INFORMATION-SET CONFOUNDING IN CONSISTENCY TESTS FOR TABULAR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work tests compatibility of univariate tabular foundation model (TFM) predictions by comparing marginals with autoregressive conditionals. The tested identities, however, combine posterior predictives based on different observed training columns. Because the law of total probability and the chain rule apply under a common information set, a nonzero cross-information gap need not imply a failure of Bayesian coherence across observed datasets. After making the conditioning explicit, we bound both reported gaps by ordinary posterior updates. We also construct an exact coherent Bayesian model for which both gaps equal $1/2$, while a predictor that ignores cross-target information attains zero gap at strictly worse joint risk. This reversal appears in a 60-task exact Bayesian benchmark, controlled experiments with TabICLv2 and TabPFNv2, and all ten classification datasets in the original study. On each dataset and for both models, a fixed mixture of the native conditional chains improves joint negative log likelihood over the zero-gap independent product; the combined pooled relative NLL gains are 10.62% and 10.17%. The queried heads can be incompatible when their different contexts are suppressed; our result shows why this does not establish that the predictor lacks a coherent Bayesian interpretation. A valid consistency test must hold the observed information fixed; practical evaluations should also include an information-blind control and held-out predictive scores.

1 INTRODUCTION

Prior-data fitted networks amortize Bayesian-style prediction across synthetic tasks and now provide strong tabular classifiers and regressors (Müller et al., 2022; Hollmann et al., 2025; Qu et al., 2026). This interpretation connects to the broader view of in-context learning as implicit Bayesian inference (Xie et al., 2022). Because their public heads are univariate, a common route to a multivariate prediction is autoregressive: predict one target, append its value as a feature, and predict the next. Whether the resulting conditionals are compatible is therefore an important question.

Klötergens et al. (2026) recently proposed two label-free checks. Marginalization consistency (C1) compares a direct marginal with a mixture of conditional predictions; factorization consistency (C2) compares the two autoregressive orders. They report nonzero gaps for every evaluated TFM and interpret them as evidence of probabilistic incoherence. The queried kernels can indeed be incompatible as marginals and conditionals of a single joint over the query targets. The further question is whether this incompatibility rules out a coherent Bayesian predictor when each kernel conditions on its own observed training data. The original study explicitly describes those different contexts; our objection concerns their interpretation as a necessary Bayesian coherence condition.

When predicting A, dropping training column B changes the context from paired observations (X_i, A_i, B_i) to (X_i, A_i). This is a change in evidence, not merely an algebraic marginalization at a test point. A coherent Bayesian posterior predictive normally changes when data are removed. C1 compares this reduced-data marginal with conditionals learned from the paired data and weights learned from a different reduced dataset. C2 likewise multiplies factors produced under different posterior information. Suppressing these information sets makes a valid probability identity appear applicable where it is not.

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.

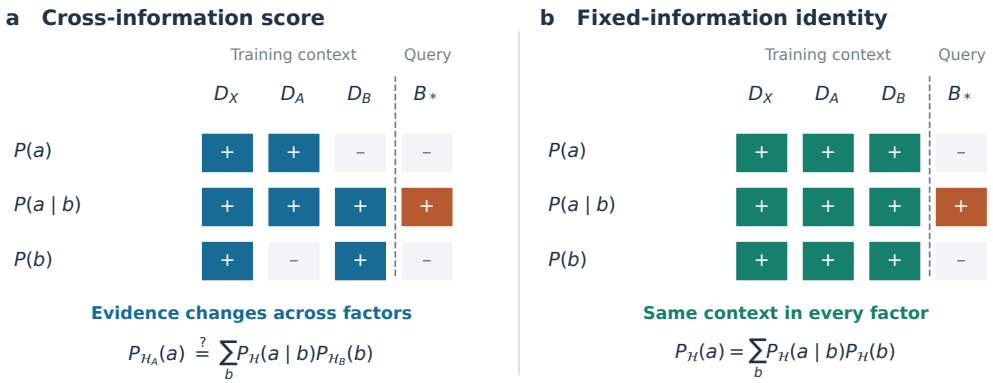

Figure 1: Training evidence and query observations are different inputs. Left: C1 combines predictions conditioned on different training columns. Right: the law of total probability holds with one fixed information set; only the query observation changes. Filled cells (+) denote observed inputs; pale cells (–) denote withheld inputs. The fixed query covariate x is omitted.

Our analysis restores the information sets and states the corresponding same-information identities (Figure 1). An information-update bound and a finite Bayesian counterexample show that a coherent predictor can have C1=C2=$1/2$. A complementary result shows that ignoring the other target forces both gaps to zero even when the discarded dependence is useful. Exact simulations and experiments with two frozen TFMs exhibit the same ranking reversal on real data. Together, these results establish the limits of the cross-information score while leaving the compatibility of current TFM heads open.

2 Where the information set enters

Let $D_X = (X_i)_{i=1}^n$, $D_A = (A_i)_{i=1}^n$, and $D_B = (B_i)_{i=1}^n$. Write

$$\mathcal{H}_A = \sigma(D_X, D_A), \qquad \mathcal{H}_B = \sigma(D_X, D_B), \qquad \mathcal{H} = \sigma(D_X, D_A, D_B).$$

For a query covariate x, let P_G denote the posterior predictive conditional on information G. The C1 identity tested for target A is, in this notation,

$$P_{\mathcal{H}_A}(a \mid x) \stackrel{?}{=} \sum_b P_{\mathcal{H}}(a \mid b, x) P_{\mathcal{H}_B}(b \mid x). \tag{1}$$

The three kernels in Eq. 1 use three information sets. In contrast, the law of total probability says, for any single G,

$$P_G(a \mid x) = \sum_b P_G(a \mid b, x) P_G(b \mid x). \tag{2}$$

Likewise, the tested C2 compares

$$P_{\mathcal{H}}(a \mid b, x) P_{\mathcal{H}_B}(b \mid x) \stackrel{?}{=} P_{\mathcal{H}}(b \mid a, x) P_{\mathcal{H}_A}(a \mid x), \tag{3}$$

whereas the chain rule equates the corresponding products only when all factors use the same G.

Proposition 1 (Information-update bound)**.** *Assume all displayed kernels are regular conditionals of one coherent Bayesian model, with a fixed measurable version of each conditional kernel on its entire state space. The claims are conditional on the realized training information and query covariate, wherever the posterior predictives are defined. Let $C_1^\times$ be the TV distance between the sides of Eq. 1, and $C_2^\times$ the TV distance between the two joints in Eq. 3. Then*

$$C_1^\times \le \mathrm{TV}(P^A_{\mathcal{H}_A}, P^A_{\mathcal{H}}) + \mathrm{TV}(P^B_{\mathcal{H}_B}, P^B_{\mathcal{H}}), \qquad C_2^\times \le \mathrm{TV}(P^A_{\mathcal{H}_A}, P^A_{\mathcal{H}}) + \mathrm{TV}(P^B_{\mathcal{H}_B}, P^B_{\mathcal{H}}).$$

Thus either gap can be nonzero solely because observing the other training column updates the posterior.

The proof, given in Section 4, inserts the full-information marginal or joint and then applies the triangle inequality and contraction of TV under a Markov kernel. The bound controls the gap by changes in posterior information; it does not establish that the bound is attained. The constructions below demonstrate nonzero gaps in coherent Bayesian models.

What can be tested. Conditional compatibility is well-defined after fixing G (Arnold & Press, 1989; Majid et al., 2025). A same-information test must query every marginal and conditional under that fixed evidence. Standard TFM APIs generally cannot retain observed D_B while treating query B_* as unobserved, so the desired full-information marginal may be unavailable. That interface limitation does not license replacing it by $P_{\mathcal{H}_A}$. In practice, cross-information gaps can be reported as sensitivity diagnostics, accompanied by an information-blind control and held-out joint scores. Without that qualification, they do not distinguish incompatibility of the extracted kernels from failure of a coherent model over training and query variables.

2.1 Three distinct questions

The distinction is between two different joint-distribution questions. A single joint over (A_*, B_*) need not realize all the extracted kernels after their context labels are suppressed. A larger coherent joint over training data and query variables can nevertheless realize each kernel conditional on its stated information set. Our counterexample establishes the latter possibility; it does not make the extracted kernels compatible under one G. Separately, held-out joint scores assess predictive utility rather than either form of compatibility.

The distinction matters if the score is used as an optimization target. When B carries evidence about a latent task parameter, an in-context learner should update its prediction of A after observing D_B; penalizing the update can favor a less responsive model. Forcing every head to ignore appended targets, for example, makes both identities hold by construction. Thus "lower is better" has no operational meaning without appropriate controls.

For an approximate TFM, an observed gap may contain both ordinary information sensitivity and genuine incompatibility, and our bound does not identify their proportions. Without a fixed-information reference or an explicit model of the update term, the scalar gap cannot reveal which mechanism produced it. Both components remain possible in the reported measurements.

3 Two exact reversals

3.1 A coherent Bayesian predictor fails both checks

Let $\theta \sim \mathrm{Bernoulli}(1/2)$. At a training covariate $x = 0$, draw $A_0 \sim \mathrm{Bernoulli}(1/2)$ independently and set $B_0 = \theta$. At query covariate $x = 1$, set $A_* = B_* = \theta$. This defines an ordinary joint distribution over all variables.

Under $\mathcal{H}_A = \sigma(A_0)$, $P(A_* = 1 \mid \mathcal{H}_A) = 1/2$. Under $\mathcal{H}_B = \sigma(B_0)$ and $\mathcal{H} = \sigma(A_0, B_0)$, θ is known and all query predictions are point masses at B_0. Substitution into Eqs. 1–3 gives

$$C_1^{\times} = C_2^{\times} = \tfrac{1}{2},$$

for every (A_0, B_0). The aligned identity in Eq. 2 has gap zero. Hence even exact Bayesian posterior predictives can give both cross-information gaps equal to one half in this binary construction; no claim of a universal maximal TV gap is needed.

The full-information conditional $P_{\mathcal{H}}(B_* \mid A_* = a)$ is arbitrary when $a \neq B_0$, a null event under $P_{\mathcal{H}}$. The C2 value is nevertheless version-invariant. The right chain assigns total mass $1/2$ to the slice $a \neq B_0$, which is disjoint from the left chain, and assigns mass $1/2$ rather than one to (B_0, B_0); any distribution of the null-slice mass therefore gives TV $1/2$. Section 4 records the calculation.

3.2 Full-support reversal

The reversal also occurs without null events. Let θ, A_0 be independent fair bits, and let B_0, A_*, B_* independently flip θ with probability $1/4$. All 32 joint atoms have positive mass, at least $1/256$. Conditional on $A_0 = B_0 = 0$, the full-information joint and the two chains are

$$J_F = J_{BA} = \begin{pmatrix} 7/16 & 3/16 \\ 3/16 & 3/16 \end{pmatrix}, \qquad J_{AB} = \begin{pmatrix} 7/20 & 3/20 \\ 1/4 & 1/4 \end{pmatrix}.$$

The marginal heads are $(1/2, 1/2)$ and $(5/8, 3/8)$. The A-side cross C1 is $1/8$, its B-side counterpart is $1/40$, and cross C2 is $1/8$; the symmetrized C1 used in experiments is therefore $3/40$. Same-information and blind gaps are zero. The fixed mixture has excess Brier risk $27/6400$, versus $1/32$ for

the blind product. Exact rational enumeration verifies these values for all four observations (A_0, B_0), with the joint tables permuted when $B_0 = 1$.

Symmetrization does not align the evidence. Averaging the directional C1 values gives $3/40$: the coherent model still has a positive gap. The full-support construction therefore separates the information mismatch from both target-order asymmetry and choices of conditionals on null events. A fixed chain mixture addresses predictive order selection, not this mismatch.

3.3 Ignoring dependence passes

Proposition 2 (Blindness degeneracy)**.** *For arbitrary marginal heads $q_A(a \mid x)$ and $q_B(b \mid x)$, define both conditional heads to ignore the appended target: $q_A(a \mid b, x) = q_A(a \mid x)$ and $q_B(b \mid a, x) = q_B(b \mid x)$. Then C1 and C2 are exactly zero, and both orders equal $q_A q_B$.*

The proof is direct normalization. In the deterministic witness, this zero-gap product has squared ℓ_2 joint risk $1/2$, while the full-information chain has risk zero. Therefore the score alone cannot rank held-out joint predictive utility: its optimum can discard precisely the dependence that motivates autoregressive use.

Order dependence remains possible. The two native chains may still disagree after their information sets are aligned, in which case the disagreement is a genuine defect. The counterexample shows that a positive cross-information gap is compatible with one Bayesian model over training and query variables. Whether current TFM heads are compatible after fixing their information set remains an empirical question.

4 Proofs

Proof of Proposition 1. Let $K(a \mid b) = P_{\mathcal{H}}(a \mid b, x)$. By the full-information law of total probability, $P^A_{\mathcal{H}} = KP^B_{\mathcal{H}}$. Hence

$$\mathrm{TV}(P^A_{\mathcal{H}_A}, KP^B_{\mathcal{H}_B}) \le \mathrm{TV}(P^A_{\mathcal{H}_A}, P^A_{\mathcal{H}}) + \mathrm{TV}(KP^B_{\mathcal{H}}, KP^B_{\mathcal{H}_B}) \le \mathrm{TV}(P^A_{\mathcal{H}_A}, P^A_{\mathcal{H}}) + \mathrm{TV}(P^B_{\mathcal{H}}, P^B_{\mathcal{H}_B}),$$

where the last step is TV contraction. For C2, insert the full-information joint $P^{AB}_{\mathcal{H}}$ between the two cross-information chains. Marginal replacement followed by a conditional kernel changes TV by at most the TV between the corresponding marginals. The triangle inequality gives the stated bound. □

Proof of Proposition 2. Marginalizing $q_A(a \mid x)q_B(b \mid x)$ over either variable returns the other normalized marginal. Both factorization orders are the same product. □

Null-event invariance in the witness. Let b_0 be observed. The left C2 chain is $\delta_{(b_0,b_0)}$. On the right, the supported conditional at $a = b_0$ must assign $B_* = b_0$, producing mass $1/2$ at (b_0, b_0). The remaining mass $1/2$ lies somewhere in the disjoint slice $a \neq b_0$, regardless of the chosen regular-conditional version there. The ℓ_1 difference is therefore $|1 - 1/2| + 1/2 = 1$, so TV is always $1/2$.

5 Experiments

We test the reversal with exact enumeration, controlled synthetic data, and real datasets. The supplement contains the protocols, seeds, decision thresholds, model configurations and raw predictions. No model was trained or tuned on the reported outcomes.

5.1 Exact finite-prior benchmark

We sample a prior of 16 binary linear classifiers in four dimensions. Two noisy labels A, B share the latent classifier; observing either target therefore updates the posterior. Across three fixed context/noise settings, 60 tasks, and 7,680 queries, we enumerate the exact posterior under $\mathcal{H}_A, \mathcal{H}_B, \mathcal{H}$. The mean coherent cross-information C1 is 0.0666 and 59/60 task means exceed 0.01. Same-information and

blind gaps are zero to numerical precision. The fixed 50:50 mixture of the coherent chains has lower excess joint Brier risk on 60/60 tasks and reduces mean excess risk by 81.95% over the blind product. Here excess risk is $\mathbb{E}_X\|q - p_{\mathrm{truth}}\|_2^2$; full expected Brier additionally includes $\mathbb{E}_X[1 - \|p_{\mathrm{truth}}\|_2^2]$.

5.2 Retrospective fixed-mixture reanalysis

The exact and synthetic analysis uses the same fixed 50:50 mixture as the real-data analysis. This estimand was selected after inspecting oracle comparisons that used population truth to choose the lower-risk chain. The resulting analysis is retrospective, not a prospective confirmation. We repeated all 32 synthetic model/task cells with unchanged models and seeds, saved all four probability heads, and independently reconstructed the reported scores. The original scalar risks and consistency gaps reproduced within the frozen tolerances. Oracle-selected risks are reported separately and are not estimates of the fixed-mixture risk.

Table 1: Full expected Brier, including irreducible uncertainty. Relative reductions in the main synthetic table concern excess risk, not these full scores.

Setting	Blind	Fixed mixture
Exact finite prior	.61817	.58953
TabICLv2 dependent	.58669	.54744
TabPFNv2 dependent	.58431	.54652

5.3 Controlled TFM evaluation

For each model, we generate four independent-target controls and 12 dependent tasks from disjoint seeds. Each task has 512 context rows, 256 query rows, and four Gaussian features. A follows a logistic model; B adds a coupling term $c(2A - 1)$, with $c = 0$ for controls and $c = 1.5$ otherwise. The population joint is available in closed form. We fit four frozen heads ($A|X$, $B|X$, $A|B, X$, $B|A, X$) for TabICLv2 and TabPFNv2.

All 12 dependent tasks have C1> 0.01 for both models. In every dependent task, the native heads have nonzero consistency gaps while their fixed mixture incurs less than half the blind product's excess risk.

5.4 OpenML experiments

We use six classification datasets and target pairs from Klötergens et al. (2026): Anneal, Credit, Phishing, Customer, Car, and Marketing (Feurer et al., 2021). Each dataset uses three deterministic 75:25 repeated holdouts, contexts capped at 768, and evaluation sets capped at 512. Rare classes are retained in context. Numeric missing values use context medians; categorical levels use context-fitted integer codes, and conditioning targets are one-hot encoded. We again fit four heads per model and compare the independent product with a fixed 50:50 mixture of the two autoregressive chains. The mixture avoids selecting an order from test outcomes. We evaluate joint NLL and Brier score on held-out paired labels.

For symmetry, our reported C1 is the mean of the A-side Eq. 1 TV and its $A \leftrightarrow B$ counterpart, averaged over query rows. We compute C2 as the row-mean TV between the two full categorical joint tables. The blind control uses the same fitted marginal heads as the native chains; only its conditional heads are defined to ignore the appended target. Thus the joint NLL comparison holds fitted marginal heads fixed. It does not separately identify contributions from modeled target dependence and additional training evidence available to conditional heads.

Both models improve dataset-mean NLL on 6/6 datasets. Across the 18 dataset/repeat blocks, the reversal occurs in 17 blocks for TabICLv2 and all 18 for TabPFNv2. Because repeats from one dataset are dependent, a post hoc hierarchical bootstrap resamples the six datasets and then the repeats within each selected dataset. The resulting 100,000-draw 95% intervals for macro-average normalized NLL gain are [1.56%,16.48%] and [1.33%,15.83%]. Joint Brier score improves on every dataset for both models, and the largest numerical C1/C2 of the blind construction is 1.18×10^{-16}. The pooled

Table 2: Controlled dependent-target results. Risk is excess joint Brier risk; gain compares the fixed 50:50 chain mixture with the blind product.

Model	Dep. C1	Control C1	Mixture risk	Blind risk	Gain
TabICLv2	.0202	.0037	.00677	.04603	85.28%
TabPFNv2	.0303	.0081	.00646	.04424	85.40%

Table 3: Primary real-data results. C1 is the mean native cross-information gap; ΔNLL is the percentage improvement of the fixed native mixture over the exactly zero-gap blind product. Values average three repeats.

	TabICLv2		TabPFNv2	
Dataset	C1/C2	ΔNLL	C1/C2	ΔNLL
Anneal	.0086/.0137	5.50%	.0132/.0223	3.77%
Credit	.0293/.0469	2.35%	.0440/.0697	1.99%
Phishing	.0349/.0521	17.69%	.0464/.0672	17.41%
Customer	.0117/.0211	0.44%	.0182/.0376	0.48%
Car	.0507/.0756	24.78%	.0552/.1000	23.80%
Marketing	.0225/.0426	0.30%	.0276/.0508	0.48%
Pooled	–	8.17%	–	7.79%

entries in Table 3 use a different estimand: the percentage reduction between mean native and blind NLL over all blocks.

The comparison exhibits a zero-gap construction dominated in predictive risk. It does not determine how much of the native TFM gap comes from information sensitivity or from incompatibility under a common information set. That distinction is established by the exact Bayesian examples, not inferred from a real-data NLL gain.

5.5 Prospective breadth extension

After the six-dataset primary test passed, we froze a separate extension to the other four classification datasets in the source study: MIC, Wine, Nursery, and Diamonds. We declared success before seeing their outcomes as at least three dataset wins and positive pooled gain. Both models win on all four (Table 4). Together, the primary and extension results give 10/10 NLL and 10/10 Brier wins for each model. A post-extension hierarchical bootstrap over all ten dataset clusters gives 100,000-draw macro-average normalized-gain intervals of [2.26%,17.87%] for TabICLv2 and [1.94%,17.13%] for TabPFNv2. The reversal occurs in 29/30 blocks for each model.

Figure 2 retains all repeats and makes the concentration of large gains visible. Nursery, whose targets are strongly dependent, accounts for most of the extension aggregate. Gains on MIC, Wine, and Diamonds are positive but below 2%, so the effect is not uniformly large.

6 Preprocessing and dataset influence sensitivity

We recompute the saved joint distributions from their four predictive heads, verify disjoint context and test row identifiers, and independently reconstruct all per-example losses. Positive Δ below denotes blind-product NLL minus the fixed 50:50 native-chain NLL. Across the ten selected datasets, dataset-cluster bootstrap intervals (10,000 resamples) are descriptive; they do not establish a population-wide guarantee.

Model	Macro Δ	Dataset-cluster 95% CI	Min. leave-one-out Δ
tabicl	0.1392	[0.0171, 0.3116]	0.0695
tabpfn	0.1373	[0.0174, 0.3066]	0.0688

For myocardial infarction complications (MIC), we additionally select features using context rows only (at most 2% missing), fit imputations or categorical mappings on context rows, and retain

Table 4: Prospectively frozen breadth extension. Entries are native cross-information C1 and joint NLL gain over the blind product, averaged over three repeats.

	TabICLv2		TabPFNv2	
Dataset	C1/C2	ΔNLL	C1/C2	ΔNLL
MIC	.0154/.0279	0.57%	.0171/.0323	0.97%
Wine	.0158/.0299	1.50%	.0255/.0493	1.79%
Nursery	.0553/.0873	38.93%	.0771/.1185	37.26%
Diamonds	.0428/.0808	0.67%	.0686/.1138	0.78%
Extension pooled	–	13.65%	–	13.04%
All ten pooled	–	10.62%	–	10.17%

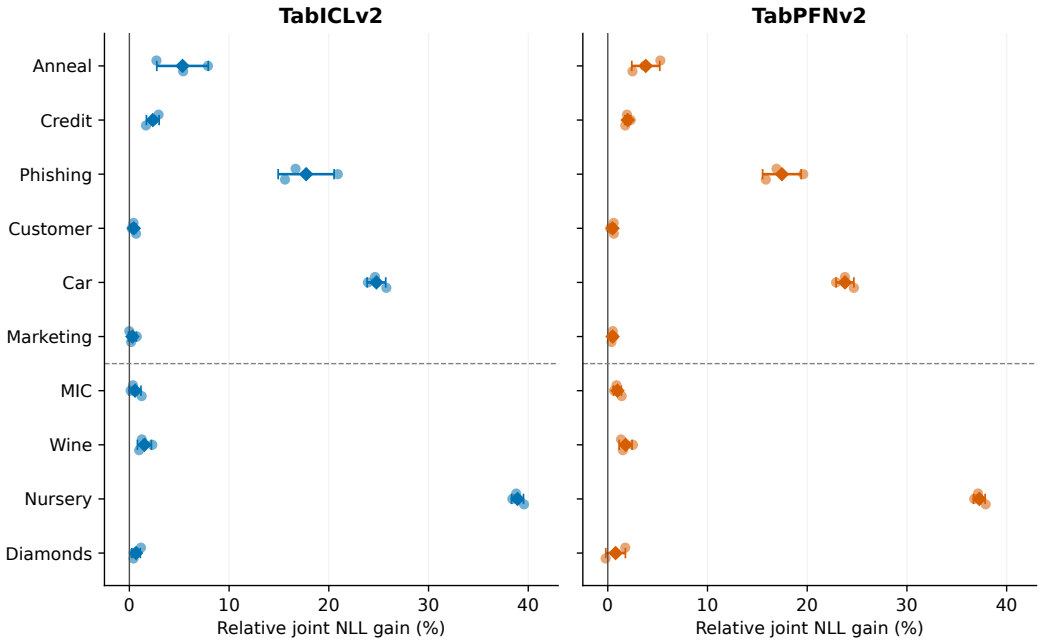

Figure 2: Joint NLL gains across all ten datasets. Small points are the three repeated holdouts; diamonds and bars show their mean and sample standard deviation, not a confidence interval. Each gain is $100(L_{\text{blind}} - L_{\text{native}})/L_{\text{blind}}$ for that repeat. The dashed line separates the six primary datasets from the four later prospectively frozen extension datasets. These per-dataset mean relative gains differ from the pooled ratio reported in the tables.

all 414 test rows in each of three repeats. Only missing targets are removed before splitting. The blind-minus-native NLL differences are tabicl: 0.0075, 0.0036, 0.0077, respectively, tabpfn: 0.0102, 0.0021, 0.0044. All six differences remain positive. This control changes the eligible row universe as well as preprocessing: only 111, 109, and 96 test identifiers overlap the old split in the three repeats. Shared-row differences are archived for transparency and are not a pure estimate of a feature-selection intervention. Neither these sensitivities nor zero blind-product gaps imply that a deployed TFM is coherent under every common information set.

7 Numerical evaluation and dataset construction

Both pinned checkpoints use four estimators and FP32 inference; their exact configurations are supplied with the evaluation code. The exact witness uses rational arithmetic. API probability vectors are converted to float64 and renormalized before probability algebra; the unnormalized float32 outputs can miss unit row sums by about 10^{-7}.

Table 5: OpenML classification tasks and evaluation sizes. Each entry is repeated three times.

Dataset	OpenML ID	Rows	Features	Classes (A, B)
Anneal	46906	898	37	(5,8)
Credit	46918	1,000	19	(2,4)
Phishing	46963	1,353	8	(3,3)
Customer	46938	1,723	12	(2,6)
Car	40975	1,728	5	(4,3)
Marketing	46940	2,240	24	(2,8)
MIC	46980	1,596	51	(8,2)
Wine	40691/40498	6,497	11	(8,2)
Nursery	26	12,960	7	(5,3)
Diamonds	42225	53,940	8	(5,7)

Following the source study, MIC drops columns with more than 2% missing values before removing incomplete rows. The column filter is computed on the full dataset before the repeated holdout split, so test-set missingness can affect which columns are retained. Value imputation and categorical fitting use context data only. Wine aligns the red and white OpenML schemas by documented column order and adds source color.

8 Related work

Compatibility of a family of conditionals with one joint is a classical problem and remains active in masked and discriminative modeling (Arnold & Press, 1989; Majid et al., 2025). These formulations hold the conditioning specification fixed. Any-order autoregressive models can disagree across orders even when trained by one network (Shih et al., 2022). Such fixed-context disagreement is distinct from changing the observed training data between factors.

Martingale-based accounts formulate predictive coherence as a sequential condition on the information observed so far (Fong et al., 2023; Falck et al., 2024). They average over future observations rather than data that have already been observed and then selectively deleted. Conditional neural processes map context sets to predictive distributions (Garnelo et al., 2018); Young (2026) measures how those predictions change when a point is added to the context. There, the context update is the object being measured, rather than an algebraic identity imposed across the two contexts.

The closest empirical comparison is Klötergens et al. (2026). That study explicitly specifies different context datasets; our criticism concerns the interpretation of the cross-context identity as a necessary coherence condition, not an omitted context label. On overlapping models and datasets, we obtain nonzero classification gaps of a similar scale. Once the information sets are explicit, we interpret these measurements as cross-context sensitivity; a nonzero value alone does not prove that no joint exists.

9 Limitations and implications

Our real-data study covers classification, two open checkpoints, and all ten classification datasets in the source study. It does not reassess continuous-target discretization or every TFM. The first six datasets form the pre-specified primary test; the remaining four belong to a later prospective extension. We use integer encoding instead of each model's native categorical interface, which can change absolute predictive quality. Every compared construction nevertheless shares the same encoded marginals, and conditioning targets are one-hot encoded. Because the real-data joint is unknown, held-out predictive scores measure utility rather than global coherence.

A fully aligned diagnostic needs an interface that distinguishes training evidence from query-time observations. The following contract specifies the required queries. When the interface cannot express them, cross-context sensitivity, the blind control, and held-out joint scores remain informative, but fixed-information compatibility is untested.

Table 6: Required queries for one full-information marginalization check. All rows retain the same paired training columns. The table specifies an evaluation contract, not a claim that current APIs implement it.

Requested distribution	Revealed query targets	Training evidence
$P_{\mathcal{H}}(A_* \mid x)$	Neither target	D_X, D_A, D_B
$P_{\mathcal{H}}(B_* \mid x)$	Neither target	D_X, D_A, D_B
$P_{\mathcal{H}}(A_* \mid B_* = b, x)$	$B_* = b$	D_X, D_A, D_B
$P_{\mathcal{H}}(B_* \mid A_* = a, x)$	$A_* = a$	D_X, D_A, D_B

9.1 A FIXED-INFORMATION EVALUATION CONTRACT

An evaluator must specify two objects before constructing either identity: the training evidence G and the set of query outcomes revealed to the prediction interface. These are distinct masks. Holding the training rows and their observed target columns constant fixes G; revealing $B_* = b$ changes the query conditioning event. Deleting D_B instead changes G and returns to the cross-information comparison. This remains true when the same model weights, random seed and query covariates are reused.

For finite target spaces, the first check enumerates b, multiplies each conditional vector by its same-information weight, and compares the sum with the direct marginal. The order check enumerates both target coordinates and compares two normalized joint tables. Every probability vector should be checked before these operations: normalization errors are implementation errors rather than evidence about compatibility. Unsupported query events additionally require a stated conditional version; the full-support construction above shows that the substantive reversal does not depend on that technical issue.

A native joint head provides an implementation control: deriving all four objects from its table makes the identities hold by construction. Passing then verifies the probability algebra, not predictive quality. Joint NLL or Brier scores answer the complementary utility question.

If the interface cannot express one row of the table, the corresponding aligned check is unavailable. Substituting a reduced-context marginal would change the conditioning specification rather than complete that check.

9.2 EXACT RESULTS ACROSS TRAINING REALIZATIONS

Table 7: Every training realization has the same exact result.

A_0	B_0	Cross C1	Cross C2	Aligned C1	Blind risk
0	0	1/2	1/2	0	1/2
0	1	1/2	1/2	0	1/2
1	0	1/2	1/2	0	1/2
1	1	1/2	1/2	0	1/2

The four rows enumerate training realizations, not independent Monte Carlo replicates. Their agreement follows from the construction's symmetry. Thus the example needs no confidence interval: uncertainty about a finite sample is separate from the exact identity being tested.

10 CONCLUSION

The conditioning sigma-field is part of a conditional distribution. Dropping an observed training column changes it, so probability identities cannot be applied across the resulting contexts without qualification. The exact examples show that cross-information C1/C2 can be positive for a coherent Bayesian predictor. The empirical comparisons establish a separate limitation: a zero-gap blind product can have worse joint risk than a nonzero-gap mixture. The actionable criterion is an information contract: every prediction in a consistency identity must condition on the same evidence. This criterion repairs the interpretation of the test; it does not certify the evaluated models as coherent.

Reproducibility Statement

The supplement contains the exact rational witness, frozen protocols, OpenML IDs and preprocessing, complete inference runners, model configurations, per-repeat predictions and statistical analysis code. The main-text reproducibility details give model settings and dataset sizes. No private training data or proprietary model API is used.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Barry C. Arnold and S. James Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989. doi: 10.1080/01621459.1989.10478750. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478750.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12784–12805, 2024. URL https://proceedings.mlr.press/v235/falck24a.html.

Matthias Feurer, Jan N. van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Müller, Joaquin Vanschoren, and Frank Hutter. OpenML-Python: An extensible python api for openml. *Journal of Machine Learning Research*, 22(100):1–5, 2021. URL https://www.jmlr.org/papers/v22/19-920.html.

Edwin Fong, Chris Holmes, and Stephen G. Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1357–1391, 2023. doi: 10.1093/jrsssb/qkad005. URL https://academic.oup.com/jrsssb/article/85/5/1357/7597700.

Marta Garnelo, Dan Rosenbaum, Chris Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1704–1713, 2018. URL https://proceedings.mlr.press/v80/garnelo18a.html.

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

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

Mahbod Majid, Rattana Pukdee, Vishwajeet Agrawal, Burak Varıcı, and Pradeep Kumar Ravikumar. On the consistent recovery of joint distributions from conditionals. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4708–4716, 2025. URL https://proceedings.mlr.press/v258/majid25a.html.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63874.

Andy Shih, Dorsa Sadigh, and Stefano Ermon. Training and inference on any-order autoregressive models the right way. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2762–2775, 2022. doi: 10.52202/068431-0200.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Robin Young. On the conditioning consistency gap in conditional neural processes. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=rLJ5Hm5vbG.