.9 SE. TabICLv2 and v2.5 lie within 1.0 SE of their controls. On matched $n \leq 512$ windows, full scope TabFM differs by 5.0 and 2.0 SE. Its trunk copy is within 0.3 SE on both datasets, with the higgs ordering reversed. One longer window exception is TabICLv2 on magic: its trunk copy decays faster over 256–2048 by 0.11 ± 0.03 (Appendix E).

Figure 5 reports median absolute margin changes after flipping all labels, on 200 higgs queries at $n \in \{64, 256, 512\}$. Table 28's response ratios use unrounded values. No measured randomized response is within a factor of two of trained. Full scope TabICLv2 gives 0.004, 0.006, and 0.010, versus trained responses of 1.066, 1.455, and 1.359 measured in the same run: ratios of 252, 244, and 136. Across the five tested copies, accuracy is 0.485–0.51 throughout. Full scope TabICLv2 has accuracy 0.49 at every size, versus 0.50, 0.65, and 0.70 for trained. Their small label responses and chance accuracy make these controls unsuitable for separating architecture from learning.

Full scope TabICLv2's signed deletion slopes, measured in the Linux environment, are -0.99 (0.02) on higgs and -1.04 (0.01) on magic. Their negated exponents exceed $1/2$ by 30 and 45 SE and lie within 1 SE of trained, matching the trunk copy pattern. The randomized v2.5 deletion measurement predates the order sensitivity diagnosis and does not use scheme B, unlike its trained counterpart. Finally, trained median absolute margins change with context size by up to 2.4 times: TabFM gives 0.31, 0.54, 0.73 and Mitra 0.43, 0.42, 0.62 at 64, 256, 512. These margins are distinct from the label flip responses and quantify the scale drift noted in the limitations.

K Reproducibility and Supporting Records

The run reference bundle contains 422 directories with configurations, seeds, cell metrics, and available source and result hashes. The code archive provides `runref_index.csv`, a claim to run index, and `docs/registration_outcomes.csv`, the complete registration ledger. The accompanying detailed reproducibility record retains the registration history, amendments, exclusions, and numerical checks.

Registration and exclusions. The 64 registrations with scored predictions contain 443 units: 323 hits, 104 misses, 15 unevaluable units, and one evaluated on a restricted model subset (supplementary registration ledger). The historical record also includes one descriptive registration. The exact tree addition is a separate descriptive protocol. Results for 23 registrations are recorded in separate results notes. Four predictions were scored retrospectively from complete artifacts. Five proposed units were never scored and support no claim. The ledger identifies these units and lists all missed and unevaluable predictions.

Some registrations are supported by git commits preceding measurement. Others are dated files whose initial timing cannot be independently verified after amendments. Neither file timestamps nor recorded hashes provide tamper-evident registration. The magic influence predictions in Table 3 have no independently established prior timing. For the v3 deletion scheme rerun, the registration precedes the deciding probe, but the rule used to choose the rerun is first visible in git after that probe. Its earlier timing rests on the note itself. Fit window differences are stated in Appendix E.

Computing environments. Historical unmarked measurements use ARM64 macOS with CPU or Metal Performance Shaders (MPS), as specified for each model in Table 4. Measurements marked p use Linux instances with NVIDIA A100 80 GB accelerators, except Mitra measurements explicitly run on CPU. The historical CUDA runs use driver 550.127.05, fp32, TF32 disabled, and deterministic algorithms unless stated otherwise. Package versions are pinned in `environments/` and `environments_platform/`. Table 4 identifies checkpoints and configurations.

Repeated numerical checks distinguish two Linux environments, P1 and P2, despite matching driver versions. Their reference cell values differ by 3.2×10^{-3} relative ($z = 0.073$), with order floors differing by 8.3×10^{-2} relative. Repeated calls within each environment agree exactly. Here z is the cell difference divided by the root sum of squared cell SEs, computed over queries. Each fitted series uses one environment. Cross-environment comparisons are limited to the checkpoint versus refit spread differences and specified vendor default comparisons. The reference checks imply slope changes near 0.001, against interval widths of 0.05–0.08. The detailed record identifies the environment for each measurement.

Some numerical checks exceeded the registered $z \leq 0.01$ threshold. For the initial CUDA comparison, the two spread cells gave $z = 0.019$ and 0.005, and their order floors gave $z = 2.2$ and 0.19. Measurements were retained under the registered exception requiring disclosure and an explicit decision, because these numerical differences were much smaller than the reported intervals. TabICLv2's cross-environment slope changes are below 0.0004. TabFM reproduces within 2×10^{-5} and Mitra's deterministic harness within 3×10^{-5} across CPUs.

TabPFN is more sensitive to the environment when fingerprints are enabled. Identical raw arrays produced different fingerprints for 21 of 256 rows (8.2%) in one stored support. Shipped cell values change by up to 0.40 relative. Among the two registered influence statistics, the largest exponent change is 0.29 SE. One full window classification changes: v3-magic's sampled pair maximum shifts from 0.354 (0.065), $t = -2.2$, to 0.348 (0.048), $t = -3.2$. The manuscript retains the original estimate. The post hoc $Q_{\max}$ also changes classification, from $t = -2.8$ to -3.9. Fingerprint-off v3-magic reproduces its two influence exponents at $z = 0.0006$ and 0.0005.

Prediction level evidence. The supplement includes probability matrices for the principal spread and large query accuracy analyses, 96 classical deletion NPZ files, six prior match data NPZ files, and 48 original TabICLv2 RMS deletion matrices. The RMS matrices reproduce the TabICLv2 cell statistics and fitted exponents. The principal August fingerprint ladder did not store deletion matrices; other historical foundation model matrices remain outside the submission ZIP. Their distributed cell summaries support refits, not prediction-level reconstruction. The separate common query CUDA grid adds 192 complete prediction cells and 12,672 per-call matrices with supports, queries, deletions, source snapshots, logs and hashes (Section G.1). These do not recover missing historical matrices. No missing prediction was inferred or fabricated. The new evidence audit and post hoc spread sensitivities have separate source files, input hashes, and outputs under `followup/`. They leave historical measurements unchanged.

Record completeness. Of the 419 historical run directories, 406 record a git revision with tracked, unmodified producing source, including the historical main result runs. A 420th directory adds the registered exact tree comparison, with hashed run time source snapshots, registration receipt, complete environment, raw probabilities, queries, and logs. Its preceding registration uses a local timestamp, not an independent timestamp authority. Two further directories complete the separately registered classical all-deletion grid, preserving the preceding 1,853 evidence files and storing all 96 deletion matrices with constructor settings, data hashes, source snapshots, warnings and logs. Ten early diagnostic directories lack recoverable run time source. Their results and logs survive, but exact source reconstruction is impossible. These include transfer depth, depth truncation, and full scope TabFM measurements. No Table 3 row depends on them, but the full scope TabFM

comparison in Section 4.4 does. The other three exceptions are an imported pilot, a diagnostic made with a disclosed script copy, and a descriptive addendum whose source entered git unchanged in the next commit. Fifteen diagnostic directories lack individual execution logs. Their reported summaries can be recomputed from stored metrics. The deciding v3 deletion scheme probe is stored separately from the run directories.

The shuffled TabICLv2 trunk and full copies and the Mitra copy reconstruct exactly from the supplied seed: all 391, 391, and 392 tensors match, respectively. Mitra's checkpoint file is byte-identical. The TabICLv2 files differ in serialization only. The bundle records the original SHA-256 values and explains which text fields were anonymized for review.

Checkpoint access. Checkpoints are obtained from their original distributors and are not included in the supplementary archives. The model cards specify tabpfn-2.5-license-v1.1 and tabpfn-3-license-v1.0 for v2.5 and v3, BSD-3-Clause for TabICLv2, Apache-2.0 for Mitra, and tabfm-non-commercial-v1.0 for TabFM. TabPFN v2.5 and v3 downloads require acceptance through a vendor account. The v1 checkpoint's licence is not resolved by the available repository and package records. The supplementary access instructions document that discrepancy.

L CONFIGURATION AND MEASUREMENT SCOPE REFERENCE

Section 2 defines the statistics and signs. Appendix E specifies historical windows and multiplicity, and Section G.1 specifies the controlled intervention.

Labels. *Shipped* retains default preprocessing, usually with one estimator/view and fp32. *Fingerprint off* ("defing.") changes `FINGERPRINT_FEATURE=False`. Vendor defaults use eight TabPFN/TabICLv2 estimators or 32 TabFM views. Mitra spread and influence disable fine-tuning. Its separate accuracy comparison labels the fine-tuning arm. Superscript p marks Linux measurements. P1/P2 numerical checks and z are defined in Appendix K. $P(\mathrm{attain})$ is a bootstrap fraction, not a null hypothesis p-value. Historical fingerprint influence ratios are geometric means of seed and size paired conditional statistics, not fractions of an effect explained. The variance overlap setting $\kappa = 1$ is a sample mean sensitivity reference, not an established bound.

Measured ranges. All datasets are binary classification. TabPFN v2 was not measured.

- **Influence:** the main grid is $64, 128, 256, 512$. Extensions reach 2048 for v2.5, TabICLv2, classical refits, TabFM randomization scopes, and Mitra-higgs. All-deletion runs reach 512, except Mitra-higgs at 256. Mitra-magic's RMS also ends at 256 because its 512 matrices use another machine. TabPFN vendor ensembles cover both datasets through 512. The common query intervention uses one v2.5 checkpoint, two datasets, six seeds, these four main grid sizes, and 50 fixed unfiltered queries per dataset.
- **Spread:** all six checkpoints cover higgs and magic through 1024. Overlapping extensions reach 8192 for v2.5 and v1 on both datasets (v1 beyond its 1024 cap), and 4096 for TabICLv2-higgs and Mitra-higgs. Disjoint partition tests reach 4096 on magic and 8192 on higgs for v1/v2.5. Disjoint higgs reaches $32{,}768$ for v2.5, v3, v1, TabICLv2, and TabFM, including the latter two vendor defaults. Mitra ends at 8192, logistic regression at $65{,}536$. V2.5/TabICLv2 cover twelve additional datasets through 512. Vendor default v2.5 reaches 4096. Seeded Mitra-stack runs reach 1024, except macOS higgs at 256.
- **Replication:** influence uses six seeds, except three in legacy $k = 128$ sweeps and attack cost tests. Spread uses 24–48 draws, with 16 at disjoint higgs $32{,}768$. Overlap tests use 48–116 blocks across 4–32 partitions, and v1 synthetic spread uses 200–500 draws. Most analyses use 50 queries. Accuracy/label flip checks use 200, disjoint accuracy 2000, order probes 10, and v1 synthetic spread 100.

Spambase's control band failure excludes it from the main comparisons. Bank passes at a matched window. The TabICLv2 RMS deletion matrices are supplied; the other historical foundation model matrices are outside the ZIP (Appendix K). The complete outcome ledger, `docs/registration_outcomes.csv`, retains misses and restricted or unevaluable units. The specified label flip budget measurements make no defense or general attack success claim.