PER-INPUT PRIOR ADEQUACY: SOURCES OF INFORMATION AND CONDITIONS OF FAILURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular foundation models (TFMs) acquire their entire competence from a synthetic prior sampled at meta-training time, and their prediction errors are far from uniform across inputs. Prior evaluation scores a whole dataset with one number, and such a score cannot say where a model is misaligned. Applying those scores per input does not help either. A coverage-style score is uninformative, and a paired improvement ratio is blind to uniform prior gaps because it is a rate statistic. To address this, we propose the reference-averaged log-error ratio ($\bar{\ell}$, RA-ratio), a per-input, training-free diagnostic composed of two components, the excess loss of the model over a Bayesian reference and an average of that excess over the reference's prior strength. First, for each input we take the model's log error in excess of that of a closed-form Bayesian reference, the Bayes gap of in-context learning theory. Second, we average that excess over a grid of reference strengths, which removes the reference's strength as a tuning knob rather than selecting it. Third, we validate the diagnostic causally, injecting meta-training tasks from the region it flags and verifying that the error falls there and not in an off-target region. Notably, the RA-ratio attains a mean rank correlation of $|\rho| = 0.76$ with the per-point error on 22 regression datasets with 3 seeds, ahead of every reference-free baseline we tried including three conformal variants, and replicates on a second architecture. On a held-out classification family, prior-based per-input diagnostics help where the model's own uncertainty has stopped resolving difficulty, and hurt where it already succeeds.

1 INTRODUCTION

Tabular data is the dominant modality in applied machine learning, and the field has long been served by gradient-boosted trees. Tabular foundation models (TFMs) have recently changed that practice. A TFM performs in-context learning. Given a support set of labelled rows it predicts at new rows without any gradient update, and its entire competence comes from a synthetic prior over tasks that was sampled at meta-training time (Garg et al., 2023; Qu et al., 2025). Evaluating that prior is therefore the central question for the field, since a prior that is adequate on average but wrong on a region of input space gives a practitioner no way to tell where the model may be trusted (Figure 1).

The dominant approach has been to ask how similar synthetic tasks are to real ones. Recent work found this unproductive. Davies et al. (2026) searched $86{,}000$ prior configurations and could neither close the gap between synthetic and real tasks nor detect any effect of that gap on downstream performance; Türkmen et al. (2026) concluded that data-level similarity only partially explains downstream behaviour; Zhao et al. (2026) built a 21-dimensional structural coverage measure but stated its association cannot be read causally. We read these results as a statement about the resolution of the measurement rather than about priors, and we take the obstruction to be structural.

Relation to prior work. Scoring points by how badly a model does, and validating that score by a within-task rank correlation with the error, is established practice in failure analysis (Thiagarajan et al., 2023), instance-difficulty estimation (Lavi et al., 2026) and conditional-coverage diagnostics (Braun et al., 2025). Our level statistic is the excess-loss-over-Bayes, or Bayes gap, of in-context learning theory (Wakayama & Suzuki, 2025), and averaging out a prior-strength hyperparameter is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

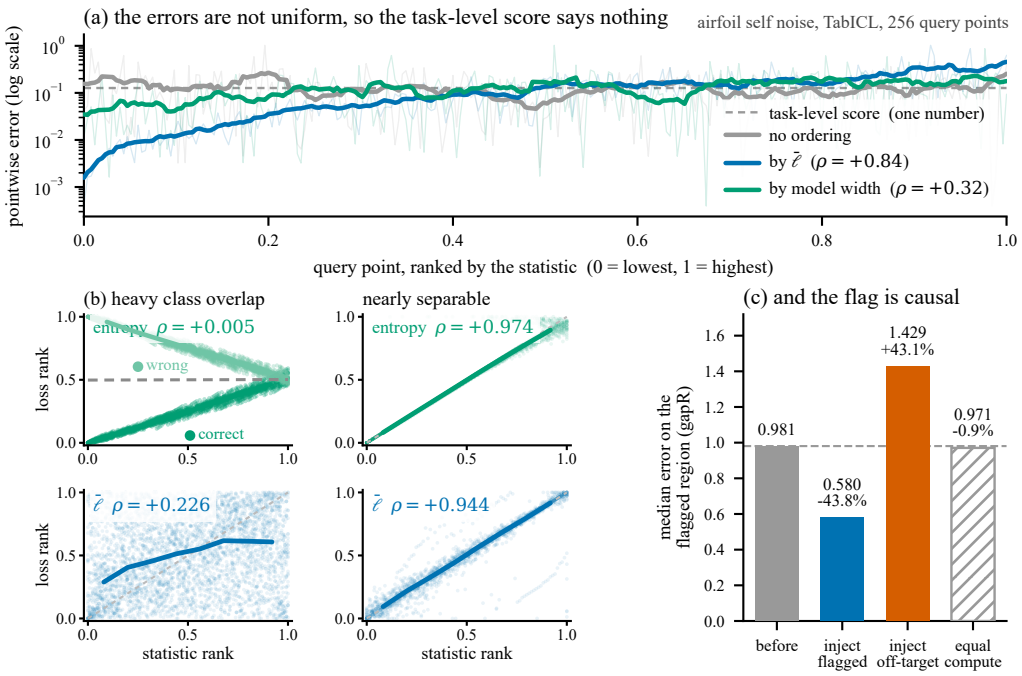

Figure 1: Per-input prior adequacy. (a) The errors of one dataset are far from uniform, yet a task-level score is a single flat line through them. The RA-ratio orders them ($\rho = +0.84$); the model's own predictive width does not ($+0.32$). (b) Per-point loss against each statistic at the two extremes of the class-overlap grid, as within-task ranks pooled over the 30 tasks of a cell and restricted to points not at the Bayes floor; the line is the pooled median loss rank per decile of the statistic, and the number is the mean within-task rank correlation, so no linear form is assumed. Where classes overlap heavily the model's own uncertainty is blind and the RA-ratio still orders points; where classes are nearly separable entropy already succeeds and the RA-ratio adds nothing. Shading in the left panel separates whether the class the model favours is the true one; the grey dashed line there is the pooled least-squares fit over both arms at once, and it is flat. (c) Injecting meta-training tasks from the region the diagnostic flags cuts the error there by $43.8\,\%$; injecting an off-target region makes it $43.1\,\%$ worse, and an equal-compute control does nothing.

what a hyper-g prior does in the same conjugate setting (Bové & Held, 2010). Our contribution is the assembly, per-input, training-free and reference-normalised, with the reference's nuisance strength marginalised out and the causal claim tested by injection. We write "per-input" rather than "pointwise", which already denotes pointwise coverage (Sluijterman et al., 2021).

A task-level score is constant within a task by construction, since it maps a task to a single number. It therefore cannot, even in principle, say where a model is misaligned. The distinction is not academic. Prior evaluation currently returns one number per dataset, while the errors of a TFM on that dataset are strongly non-uniform across inputs, so the same score is consistent both with a model whose errors are a uniform floor and with one whose errors are concentrated in a small region of input space. The second case is the one a user needs to know about, and it is the case the score cannot express. The literature above measured within-task correlations of $\approx +0.02$, which is the empirical face of this limitation. The argument is the marginal-versus-conditional distinction, and it is documented from several directions: exact conditional guarantees are impossible without distributional assumptions (Barber et al., 2019), the evaluation of regression uncertainty carries a critique along the same lines (Sluijterman et al., 2021), and the same argument has been made for TFMs whose uncertainty is benchmarked at the dataset level (Costa et al., 2026).

We therefore ask a different question. Instead of scoring tasks, we score points. For one trained TFM and one target task, which inputs is it misaligned on? The question matters beyond diagnosis. If a prior gap is localisable, then injecting meta-training tasks drawn from the flagged region should repair that region specifically, which makes the diagnostic falsifiable by intervention rather than only by correlation.

A useful per-input diagnostic has to be strong, and its scope has to be known. We first show that the natural per-input statistics are weak, and we identify the reason, namely that a rate statistic cancels a uniform prior gap by construction. The consequence is measurable, and it is stark: a trained model can sit above the reference at every support size while the rate statistic reports nothing. We then build a per-input diagnostic that is strong and reference-insensitive, validate it causally, and replicate it across two models. The scope condition matters as much as the diagnostic itself. Whether a prior-based per-input diagnostic is worth using depends on the ranking quality of the model's own uncertainty, a quantity already known to govern what a confidence signal buys in selective prediction (Rabanser & Papernot, 2025; Galil et al., 2023). We validate that rule out of sample, and we falsify the intuitive alternative, that the diagnostic helps more as the epistemic share grows.

Contributions.

1. **We show that the natural per-input extensions fail.** Per-input coverage scores $|\rho| = 0.081$ and ψ scores 0.310, indistinguishable from the model's own predictive width, and ψ is blind to uniform prior gaps, since it is a rate statistic (Thiagarajan et al., 2023; Sluijterman et al., 2021). On a trained model the effect is direct, since a model deliberately given a uniform gap is ranked by ψ below a gap-free one while the level statistic separates them cleanly.
2. **We introduce the RA-ratio, a per-input, training-free diagnostic.** Its level form is the Bayes gap (Wakayama & Suzuki, 2025) and its averaging step a hyper-g-style marginalisation of the reference's prior strength (Bové & Held, 2010; Cousins & Wasserman, 2024). We validate it end to end. It attains $|\rho| = 0.759$, ahead of every reference-free baseline we tried including three conformal variants, survives a matched averaging control, and replicates on a second architecture and two non-PFN learners.
3. **We validate the diagnostic causally, and derive the scope rule that governs its use.** Injection helps exactly where the diagnostic points, with specificity against off-target and equal-compute controls—an established pattern, applied here to a synthetic prior at per-input resolution (Gupta et al., 2026; Jha, 2026; Sabry & Belz, 2025). On closed-form Bayes posteriors, the ranking strength of the model's own uncertainty rather than the epistemic share predicts the sign of the increment (AUC 0.888 vs 0.524, $1/18$ held-out cells).

2 Related Work

Prior evaluation and per-input reliability. Prior evaluation for TFMs is active and dataset-level: structural coverage over data-side descriptors (Zhao et al., 2026); discriminator AUC and k-NN coverage between synthetic and real corpora (Davies et al., 2026); data-space and performance-space similarity (Türkmen et al., 2026); and prior design, where mechanism diversity, realism and shift-aware stress contribute independently (Bouadi et al., 2026). Adjacent work finds TFMs internally inconsistent (Klötergens et al., 2026), argues for proper scoring rules in TFM benchmarks (Landsgesell et al., 2026a), and uses prior distance as a cleaning diagnostic (Berti-Equille, 2026). Per-input reliability is established as well, and so is our generic negative result: intervals measure only marginal coverage unless pointwise coverage is demanded (Sluijterman et al., 2021), regression uncertainty is necessary but insufficient to characterize per-instance failure (Thiagarajan et al., 2023), instance difficulty is scored by within-task rank correlation on tabular data (Lavi et al., 2026), and per-prediction audits find high-confidence brittleness that calibration misses (Cenacchi et al., 2026). The object here differs: the adequacy of a synthetic meta-training prior for an in-context model, against a closed-form Bayesian reference and without training a scorer.

Conditional coverage, and why we rank rather than guarantee. Exact distribution-free conditional coverage is impossible (Barber et al., 2019), so conditional reliability is approached through classes of covariate shifts (Gibbs et al., 2023), non-asymptotic decompositions of miscoverage (Min et al., 2026), or a per-point excess risk of the target coverage (Braun et al., 2025). That is the sharper, higher-power instrument; we rank instead, because our object is the ordering of inputs by error, which is what a decision to acquire data acts on, and ranking carries no coverage guarantee. CQR (Romano et al., 2019) and locally-adaptive conformal (Gibbs et al., 2023) are our reference-free competitors and score 0.084–0.128; TFM miscalibration is documented independently (Costa et al., 2026; Izbicki & Rodrigues, 2026; Landsgesell et al., 2026b; Johnson et al., 2026).

Provenance of the components. Our level statistic is the excess loss over a Bayesian reference in log form—the Bayes gap of in-context learning theory (Wakayama & Suzuki, 2025)—and PFN theory already predicts that the bias vanishes only if the predictor localizes around the test feature (Nagler, 2023), with prior-induced bias that data does not remove (Melnychuk et al., 2026). Removing a prior-strength hyperparameter by averaging is what a hyper-g prior does in the same conjugate setting (Bové & Held, 2010), marginalising and profiling can diverge (Cousins & Wasserman, 2024), prior sample size is how prior strength is quantified (Reimherr et al., 2014), and prior–data conflict is a developed diagnostic (Nott et al., 2016). Our aleatoric/epistemic split is the standard proper-scoring-rule decomposition (Hofman et al., 2024; 2025), already applied to TabPFN where aleatoric uncertainty dominates near overlapping classes (Fortini et al., 2026), and additive decompositions can break down under mismatched predictives (Gillis et al., 2026).

Ranking quality, guarantees, and targeted repair. What makes a confidence signal useful is its ranking quality rather than its magnitude (Rabanser & Papernot, 2025), how well it ranks errors varies across models and regimes (Galil et al., 2023), and which uncertainty source to report is task-dependent (Hofman et al., 2025); our scope rule applies that principle and falsifies the epistemic-share alternative. A Bayesian reference is the right yardstick because a transformer can implement Bayesian inference in-context (Garg et al., 2023), PFN predictions admit a martingale posterior reading (Nagler & Rügamer, 2026), and a deterministic-limit pathology is documented (Tenachi et al., 2026). Diagnose-then-target repair with specificity controls is established (Gupta et al., 2026; Sabry & Belz, 2025), including the finding that targeted repair can lose to equal-compute untargeted resampling (Jha, 2026)—the bar our control arms must clear. Finally, decision distance vectors compare models (Li et al., 2021), whereas we measure one model's per-input reliability with support size as the independent variable; TabM (Gorishniy et al., 2025), TALENT (Liu et al., 2024) and TabArena (Erickson et al., 2025) are the toolboxes and benchmark we build on, and since every baseline we beat is model-agnostic we re-run the protocol on a gradient-boosted learner and on ridge regression (Section 5).

3 METHOD

3.1 SETUP AND NOTATION

A TFM observes a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{m}$ and predicts at query points x. We probe the model at a nested grid of support sizes $m \in \{16, 24, \ldots, m_{\max}\}$ (nine points up to $m_{\max} = 256$; the replication model uses a reduced three-point grid, and the classification arm probes $\{64, 192, 512\}$) and record its predictive summary $p_{\text{model}}(y \mid x; m)$, a predictive median and IQR for regression and the class distribution for classification. All statistics below are per-input and are evaluated by their within-task rank correlation with the per-input error.

Errors are quoted in each dataset's own units. For regression, e is the model's absolute deviation from the target divided by the spread of the support targets (their interquartile range, $Q_3 - Q_1$), $e(x; m) = |\hat{y}(x; m) - y|/\sigma_T$; for classification it is the Brier loss $(p(y \mid x, m) - y)^2$. The asymmetry is deliberate: classification is scored on its predictive distribution, regression on its median. Model and reference share the support set and the query points, so every comparison below is paired point by point. The reference is a Bayesian linear model with a normal-inverse-gamma prior, $y = X\beta + \epsilon$ with $\beta \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2 \kappa_0^{-1} I)$ and $\sigma^2 \sim \text{InvGamma}(10^{-2}, 10^{-2})$, fitted to standardised features and read through its Student-t predictive distribution; in the classification arm the same model is applied to the $0/1$ target and clipped to $[0, 1]$.

3.2 WHY RATE STATISTICS FAIL

Define the paired improvement ratio

$$\psi(x) = \big[\log e_{\text{model}}(x; m_{\min}) - \log e_{\text{model}}(x; m_{\max})\big] - \big[\log e_{\text{ref}}(x; m_{\min}) - \log e_{\text{ref}}(x; m_{\max})\big], \quad (1)$$

where e is absolute error in task-scale units (regression) and e_{ref} is the same error under a Bayesian reference. ψ compares how much the model and the reference improve as evidence accumulates. If the model is uniformly worse than the reference at every m (a uniform prior gap), both improve by similar amounts, the difference cancels, and $\psi \approx 0$ even though the model is badly misaligned. This cancellation is an identity of log-differences rather than a finding: if $e_{\text{model}}(x; m) = c\,\tilde{e}(x; m)$

with c free of m, then $\log e_{\text{model}}(x; m_{\min}) - \log e_{\text{model}}(x; m_{\max})$ equals the same expression in $\tilde{e}$, term by term, and the reference side behaves identically. That scale-freeness lets ψ compare datasets whose errors live in different units, and it is also what hides a gap uniform in m. We therefore state it as an identity and claim only the measurement of what it costs in practice, namely that a trained model can be misaligned at every support size while ψ reports nothing. Section 5 measures that by intervention.

3.3 The RA-ratio, a reference-averaged log error ratio

The level statistic compares the model's error to the reference's at the same evidence level:

$$\text{level}(x; \kappa_0) = \log e_{\text{model}}(x; m_{\max}) - \log e_{\text{ref}}(x; m_{\max}; \kappa_0), \tag{2}$$

where κ_0 is the reference model's prior strength. In log form this is the excess loss of the model over the Bayesian reference, i.e. the Bayes gap of in-context learning theory (Wakayama & Suzuki, 2025), and we do not present it as a new object. It detects uniform gaps, but it is also sensitive to κ_0 when the model and task are held fixed, because it absorbs the calibration of the reference. In a factorial sweep that moves only κ_0, $|\rho|$ for a single reference swings from 0.073 to 0.680 (Section 5), so the statistic's usefulness depends on a knob that has nothing to do with the model.

That sensitivity is not incidental, and removing it does not require a new idea. Write the quantity of interest as $g(x) - h(x; \kappa_0)$, where $g(x) = \log e_{\text{model}}(x; m_{\max})$ does not depend on κ_0 at all and h does. Every bit of the κ_0-dependence therefore lives in h alone, so the smallest correction that leaves the model side untouched is to replace h by a κ_0-free summary of it. Averaging is the natural summary here because the statistic is already a difference of log-errors. The mean is the only summary that keeps $\bar{\ell}$ on exactly the same scale as the level it corrects, so its value can still be read as a log error ratio, which a median or a maximum would not preserve. This is the ordinary move of removing a nuisance hyperparameter by averaging it out—a discrete hyperprior on the reference's strength, in the spirit of the hyper-g prior for a conjugate regression (Bové & Held, 2010); profiling rather than marginalising is known to behave differently outside regular asymptotics (Cousins & Wasserman, 2024). The construction is therefore a parameter elimination rather than a modelling choice, and its value is that no reference strength has to be picked:

$$\begin{aligned} \bar{\ell}(x) &= \frac{1}{|\mathcal{K}|} \sum_{\kappa_0 \in \mathcal{K}} \left[\log e_{\text{model}}(x; m_{\max}) - \log e_{\text{ref}}(x; m_{\max}; \kappa_0) \right] \;=\; g(x) - \bar{h}(x), \\ \mathcal{K} &= \{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}. \end{aligned} \tag{3}$$

We refer to $\bar{\ell}$ as the RA-ratio from here on. The set $\mathcal{K}$ and its equal weights are a discretisation of a hyperprior on the reference's strength, and a smooth integral is the principled limit of what we compute; the step buys invariance to the reference's strength rather than a particular grid. We test that the set is not doing hidden work in two ways. The matched control below applies the identical averaging to the reference's own width, and Section 5 reports that the RA-ratio responds monotonically to the gap while staying fixed under the reference knob. Two further properties matter. First, because the model term does not depend on κ_0, the RA-ratio responds to task-relevant variation and to the reference-side knob asymmetrically, which the factorial quantifies. Second, because the reference is available in closed form (a normal-inverse-gamma linear reference), the six terms cost essentially nothing, because the model probe grid is run once.

The strength κ_0 behaves as a prior sample size, since the posterior precision is $\kappa_0 I + X^\top X$: small values let the support set dominate, large ones hold the reference near its prior, and that is why sweeping it moves the single-reference level statistic.

Matched control. The critical alternative explanation is that averaging per se manufactures discrimination. We therefore apply the identical averaging operation to the reference's own width, $\text{width_multi_ref}(x) = \frac{1}{|\mathcal{K}|} \sum_{\kappa_0} w_{\text{ref}}(x; \kappa_0)$. If averaging were the mechanism, this control would score as well as the RA-ratio.

4 Experimental Setup

Models. TabICL (Qu et al., 2025) is the primary model. TabPFN 2.5 is used for replication, with a reduced 3-point support grid $\{16, 64, 256\}$ because its per-call cost is ≈ 30 s independent of the

number of query points. level and $\bar{\ell}$ are read at $m_{\max}$ and are therefore comparable across the two grids. Both models are PFN-family, so our claims are scoped to that family. The diagnostic is defined against a Bayesian reference and is therefore in principle most natural for models whose predictive distribution can be read as a posterior. We return to what this does and does not license in Appendix B.

Regression data. 22 regression datasets from BeyondArena (Erickson et al., 2025), 3 seeds each (66 rows in the main table). Each row contributes one within-task rank correlation between a statistic and the per-input error, so the unit of analysis is one dataset and seed combination, and all reported intervals are paired bootstrap intervals over those units.

Baselines. The model's own predictive width (IQR); a coverage-style centroid-distance score; split conformal (excluded as degenerate because it assigns one constant width per task and therefore cannot rank); CQR; locally-adaptive conformal (LACP); and the matched averaging control above.

Controlled priors. A 1-D PFN trained from scratch in four arms that differ only in their meta-training distribution, and a factorial design over 14 cells spanning a noise axis ($\sigma \in \{0.05, \ldots, 2\}$), a dimensionality axis and a nonlinearity axis.

Classification. 11 real datasets, plus a synthetic family with a closed-form Bayes posterior (Section 5.5), a two-Gaussian generative model with a class-overlap knob, 6 overlap levels $\times$ 3 context sizes, 30 tasks per cell (540 tasks).

5 Results

5.1 The natural per-input statistics are weak

Reconstructing N1-style coverage faithfully, with the original read-out function and a synthetic cloud of 64 tasks drawn from TabICL's exact pre-training prior, gives $|\rho| = +0.081$ against -0.287 for ψ; the paired difference is $+0.368$ $[+0.274, +0.460]$, ψ better on $21/21$ datasets. The synthetic cloud is the one stochastic ingredient, so what we rely on is the draw-invariant paired gap rather than the 0.081 point value; nor is ψ an improvement on what a user already has, since its 0.3099 is statistically indistinguishable from the model's own predictive width 0.2733. Both comparisons, and their cloud-draw and task-level checks, are in Appendix B.

5.2 Why ψ fails, because a uniform gap is invisible to a rate statistic

A controlled 1-D PFN trained from scratch in four arms that differ only in their meta-training distribution isolates the mechanism (Table 3, Appendix B, and Figure 2). On the *gapR* arm the model is uniformly worse than the reference at every m ($+74.6\,\%$ relative error), so both improve by similar amounts and the rate ratio collapses: ψ does not merely fail to flag *gapR*, it ranks it below the non-gap *covered* (-0.196 vs -0.009), while the level statistic separates the two cleanly ($+0.826$ vs -0.004). The error curves show why: on *gapR* the model sits above the reference at every m, an offset rather than a divergence, which is exactly what a difference of improvement rates cannot see.

The fix. Averaging the level statistic over reference strengths removes exactly this knob: across the sweep the single-reference level statistic moves over 0.073–0.680, while $\bar{\ell}$ is invariant to κ_0 and still falls monotonically with the noise σ that should matter (Figure 3).

5.3 Main result

Controls. The RA-ratio survives conditioning on distance-to-support (0.7600) and partialling out both the model's width and a coverage-style centroid score (0.7595), and its advantage is not explained by the width ratio (Spearman $= +0.063$); the centroid score itself reaches 0.242 on TabICL and 0.252 on TabPFN 2.5. The matched averaging control rules out the "averaging artefact" explanation: the identical averaging applied to the reference's own width yields 0.084, so the gain comes from the quantity being averaged, not from the averaging. A flexible reference does cost strength—at matched effective degrees of freedom, a quadratic and two random-Fourier-feature references move the mean $|\rho|$ from 0.867 to 0.835 and 0.625–0.663 (Appendix B)—so we read 0.76 as a lower bound and describe the statistic as measuring prior-relative adequacy rather than claiming

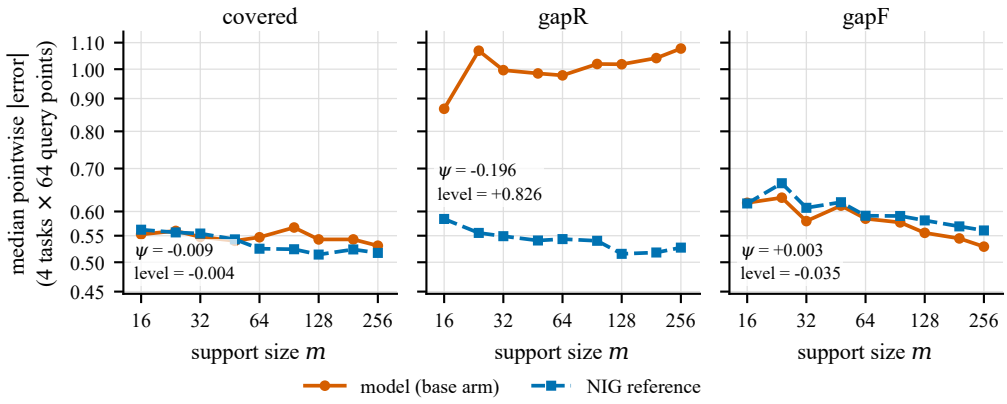

Figure 2: Why a rate statistic is blind to a uniform prior gap. Median per-input error (4 tasks $\times$ 64 query points) against support size m for the *base* arm (solid) and the NIG reference (dashed), one panel per mechanism. On *gapR* the model sits above the reference at every m, an offset of $+0.40$ to $+0.83$ in log error rather than a difference in rate, so the paired improvement ratio cancels ($\psi = -0.196$) while the level statistic detects the gap ($+0.826$). On *covered* and *gapF* the two curves nearly coincide.

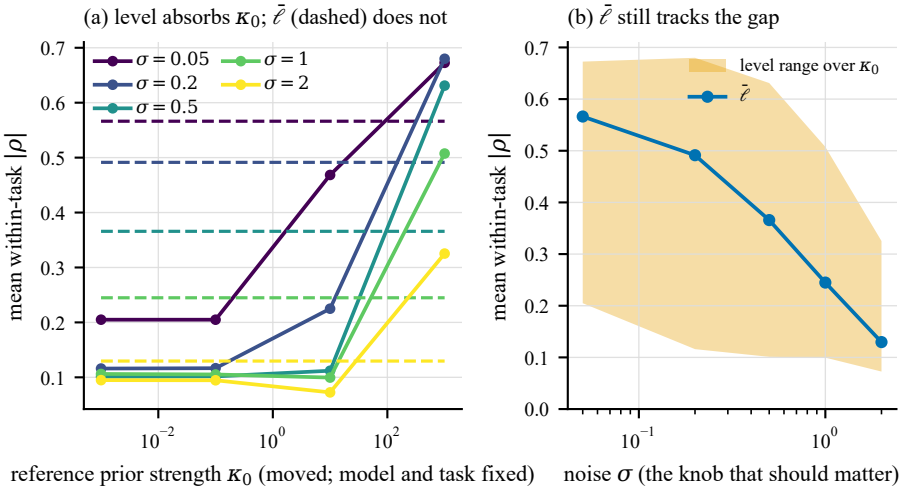

Figure 3: The reference knob moves the level statistic and leaves $\bar{\ell}$ alone. (a) Sweeping only the reference prior strength κ_0 while model and task are held fixed. The single-reference level statistic (solid) fans out from $|\rho| = 0.205$ to 0.673 at $\sigma = 0.05$, whereas $\bar{\ell}$ (dashed, same colour) is constant within each noise level. (b) The same cells ordered by the knob that should matter. $\bar{\ell}$ falls monotonically with σ, from 0.566 to 0.130, while the shaded band shows the range the single-reference version spans over κ_0.

that it isolates the prior gap alone. Figure 4 (Appendix B) shows the per-dataset paired differences behind the ordering in Table 1.

5.4 Causal validation

Table 2 reports the three arms (targeted, off-target, and an equal-compute control on a doubled covered prior).

Specificity (paired, negative = helps its own target more) is -0.426 $[-0.527, -0.332]$ and -0.459 $[-0.524, -0.388]$.

The flag itself moves. The injection experiment is also a direct read-out of the statistic. On the *base* arm the RA-ratio's median level is $+1.40$ on *gapR* against -0.29 on *covered*; training on *gapR* drops

statistic	TabICL $\lvert\rho\rvert$	95 % CI	paired vs $\bar{\ell}$	TabPFN 2.5
$\bar{\ell}$ **(ours)**	$\mathbf{0.7592}$	$[0.7141, 0.8001]$	—	$\mathbf{0.7446}$
level (single κ)	0.5608	$[0.5042, 0.6147]$	$+0.198$	0.5422
ψ	0.3099	$[0.2478, 0.3676]$	$+0.449$	0.2682
width	0.2733	$[0.2070, 0.3436]$	$+0.486$	0.2536
LACP (conformal)	0.1276	$[0.0888, 0.1739]$	$+0.632$	0.1514
CQR (conformal)	0.0857	$[0.0572, 0.1197]$	$+0.674$	0.0843
width_multi_ref (matched control)	0.0840	$[0.0541, 0.1177]$	$+0.675$	0.0847

Table 1: Mean within-task $\lvert\rho\rvert$ over 22 regression datasets. The last column is the replication on TabPFN 2.5 under the identical protocol on a coarser support grid; the paired column is the paired bootstrap difference from $\bar{\ell}$ on TabICL, with $\bar{\ell}$ better on all of them and on $22/22$ datasets. The matched control applies the identical averaging to the reference's own width.

arm	on covered	on *gapR*	on *gapF*
inject *gapR*	$-0.7\,\%$	$-43.8\,\%$ $[-53.5, -33.6]$	$-0.5\,\%$ $[-0.75, +0.18]$
inject *gapF*	$+0.4\,\%$	$+43.1\,\%$ $[+36.3, +48.0]$	$-6.2\,\%$
equal-compute control	$-0.003\,\%$	$-0.9\,\%$	$+0.02\,\%$

Table 2: Relative change in per-input error for each arm: targeted (*gapR*), off-target (*gapF*) and equal-compute. Only the targeted arm moves its own region.

that level by 1.30–1.53 over three independent task draws, injecting the off-target region raises it by $+0.30$ to $+0.38$, and the equal-compute control moves it by -0.01. The sign of the causal claim is reference-invariant; its magnitude is not (Appendix B).

5.5 Conditions of failure in classification

On 11 real classification datasets the model's own predictive entropy explains the per-input loss almost completely ($\lvert\rho\rvert = 0.822$), and partialling it out leaves -0.055 $[-0.188, +0.085]$. We therefore move to a synthetic family where the decomposition is exact: with a closed-form Bayes posterior p^*, the aleatoric (Bayes risk) and epistemic (excess over Bayes) parts are $\ell_{\text{ale}}(x) = \frac{1}{C}(1 - \sum_c p_c^*(x)^2)$ and $\ell_{\text{epi}}(x) = \frac{1}{C}\sum_c (p_c(x) - p_c^*(x))^2$, with $\ell_{\text{ale}} + \ell_{\text{epi}} = \mathbb{E}_y[\text{brier}(p, y)]$ verified to 10^{-5} at every separation, so no proxy is involved. It is exact here because the Bayes posterior is known; additive decompositions can break down under mismatched predictives (Gillis et al., 2026), so we scope it to this family.

A controlled grid (6 class-overlap levels $\times$ 3 context sizes, 30 tasks per cell) gives a sharply powered boundary on the aleatoric level, with the opposite sign to the natural hypothesis (Table 7, Appendix C).

Splitting at the median aleatoric level gives 9/9 high-aleatoric cells with a positive increment and 0/9 low-aleatoric cells with none (Mann–Whitney one-sided p = 0.0002); Spearman(overlap, increment) $= -0.843$, $p < 0.0001$.

The aleatoric level is the wrong name for this boundary. It and entropy strength are near-collinear here (Spearman -0.938), and partialling each out of the other on all 540 tasks leaves the aleatoric level carrying nothing (partial -0.064) while entropy strength retains its signal (partial -0.422). The boundary is about the model's self-knowledge, not about the noise floor: the increment is positive where the model's own uncertainty has stopped resolving per-input difficulty, and negative where it already resolves it. Aleatoric uncertainty dominates where classes overlap is known (Fortini et al., 2026); it is not, however, the variable that governs whether a reference-based diagnostic is worth using. Since the two are near-collinear, our evidence for separability is the partial correlation rather than a proof.

The 2×2 median split over the 540 tasks puts the largest gain where the diagnostic is strong and entropy weak ($+0.421$) and the largest loss where entropy is strong (-0.637).

The sign is set by entropy strength alone; the diagnostic's own strength only sets the size of the penalty when entropy wins. The result is a scope rule, and it is where the original hypothesis fails.

Whether a confidence signal buys anything is set by how well it ranks errors, not by its magnitude (Rabanser & Papernot, 2025; Galil et al., 2023); the new evidence is the application to a reference-based per-input diagnostic and the falsification, on held-out cells, of the competing epistemic-share explanation. Figure 5 (Appendix C) shows the same boundary per controlled cell and as the resulting 2×2 rule.

Out-of-sample check. We pre-registered G1–G5 and scored three hypotheses by the same protocol, learning both the sign convention and the decision threshold on training cells (the scope rule is not centred on zero, so a threshold-free ranking statistic cannot predict a sign at all). Table 6 (Appendix C) reports the three protocols. H_{gate}, the entropy-strength statistic, passes all five criteria, and H_{orig} predicts sign worse than chance on held-out data ($1/18$ cells), the cleanest form the falsification could take. H_{ale} matches H_{gate} at cell resolution ($18/18$) and trails it on every row-level measure, so we report the two as near-equivalent predictors and name the boundary by the mechanism rather than by either proxy. We avoid the term confidence gate, which denotes monotone abstention on ranked decision systems (Doku, 2026).

The applicable domain is therefore the opposite of the intuitive one. When classes are nearly separable the model's entropy already identifies difficulty almost perfectly ($|\rho| = 0.87$–0.98) and any reference-based statistic is noise, so the RA-ratio is worse by up to 0.74. When classes overlap heavily, entropy collapses ($|\rho| = 0.005$–0.031) while the RA-ratio retains $|\rho| = 0.23$–0.42. But the quantity that decides which regime a task is in is not the noise floor. It is whether the model's own uncertainty is still informative. A reference-based diagnostic is a **conditionally available instrument**. It should be reached for where the model's self-reported uncertainty fails, and ignored where it already succeeds.

5.6 Replication on a second model

TabPFN 2.5 costs a fixed ≈ 30 s per *fit+predict* independent of the number of query points, so the replication uses the coarser grid $\{16, 64, 256\}$ with 3 seeds, with level and $\bar{\ell}$ read at $m_{\max}$ in both models and therefore comparable (Table 1). Two differently-architected models with different priors order the statistics almost identically (Spearman $+0.976$), and the matched control takes nearly the same value on both, so the strength relations are properties of the measurement rather than of one model; $\bar{\ell}$ beats ψ pairwise on $22/22$ datasets on the second model too ($+0.476$ $[+0.404, +0.549]$). Re-running the identical protocol on a histogram gradient-boosted learner and on ridge regression leaves the ordering intact ($|\rho| = 0.733$ and 0.753, top statistic on both) but weakens the effect size, so the ordering transfers and the magnitude does not (Appendix B).

6 Conclusion

Per-input prior adequacy is measurable, but only if the statistic is chosen for the failure mode that matters. The pieces are known, and the contribution is the combination: a rate-based statistic is blind to uniform prior gaps and barely beats the model's own uncertainty; a level-based statistic is strong but absorbs the reference's calibration; averaging the level over reference strengths resolves both, giving $|\rho| = 0.759$ with a matched control that rules out the averaging explanation. Targeted prior injection then confirms the diagnostic causally, with specificity, and that injection experiment, applied to a synthetic meta-training prior at per-input resolution, is the part of this work with the least precedent.

The scope condition matters as much as the method: these diagnostics earn their keep precisely where the model's own uncertainty has stopped being informative, and hurt where it already succeeds. Under heavy class overlap $|\rho_{\text{entropy}}|$ falls to 0.005–0.031 while the RA-ratio keeps 0.23–0.42; where the model is well separated and self-aware ($|\rho_{\text{entropy}}| \approx 0.98$) it is worse by up to 0.74. That a confidence signal's usefulness is governed by its ranking quality rather than its magnitude is known from selective prediction; what is new is the application to a reference-based prior diagnostic and the falsification of the competing explanation, which survives calibration at one evidence level and application at another (AUC 0.888).

Limitations. $\bar{\ell}$ is a within-task ranker, not a cross-task score; the regression arm needs $m_{\max} \geq 192$; the reference is weak by construction, so 0.76 is a lower bound; and the non-PFN replication is single-seed. Appendix B states the full scope.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing, to identify and search for relevant literature, and to implement and refactor parts of the analysis code; we made no other use of them. All research decisions, derivations, experimental designs, statistical criteria and reported numbers were specified by the authors, and every AI-assisted change was run and checked against the saved result files. We take responsibility for the final content of this work, including text, claims and artifacts.

ETHICS STATEMENT

This work studies the reliability of predictive uncertainty in tabular foundation models using public benchmark datasets and synthetic data; it involves no human subjects and releases no personal data. We see no ethics concerns specific to this work.

REPRODUCIBILITY STATEMENT

All results are produced by self-contained scripts in a single package, driven by fixed seeds and saved as plain CSV/JSON artifacts together with a machine-readable verdict file per experiment. The Appendix lists the exact commands, including a re-analysis path that re-derives the classification scope analysis without re-running the model. The statistics whose behaviour we rely on are verified against earlier published values before new columns are trusted.

REFERENCES

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference, 2019. URL https://arxiv.org/abs/1903.04684.

Laure Berti-Equille. Model-aware data cleaning for tabular foundation models, 2026. URL https://arxiv.org/abs/2604.25154.

Mohamed Bouadi, Nassim Bouarour, Varun Kulkarni, Shivam Dubey, Aditya Tanna, and Vinay Kumar Sankarapu. Shaping the prior: How synthetic task distributions determine tabular foundation model quality, 2026. URL https://arxiv.org/abs/2605.18971.

Daniel Sabanés Bové and Leonhard Held. Hyper-g priors for generalized linear models, 2010. URL https://arxiv.org/abs/1008.1550.

Sacha Braun, David Holzmüller, Michael I. Jordan, and Francis Bach. Conditional coverage diagnostics for conformal prediction, 2025. URL https://arxiv.org/abs/2512.11779.

Filippo Cenacchi, Longbing Cao, and Runze Yang. Counterfactual fragility certificates: Exposing high-confidence brittleness under structured evidence failure, 2026. URL https://arxiv.org/abs/2609.00366.

José Lucas De Melo Costa, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. High performance, low reliability: Uncertainty benchmarking for tabular foundation models, 2026. URL https://arxiv.org/abs/2605.28554.

Robert D. Cousins and Larry Wasserman. Phystat informal review: Marginalizing versus profiling of nuisance parameters, 2024. URL https://arxiv.org/abs/2404.17180.

Alex O. Davies, Telmo de Menezes e Silva Filho, and Nirav Ajmeri. Mind the gap? a distributional comparison of real and synthetic priors for tabular foundation models, 2026. URL https://arxiv.org/abs/2605.06343.

Ronald Doku. The confidence gate theorem: When should ranked decision systems abstain?, 2026. URL https://arxiv.org/abs/2603.09947.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data, 2025. URL https://arxiv.org/abs/2506.16791.

Sandra Fortini, Kenyon Ng, Sonia Petrone, Judith Rousseau, and Susan Wei. Uncertainty decomposition for bayes-filtered transformers via bayesian predictive inference, 2026. URL https://arxiv.org/abs/2602.04596.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers, 2023. URL https://arxiv.org/abs/2302.11874.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes, 2023. URL https://arxiv.org/abs/2208.01066.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees, 2023. URL https://arxiv.org/abs/2305.12616.

H. Martin Gillis, Isaac Xu, and Thomas Trappenberg. Variance-gated ensembles: An epistemic-aware framework for uncertainty estimation, 2026. URL https://arxiv.org/abs/2602.08142.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. Tabm: Advancing tabular deep learning with parameter-efficient ensembling, 2025. URL https://arxiv.org/abs/2410.24210.

Vipul Gupta, Zihao Wang, Razvan-Gabriel Dumitru, MohammadHossein Rezaei, Aakash Sabharwal, and Yunzhong He. Craft: Clustering rubrics to diagnose weak llm capabilities and generate targeted fine-tuning data, 2026. URL https://arxiv.org/abs/2607.16122.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty with proper scoring rules, 2024. URL https://arxiv.org/abs/2404.12215.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Uncertainty quantification with proper scoring rules: Adjusting measures to prediction tasks, 2025. URL https://arxiv.org/abs/2505.22538.

Rafael Izbicki and Pedro L. C. Rodrigues. Benchmarking tabular foundation models for conditional density estimation in regression, 2026. URL https://arxiv.org/abs/2603.26611.

Anik Jha. Does fault localization beat a fresh attempt? a placebo-controlled study of test-guided code repair, 2026. URL https://arxiv.org/abs/2609.00854.

Tyler R. Johnson, Kian Ben-Jacob, Nima Negarandeh, Oriol Vendrell-Gallart, and Ramin Bostanabad. On the uncertainty quantification ability of tabular foundation models, 2026. URL https://arxiv.org/abs/2606.01427.

Christian Klötergens, Vijaya Krishna Yalavarthi, Lars Schmidt-Thieme, and Tom Hanika. Do tabular foundation models agree with themselves?, 2026. URL https://arxiv.org/abs/2608.06004.

Jonas Landsgesell, Pascal Knoll, and Tizian Wenzel. Distributional regression with tabular foundation models: Evaluating probabilistic predictions via proper scoring rules, 2026a. URL https://arxiv.org/abs/2603.08206.

Jonas Landsgesell, Pascal Knoll, and Tizian Wenzel. Scoringbench: A benchmark for evaluating tabular foundation models with proper scoring rules, 2026b. URL https://arxiv.org/abs/2603.29928.

Tomer Lavi, Bracha Shapira, and Nadav Rappoport. Trajectory-based difficulty scoring for reliable learning on tabular data, 2026. URL https://arxiv.org/abs/2605.24680.

Yuanchun Li, Ziqi Zhang, Bingyan Liu, Ziyue Yang, and Yunxin Liu. Modeldiff: Testing-based dnn similarity comparison for model reuse detection, 2021. URL https://arxiv.org/abs/2106.08890.

Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and Han-Jia Ye. Talent: A tabular analytics and learning toolbox, 2024. URL https://arxiv.org/abs/2407.04057.

Valentyn Melnychuk, Vahid Balazadeh, Stefan Feuerriegel, and Rahul G. Krishnan. Frequentist consistency of prior-data fitted networks for causal inference, 2026. URL https://arxiv.org/abs/2603.12037.

Yinjie Min, Liuhua Peng, and Changliang Zou. A unified theory of conditional coverage in conformal prediction with applications, 2026. URL https://arxiv.org/abs/2605.11602.

Thomas Nagler. Statistical foundations of prior-data fitted networks, 2023. URL https://arxiv.org/abs/2305.11097.

Thomas Nagler and David Rügamer. Uncertainty quantification for prior-data fitted networks using martingale posteriors, 2026. URL https://arxiv.org/abs/2505.11325.

David J. Nott, Xueou Wang, Michael Evans, and Berthold-Georg Englert. Checking for prior-data conflict using prior to posterior divergences, 2016. URL https://arxiv.org/abs/1611.00113.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data, 2025. URL https://arxiv.org/abs/2502.05564.

Stephan Rabanser and Nicolas Papernot. What does it take to build a performant selective classifier?, 2025. URL https://arxiv.org/abs/2510.20242.

Matthew Reimherr, Xiao-Li Meng, and Dan L. Nicolae. Prior sample size extensions for assessing prior impact and prior–likelihood discordance, 2014. URL https://arxiv.org/abs/1406.5958.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression, 2019. URL https://arxiv.org/abs/1905.03222.

Mohammed Sabry and Anya Belz. Induction signatures are not enough: A matched-compute study of load-bearing structure in in-context learning, 2025. URL https://arxiv.org/abs/2509.22947.

Laurens Sluijterman, Eric Cator, and Tom Heskes. How to evaluate uncertainty estimates in machine learning for regression?, 2021. URL https://arxiv.org/abs/2106.03395.

Wassim Tenachi, Yashar Hezaveh, Laurence Perreault Levasseur, and Pierre-Luc Bacon. Do tabular foundation models know physics? contamination, units, and the deterministic limit, 2026. URL https://arxiv.org/abs/2609.02766.

Jayaraman J. Thiagarajan, Vivek Narayanaswamy, Puja Trivedi, and Rushil Anirudh. Pager: A framework for failure analysis of deep regression models, 2023. URL https://arxiv.org/abs/2309.10977.

Zeynep Türkmen, Kürşat Kaya, Alexander Pfefferle, and Frank Hutter. Towards evaluating data priors for tabular foundation models, 2026. URL https://arxiv.org/abs/2606.29241.

Tomoya Wakayama and Taiji Suzuki. In-context learning is provably bayesian inference: A generalization theory for meta-learning, 2025. URL https://arxiv.org/abs/2510.10981.

He Zhao, Ryan Thompson, Daniel M. Steinberg, Ashfaqur Rahman, Edwin V. Bonilla, and Cheng Soon Ong. From synthetic priors to model behavior: Structural coverage in tabular foundation models, 2026. URL https://arxiv.org/abs/2609.06912.

A Process and corrections

Recorded so the negative results are not lost; none of this belongs in the narrative above.

1. **A prior-art audit narrowed the claims in this paper, and we record what changed.** A literature search run after the experiments were complete found that (i) the excess of log error over a Bayesian reference is a named object, the Bayes gap (Wakayama & Suzuki, 2025); (ii) removing a prior-strength hyperparameter by averaging is what a hyper-g prior does in the same conjugate family (Bové & Held, 2010); (iii) the generic claim that uncertainty is insufficient to localize per-input error is already published for regression (Thiagarajan et al., 2023; Sluijterman et al., 2021); (iv) per-input difficulty scoring validated by within-task rank correlation, with many of our baselines, predates us on tabular data (Lavi et al., 2026); (v) the principle that a confidence signal's value is set by its ranking quality rather than its magnitude is established (Rabanser & Papernot, 2025; Galil et al., 2023);

(vi) the aleatoric/epistemic decomposition we use is standard and has already been applied to TabPFN with the overlap-regime finding (Fortini et al., 2026; Hofman et al., 2024); and (vii) diagnose-then-target repair with specificity controls is established methodology (Gupta et al., 2026; Jha, 2026). Accordingly the title, abstract, contribution list, related work, discussion and conclusion were rewritten to claim a measurement and an assembly rather than a new statistical construction or a new problem, and the word "pointwise" was replaced by "per-input" because "pointwise coverage" is an existing term for a different object (Sluijterman et al., 2021). No reported number changed.

2. An earlier formulation used a contraction exponent β from $\log W \sim -\beta \log m$. This is ill-posed. Predictive width has an irreducible aleatoric floor, so $\beta \to 0$ for any model (a correctly specified reference measured -0.016). Replaced by the paired improvement ratio ψ.
3. Split conformal initially failed a criterion because it assigns a constant width and therefore cannot rank; constant baselines are now excluded as degenerate.
4. **A hypothesis of our own was falsified.** On 8 synthetic families ψ's discrimination showed no systematic relation to the uniform gap (Spearman $= -0.024$), and that screen was confounded (level $\times$ width-ratio collinear), so it could not separate the two explanations. The factorial design was built specifically to remove that confound.
5. **Level was briefly the main metric and then partly retracted.** The first pass passed $7/7$ criteria, but the factorial showed the single-reference version is driven by a reference-only knob. Reference averaging ($\bar{\ell}$) fixed this and raised discrimination from 0.561 to 0.759.
6. **The classification hypothesis was falsified twice, in two different directions.** The first proxy correlated negatively with the increment and degenerated at high overlap; the exact decomposition then showed the epistemic fraction was not significant ($+0.148$, $p = 0.56$) while the aleatoric level was ($+0.825$). The first pass reported the aleatoric level as the governing variable. That was still not the mechanism. The aleatoric level is collinear with entropy strength (-0.94) and partialling it out of entropy strength leaves -0.064, so we now name the scope rule by the mechanism and validate it out of sample. Recorded as a reminder that "which variable correlates most" was twice not the same as "what governs".

B Additional results

Experiments that support the claims of Section 5 but are not needed for its argument.

B.1 Coverage reconstruction and its cloud dependence

The synthetic cloud is the one stochastic ingredient in the comparison, and the prior sampler does not return the same cloud twice, so a single run is one draw. Across three independent clouds at a fixed data split, mean coverage moves over 0.039–0.074 (sd 0.018) while ψ is bit-identical (-0.287 every time) and the paired difference stays in $+0.327$–$+0.361$, never below $+0.29$. We therefore treat the paired gap, not the 0.081 point value, as the draw-invariant statistic and the difference between ψ and any coverage-style score as the conclusion. Table 8 gives the point values. Independent task-level numbers agree that coverage is near zero ($+0.021$ earlier, $+0.081$ here), and the paired comparison against the model's own width is $+0.037$ $[-0.030, +0.100]$, better on $13/22$ datasets—so ψ is not an improvement on a zero-cost baseline.

B.2 The mechanism, arm by arm

B.3 Main result, per dataset

B.4 The rate statistic against a zero-cost baseline

Table 4 shows that ψ's advantage over the model's own uncertainty is not statistically distinguishable from zero. Its partial correlation remains -0.322, but that removes only the linear component and does not establish a paired advantage.

mechanism	base error	relative gap	ψ	level
covered	0.5616	—	-0.009	-0.004
gapR	0.9807	$+74.6\,\%$	-0.196	$+0.826$
gapF	0.5881	$+4.7\,\%$ (not a gap)	$+0.003$	-0.035

Table 3: The controlled 1-D PFN, four arms differing only in the meta-training distribution, with ground truth from the *base* arm. The non-gap *gapF* ($+4.7\,\%$) is ranked above the real gap *gapR* by ψ and below it by the level statistic.

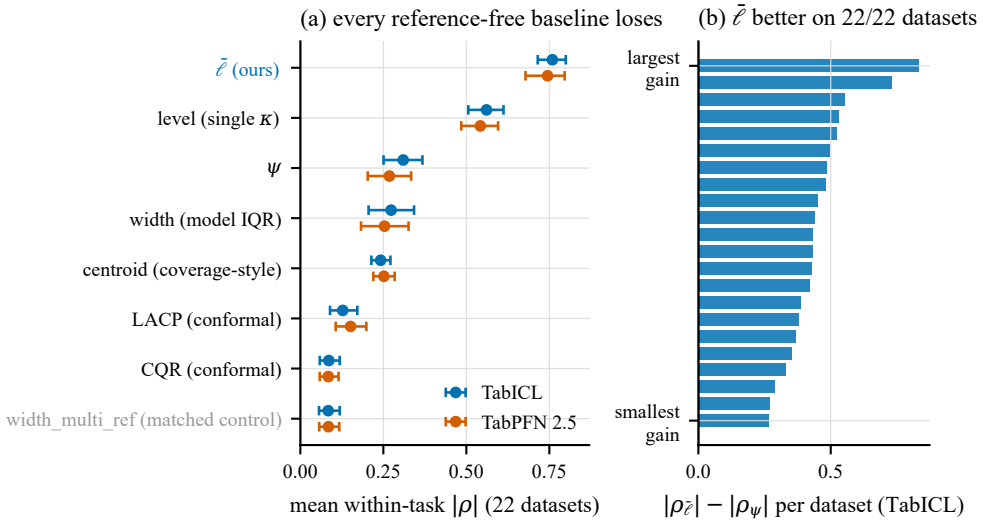

Figure 4: $\bar{\ell}$ dominates every reference-free baseline we tried. (a) Mean within-task $|\rho|$ with bootstrap CIs over the 22 datasets (per-dataset seed-averaged), for TabICL and for the TabPFN 2.5 replication. The matched control, which applies the identical averaging to the reference's own width, sits at the bottom next to CQR. (b) Per-dataset paired difference $|\rho_{\bar{\ell}}| - |\rho_{\psi}|$ on TabICL. $\bar{\ell}$ is better on $22/22$ datasets, mean $+0.449$ (a bar is one dataset).

statistic	mean $\lvert\rho\rvert$	paired vs *width*	datasets better
ψ	0.3099	$+0.0366\ [-0.030, +0.100]$	$13/22$
width (model's own IQR)	0.2733	—	—

Table 4: The paired improvement ratio against the model's own predictive width on the coverage-reconstruction suite. The advantage is not distinguishable from zero.

B.5 Reference-strength sensitivity

Sweeping a reference-only knob κ_0 while holding model and task fixed:

The RA-ratio is invariant to the reference knob while still responding monotonically to the actual gap knob σ (Figure 3), and ψ does not share that robustness: its sign flips in $28.6\,\%$ of cells across the κ_0 sweep, which is precisely the defect $\bar{\ell}$ was built to remove.

σ	level $\lvert\rho\rvert$ range over κ_0	$\bar{\ell}$ $\lvert\rho\rvert$
0.05	$0.205 \ldots 0.673$	0.566 (constant)
0.5	$0.101 \ldots 0.631$	0.366 (constant)
2	$0.073 \ldots 0.325$	0.130 (constant)

Table 5: The single-reference level statistic and the RA-ratio under a sweep of the reference's prior strength κ_0, at three noise levels. The level statistic spans a wide range within each row; the RA-ratio is constant in κ_0.

Multi-seed controlled factorial. The reduced factorial is TabPFN 2.5 over 14 named cells, originally with one seed per cell, which was the one robustness gap a reviewer could point at, since the σ-axis ordering rested on a single draw. It now carries three seeds per cell (42 rows) on the 9-point support grid. The 14 names are only 12 distinct configurations: *sigma0.5*, *d8_s8* and *nl0* share the parameters $(\sigma, d, d_{\text{signal}}, \text{nonlinear}) = (0.5, 8, 8, 0)$, and because the task sampler seeds from the configuration rather than from the cell name they produced bit-identical rows. Counting them once leaves the nonlinearity axis with two configurations, too few to rank, and the dimensionality axis with four distinct d values rather than five. Every axis statistic is therefore quoted on deduplicated cells.

With duplicates removed, only the noise axis is cleanly ordered (Table 9); ψ decreases monotonically across $\sigma = 0.05 \to 2$, with values +2.437, +1.863, +1.020, +0.509, +0.083, -0.125 (per-cell sd over seeds 0.16–0.53), a cross-cell range of 3.04 against a median within-cell CI width of 0.57. Two limitations follow. The dimensionality ordering is weak in both models ($p = 0.80$ / 0.60 at $n = 4$), so we report it as unordered, and individual ψ values are not stable across seed count: on TabPFN 2.5 the largest shift on the noise axis between the 1-seed and 3-seed runs is 0.45 and the sign of ψ at $\sigma = 1$ flips. This is why we read ψ only as a qualitative rate-side companion and never as a comparable cross-task quantity; the main table and its cross-model replication do not depend on that restriction.

B.6 Reference-family sensitivity

The reference's feature space is the other free choice, and it does move the statistic. Replacing the linear reference by a quadratic-polynomial reference and by random-Fourier-feature references, each first calibrated to the same effective degrees of freedom, moves the mean $|\rho|$ from 0.867 (linear) to 0.835 (quadratic) and 0.625–0.663 (flexible), so the RA-ratio retains most but not all of its strength once the reference is allowed to be nonlinear. The matched control stays at 0.080–0.110 in every family, and the linear reference remains the best of the four.

B.7 Causal validation: paired intervals and the reference sweep

Every paired bootstrap interval behind Section 5 excludes zero: on *gapR* training drops the RA-ratio's median level by 1.30–1.53 over the three task draws, the equal-compute control moves it by -0.01, and injecting the off-target region raises it by $+0.30$ to $+0.38$. This holds in every reference family we built, across both the published and the matched prior sweeps, so the sign of the causal claim is reference-invariant even though the magnitude of the rank discrimination is not. That bounds how much the reference-family sweep above should worry a user.

B.8 Non-PFN learners

Both replication models are PFN-family, so we re-ran the identical protocol on a histogram gradient-boosted learner with quantile loss and on ridge regression with a Gaussian predictive interval. The RA-ratio is the top statistic on both ($|\rho| = 0.733$ and 0.753, against 0.249 and 0.225 for the best reference-free baseline, the coverage-style centroid), it beats predictive width on $22/22$ datasets in each (paired $+0.541$ $[+0.483, +0.596]$ and $+0.617$ $[+0.548, +0.678]$), and its own matched control stays low (0.093 and 0.152). One further limitation: the boosted learner's own intervals are miscalibrated (central-80 % coverage 0.683), which weakens the width baseline it loses to, though the RA-ratio still wins against the stronger coverage-style centroid baseline ($+0.444$ $[+0.395, +0.492]$).

B.9 Scope and limitations

- **Scope of the statistic.** $\bar{\ell}$ is training-free and needs one probe grid plus six analytic reference fits per task, but it is not cross-task comparable: its level moves with reference strength (spread 1.478 against ψ's 0.592), so it is a within-task ranker and ψ should be reported alongside. The regression arm needs $m_{\max} \geq 192$, below which the model has not entered its contraction regime. The NIG reference is weak by construction, and at matched degrees of freedom a flexible reference costs 0.20–0.24 of mean $|\rho|$, so part of what $\bar{\ell}$ detects is "hard for a linear model" rather than the prior gap alone, and 0.76 is a lower bound on prior-relative adequacy.

hypothesis	LOCO AUC	nested AUC	row-sign acc.	cell-sign acc.
entropy strength (H_{gate})	$\mathbf{0.889}$	$\mathbf{0.888}$	$\mathbf{0.846}$	$17/18$
aleatoric level (H_{ale})	0.857	0.856	0.843	$\mathbf{18/18}$
epistemic share (H_{orig})	0.524	0.556	0.278	$1/18$

Table 6: Held-out sign prediction for the three hypotheses of Section 5.5. LOCO = learn on 17 of 18 cells and predict the held-out cell; nested = calibrate at one context size and predict the others ($m \in \{64, 192, 512\}$, nested support sets).

overlap	aleatoric level	$\lvert\rho_{\text{level}}\rvert$	$\lvert\rho_{\text{entropy}}\rvert$	increment over entropy
0.1	0.249	$+0.23 \cdots + 0.42$	$+0.005 \cdots + 0.031$	$+0.31 \cdots + 0.40$
0.6	0.229	$+0.36 \cdots + 0.42$	$+0.24 \cdots + 0.33$	$+0.08 \cdots + 0.17$
2.0	0.111	$+0.01 \cdots + 0.12$	$+0.87 \cdots + 0.88$	-0.73
3.5	0.030	$+0.56 \cdots + 0.94$	$+0.97 \cdots + 0.98$	$-0.03 \cdots - 0.38$

Table 7: The controlled class-overlap grid (6 overlap levels $\times$ 3 context sizes, 30 tasks per cell), four representative rows. Ranges are over context sizes. The increment of $\bar{\ell}$ over the model's own entropy is positive where entropy is weak and negative once entropy resolves per-input difficulty, the opposite sign to the hypothesis that the advantage grows with the epistemic share.

measure	per-input $\lvert\rho\rvert$	95 % CI	datasets negative
coverage (reconstruction)	$+0.081$	$[+0.028, +0.137]$	$5/21$
ψ	-0.287	$[-0.361, -0.217]$	$20/21$

Table 8: Per-input discrimination of the two existing measures on the coverage reconstruction suite, a faithful re-implementation of the coverage read-out with a synthetic cloud drawn from TabICL's exact pre-training prior, against the paired improvement ratio. Coverage is near zero; ψ is strongly negative, i.e. it ranks points in the wrong direction relative to the error.

- **Scope of the evidence.** Regression is 22 BeyondArena datasets (Erickson et al., 2025); classification is a scope boundary over 18 controlled cells and 540 tasks, validated out of sample but not a positive result about the RA-ratio itself. Primary results are TabICL, replicated on TabPFN 2.5 and, in the appendix, on two non-PFN learners; the non-PFN runs are single-seed and LightGBM and XGBoost are untested. We compared no k-nearest-neighbour, density or trajectory-based difficulty proxy (Lavi et al., 2026), so "is $\bar{\ell}$ merely a difficulty proxy?" is bounded rather than settled. Finally, the causal design follows established methodology (Gupta et al., 2026; Jha, 2026); equal compute here means a doubled covered prior at the same training budget, and the specificity claim is scoped to that operationalisation.

C SUPPLEMENTARY MATERIAL

The controlled-grid axis statistics of Section 5 are given in Table 9, the coverage reconstruction of Appendix B in Table 8, the classification grid in Table 7, and the held-out check of Section 5.5 in Table 6. The figure for the classification boundary of Section 5.5 follows.

axis	TabICL	TabPFN 2.5	verdict
noise ($\sigma \in \{0.05, \ldots, 2\}$, $n = 6$)	$\mathbf{-1.000}$	$\mathbf{-1.000}$	monotone
dimensionality ($d \in \{4, 16, 32, 64\}$, $n = 4$)	-0.200	-0.400	not significant
nonlinearity ($n = 2$)	—	—	not computable

Table 9: Spearman correlation between each controlled-grid axis and ψ, on deduplicated configurations (the 14 cell names are 12 configurations). Only the noise axis is cleanly ordered and exactly monotone in both models; the dimensionality ordering is not significant and the nonlinearity axis has too few configurations to rank.

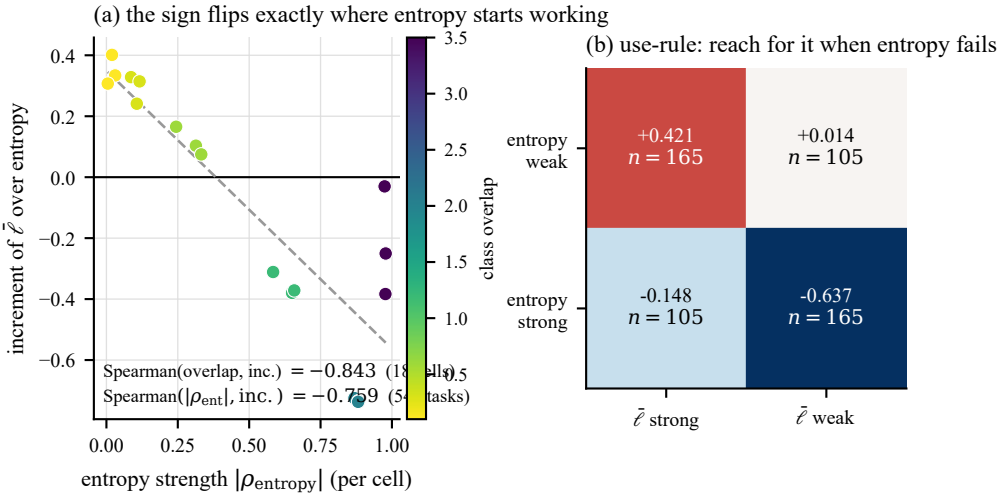

Figure 5: When the diagnostic helps, and when it does not. (a) The increment of $\bar{\ell}$ over the model's own entropy, one point per controlled cell, against entropy strength $|\rho_{\text{entropy}}|$ and coloured by class overlap. (b) The resulting use-rule on the two observables a practitioner has (median split over the 540 tasks).

experiment	TabICL	TabPFN 2.5
main per-input table	22 ds × 3 seeds, 9-point grid	22 ds, 3-point grid, 3 seeds
controlled factorial	14 cells, 3 seeds	14 cells, 3 seeds
causal / specificity	yes (own PFN)	—
classification boundary	18 cells × 30 tasks (540 tasks)	—
scope-rule out-of-sample validation	held-out cells + nested evidence levels	—
reference-strength sensitivity	154 cells	—
reference-family sensitivity	4 families × 22 ds, matched effective d.o.f.	—
non-PFN learners	22 ds, gradient boosting + ridge	—
coverage reconstruction	21 ds	—

Table 10: Replication status: the primary model carries every experiment, the second model the main table and the controlled factorial, both under the identical protocol.

D Replication status

Table 10 lists which experiment was run on which model.

On TabPFN 2.5 the ordering of all eight statistics agrees with TabICL at Spearman $+0.976$; $\bar{\ell}$ is the top statistic on both (0.759 / 0.745); ψ drops to 0.268; and the matched averaging control is essentially identical across models (0.0840 / 0.0847), which indicates that "averaging is not the explanation" is a model-independent property rather than a TabICL artefact. $\bar{\ell}$ beats ψ on $22/22$ datasets on the second model too ($+0.476$ $[+0.404, +0.549]$).

E Reproducibility

All experiments live in the *erc* package: shared core, grid and PFN utilities; the main-table pipeline (*run_level.py* with *analyze_level.py*); the reference-knob factorial (*run_separation.py*); the RA-ratio and its matched control (*run_refcal.py*); the coverage reconstruction (*run_coverage.py*); the causal and specificity experiments (*u1_causal.py* and *u1_eval.py*); the classification boundary and gate (*run_clf_decomp.py* and *validate_clf_gate.py*); the multi-seed controlled factorial (*run_g0_seeds.py*); and a correctness harness (*smoke_test.py*). Every collection script carries a verification step: it must reproduce the earlier ψ numbers ($\rho_\psi = -0.3025$, $\rho_{\text{width}} = +0.2733$, $\rho_{\text{center}} = -0.2419$, $\psi_{\text{partial}} = -0.3219$) before its new columns are trusted. Each run writes a directory under *results/erc/* holding per-row and per-cell tables, a per-dataset summary, a machine-readable verdict file and the printed analysis.

The classification arm is reproducible without re-running the model. The gate analysis is re-derived from the saved shards in analyze-only mode, with the shard count the collection used (six here), after which the out-of-sample validator scores criteria G1–G5 on the held-out cells.