Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Estimand-Relative Validity and Context-Driven Instability of Shapley Explanations for Tabular Foundation Models

Emin Talip Demirkiran

etd@eskisehir.edu.tr

Eskisehir Technical University

Research Article

Keywords: Shapley values, tabular foundation models, removal operators, in-context learning, feature attribution, explanation stability

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11112632/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Estimand-Relative Validity and Context-Driven Instability of Shapley Explanations for Tabular Foundation Models

Author: Emin Talip Demirkiran
Affiliation: Department of Computer Engineering, Eskisehir Technical University, Iki Eylul Campus, 26555 Eskisehir, Türkiye
Email: etd@eskisehir.edu.tr
ORCID: https://orcid.org/0000-0001-6063-8178

Abstract

Shapley explanations for tabular foundation models depend on a removal operator that determines how predictions are evaluated when features are absent. Although different operators induce different cooperative games, their effects on attribution validity and stability in in-context tabular models remain insufficiently understood. We compare five removal operators for TabPFN v2 and TabICL v2 using exact Shapley enumeration, controlled synthetic data with known or high-precision reference attributions, real-world tabular datasets, and a preregistered factorial design. Operator performance was strongly estimand-dependent. Mean imputation and missing-value encoding most closely approximated interventional attributions, whereas recontextualisation and conditional estimation through in-context learning yielded the lowest errors for the conditional estimand; these rankings also varied with feature dependence. Across nuisance configurations, attribution variability was dominated by the sampled in-context examples, while column order, label coding, and model seed contributed little. A planted-null-feature control further showed that both models frequently changed their predictions in response to features constructed to contain no label information. The analysis is limited to two binary-classification foundation models and exact enumeration in low-dimensional settings, so the observed rankings should not be extrapolated to larger feature spaces or other model families. These findings show that Shapley explanations for tabular foundation models are not interpretable without specifying both the removal operator and the intended estimand, and that context sensitivity should be reported as part of explanation reliability.

Keywords: Shapley values; tabular foundation models; removal operators; in-context learning; feature attribution; explanation stability

1. Introduction

Tabular foundation models perform in-context learning: a context of labelled rows and a query yield a prediction without parameter updates (Qu et al. 2025; Hollmann et al. 2025). Explanation methods have already been adapted to these models (Rundel et al. 2024; Sena and Azevedo 2026; Fonseca and Stoyanovich

2026), and Shapley-based methods are among the most widely used approaches to post-hoc feature attribution (Lundberg and Lee 2017).

However, a Shapley value is defined only relative to a *value function*, and for a supervised model there is no canonical way to evaluate a coalition in which some features are absent (Shapley 1953; Covert, Lundberg, and Lee 2021). The "explaining by removing" framework (Covert, Lundberg, and Lee 2021) makes this explicit: marginal imputation, baseline imputation, missing-value encoding, retraining and conditional estimation are legitimate removal operators that define different games. For conventional predictive models, this choice is already a recognised source of disagreement (Olsen et al. 2024). For TFMs, the model consumes an entire context at inference time, so removing a feature may mean removing it from the query, from the context, or from both. Because these models predict without conventional parameter retraining, retraining-style operators can instead be implemented through recontextualisation (Olsen and Christensen 2026; Qu et al. 2025; Hollmann et al. 2025).

These observations motivate two pre-registered research questions. First, does removal-operator choice materially change attribution correctness across dependence and data regimes? We address this question with synthetic generators whose known target functions permit exact or high-precision computation of interventional and conditional Shapley reference values (Liu et al. 2021; Olsen et al. 2024). Second, how stable are TFM attributions under changes that practitioners may regard as nuisance variation? In-context prediction is sensitive to the sampled context (Abdullah 2026), and TFMs are not guaranteed to be invariant to column order (Biloš et al. 2026) or class relabelling (Arbel, Salinas, and Hutter 2025). If their attributions inherit this sensitivity, explanation multiplicity (Hwang et al. 2026) becomes a routine reliability concern rather than an exceptional case.

This study makes four contributions:

1. A preregistered, ground-truth-referenced comparison of five removal operators on two TFMs across 12 dependence regimes, with exact Shapley enumeration rather than sampling approximations.
2. Evidence that operator correctness is estimand-relative: the best operator for the interventional game is close to the worst for the conditional game, and vice versa.
3. A fully crossed variance decomposition of attribution instability over four in-context nuisance factors, showing that the context draw dominates while the other three contribute negligible variance under the evaluated design.
4. A study-wide negative result for a standard null-feature invariance control, together with independently reproduced analyses and complete unit-level provenance.

2. Related work

Prior-data fitted networks introduced in-context prediction for tabular tasks by training transformers to approximate posterior prediction over synthetic task distributions (Müller et al. 2022). TabPFN subsequently established this approach on small tabular datasets (Hollmann et al. 2025), while TabICL extended it to larger contexts (Qu et al. 2025) and later model generations expanded scale, openness and model capacity (Grinsztajn et al. 2025, 2026; Qu et al. 2026). These systems do not automatically satisfy the invariances that tabular prediction may require. EquiTabPFN formalised target-permutation equivariance (Arbel, Salinas, and Hutter 2025), and later mechanistic analyses documented sensitivity to feature order and architectural choices (Biloš et al. 2026). This model dependence motivates treating the fitted context and representation convention as part of the object being explained.

Several lines of work now address TFM interpretability, but they target different explanatory objects. Post-hoc methods for TabPFN include recontextualisation-based Shapley values and data valuation (Rundel et al. 2024), conditional Shapley estimation using a TFM as the conditional-expectation engine (Olsen and Christensen 2026), and experimental-design methods that reduce the cost of coalition evaluation (Rundel et al. 2026). Amortised approaches learn to predict attributions directly, either through a specialised explainer or a Shapley-regression objective (Fonseca and Stoyanovich 2026; Alkhatib et al. 2025; Jethani et al. 2022), while ShapPFN incorporates attribution prediction into a small prior-fitted model (Sena and Azevedo 2026). Other approaches expose additive shape functions, kernel or nearest-neighbour weights, and distilled interactions rather than feature-removal attributions (Mueller et al. 2024; Miftachov, Charron, and Valentin 2026; Koshil, Feurer, and Eggensperger 2025; Jia et al. 2026). These methods improve access to explanations, but most evaluate agreement with another explainer or fidelity to model outputs; they do not compare removal semantics against both interventional and conditional data-generating targets.

The meaning of a Shapley attribution depends on the value function used when features are absent (Shapley 1953; Lundberg and Lee 2017). The removal-based framework separates the removal mechanism from the subsequent summarisation and allocation steps (Covert, Lundberg, and Lee 2021). Within that framework, marginal or baseline replacement and conditional expectation answer different questions. The distinction has been framed as fidelity to the fitted model versus fidelity to the data distribution (Chen et al. 2020), as a causal choice (Janzing, Minorics, and Blöbaum 2020), and as a dependence-aware estimation problem (Aas, Jullum, and Løland 2021; Frye et al. 2021; Olsen et al. 2024). Baseline selection can also change local attributions even when the allocation rule is fixed (Haug et al. 2021). The present study holds the summarisation and allocation rule constant through exact Shapley enumeration and varies the removal operator, allowing disagreement to be interpreted as a change of estimand rather than an approximation artefact.

Evaluation research likewise shows that agreement with a single reference explainer is insufficient. Randomisation tests can reveal explanations that are insensitive to the fitted model (Adebayo et al. 2018), while OpenXAI, XAI-Units and synthetic benchmark suites combine controlled tasks with multiple evaluation criteria (Agarwal et al. 2022; Lee et al. 2025; Liu et al. 2021; Barr et al. 2020). Broader evaluation frameworks distinguish application-grounded, human-grounded and functionally grounded evidence (Doshi-Velez and Kim 2017), and empirical studies show that plausible explainers can disagree substantially on the same prediction (Krishna et al. 2024). Exact-game tools such as shapiq make attribution error measurable for low-dimensional games (Muschalik et al. 2024), whereas recent work on SHAP multiplicity separates model-induced from explainer-induced variation (Hwang et al. 2026). Our design builds on these principles by using exact enumeration, explicit ground-truth games, preregistered materiality criteria and negative controls.

TFM-specific instability has so far been studied mainly at the prediction or representation level. Context resampling changes TabPFN predictive performance (Abdullah 2026), TabICL's positional encoding does not guarantee column-permutation invariance (Qu et al. 2025), and target-label equivariance requires explicit architectural treatment (Arbel, Salinas, and Hutter 2025). Mechanistic studies have probed layer contributions, internal representations, inference dynamics, attention-head function, encoded causal structure and individual neurons (Rezaei Balef, Koshil, and Eggensperger 2025, 2026; Aviral Gupta, Sethi, and Kumar 2026; Atharva Gupta et al. 2026; Swelam et al. 2025; Knauer and Rodner 2025; Biloš et al. 2026), but they do not decompose the stability of feature attributions under crossed context, column-order, label-coding and model-seed perturbations. Explanation-multiplicity work includes TabPFN but varies model and explainer seeds rather than these TFM-specific factors (Hwang et al. 2026). The remaining gap is therefore not whether TFMs can be explained, but whether their Shapley explanations are valid for a stated estimand and stable under the contextual choices that define an in-context prediction.

3. Method

3.1 Setup

For a model with fitted context D and query $\mathbf{x}$, let $p_D(\mathbf{x}) = \Pr(y = 1 \mid \mathbf{x}, D)$. The explained scalar is the positive-class logit after clipping the predicted probability at $\varepsilon = 10^{-4}$:

$$s_D(\mathbf{x}) = \operatorname{logit}[\min\{1 - \varepsilon, \max(\varepsilon, p_D(\mathbf{x}))\}] \,.$$

For a coalition $S \subseteq N$ of retained features, each removal operator o defines a value function $v_o(S; \mathbf{x}, D)$ (Shapley 1953; Covert, Lundberg, and Lee 2021). Unless the context itself is restricted, retained query coordinates remain fixed at $\mathbf{x}_S$ and only the complementary coordinates $\bar{S}$ are altered:

- **O1 — marginal imputation**: replace $\mathbf{x}_{\bar{S}}$ with the corresponding values from each of $R = 16$ context rows and average the resulting s_D values.
- **O2 — baseline imputation**: replace $\mathbf{x}_{\bar{S}}$ with context feature means.
- **O3 — missing-value encoding**: replace $\mathbf{x}_{\bar{S}}$ with NaN while leaving the context unchanged.
- **O4 — recontextualisation**: present the pretrained model with the context and query restricted to S; this changes the inference context but does not update model parameters. The empty coalition uses the context label prior (Rundel et al. 2024; Olsen and Christensen 2026).
- **O5 — conditional via ICL**: fit a TabPFN v2 regressor to $(\mathbf{X}_{\text{pool},S}, s_D(\mathbf{X}_{\text{pool}}))$ and evaluate it at $\mathbf{x}_S$, using a pool disjoint from the context and queries (Olsen and Christensen 2026). The same conditional-estimation engine is used for both classifiers to avoid coupling the operator to the model being explained.

For feature j, the attribution is the exact Shapley value over all 2^d coalitions (Shapley 1953):

$$\phi_j(v_o) = \sum_{\substack{S \subseteq N \\ j \notin S}} \frac{|S|!\,(d - |S| - 1)!}{d!} \left[v_o(S \cup \{j\}) - v_o(S) \right].$$

We evaluated the TabPFN v2 and TabICL v2 classifiers using their pinned pretrained checkpoints (Hollmann et al. 2025; Qu et al. 2025). Shapley efficiency, $\sum_{j=1}^{d} \phi_j = v_o(N) - v_o(\varnothing)$, was checked per unit; the maximum residual across the study was 2.01×10^{-6} against a 10^{-5} tolerance.

3.2 Ground truth

Synthetic data come from a Gaussian copula with equicorrelated $\mathbf{\Sigma}_\rho$ and either Gaussian or standardised log-normal marginals, with a known additive-plus-pairwise target function g^*. Synthetic benchmarks make it possible to compare explanations against an explicitly defined data-generating function rather than another explainer (Liu et al. 2021; Agarwal et al. 2022). The two reference games are

$$v_{\text{int}}(S; \mathbf{x}) = \mathbb{E}_{\mathbf{X}_{\bar{S}}}[g^*(\mathbf{x}_S, \mathbf{X}_{\bar{S}})], \qquad v_{\text{cond}}(S; \mathbf{x}) = \mathbb{E}[g^*(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S].$$

Both have closed forms under Gaussian marginals; for skewed marginals we use Monte Carlo estimates with 200,000 interventional and 100,000 conditional draws and report their standard errors. The games correspond to distinct removal semantics (Covert, Lundberg, and Lee 2021; Olsen et al. 2024). They are reference attributions of the data-generating score g^*, not of the fitted TFM itself. Accordingly, the analysis evaluates how closely an operator applied to a TFM recovers a specified data-generating estimand; it does not treat

disagreement with g^* as proof that the attribution is internally unfaithful to the TFM. For generator replicate g, regime r, model M, operator o and target game τ, agreement with the reference is measured by the mean normalised Euclidean attribution error

$$E_{g,r,M,o,\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{\left\| \phi_i^o - \phi_i^{*,\tau} \right\|_2}{\left\| \phi_i^{*,\tau} \right\|_2}.$$

3.3 Design

The study comprised 632 units: Arm A (synthetic $d = 8$, 12 regimes $\times$ 5 replicates $\times$ 2 models = 120) for H1; Arm B ($d = 6$, the full $C(4) \times P(3) \times L(2) \times E(2) = 48$ nuisance configurations per model and condition = 384) for H2; a $d = 4$ scale arm (40); a dose-response arm (24); a label-randomisation control (24); and a real-data arm over four public OpenML datasets with planted null features (40) (Vanschoren et al. 2014). Each unit used a stratified context of 512 observations, 32 disjoint query observations (16 in Arm B), and a disjoint O5 pool of 256 observations. Features were standardised using context statistics before model evaluation. The real-data arm used banknote-authentication (OpenML 1462, version 1), wilt (OpenML 40983, version 2), phoneme (OpenML 1489, version 1) and mozilla4 (OpenML 1046, version 1).

Every random draw derives from a per-unit seed by a fixed rule, so any unit is reproducible in isolation. Arm B queries, O5 pools and background indices are drawn at the *condition* level, so the nuisance factors are crossed rather than confounded with the data.

3.4 Preregistration and analysis

Before execution, we specified the protocol, statistical models, exclusion rules, materiality margins ($\delta_E = 0.10$ for error differences and $\tau_\pi = 0.10$ for variance shares), and multiplicity procedures. Holm correction was applied within each confirmatory family of eight tests (Holm 1979), and Benjamini–Hochberg correction within each secondary family (Benjamini and Hochberg 1995). H1 uses the linear mixed-model specification

$$E \sim o \times \rho + I + G + (1 \mid r{:}g),$$

where o denotes the removal operator, ρ feature dependence, I the interaction regime, G the marginal distribution and $(1 \mid r{:}g)$ a random intercept for generator replicate nested within regime. Likelihood-ratio tests and percentile cluster-bootstrap intervals over regime–replicate units provide inference (Efron and Tibshirani 1993).

For H2, each attribution vector is standardised as $\mathbf{z} = \boldsymbol{\phi}/\|\boldsymbol{\phi}\|_2$; vectors with norm below 10^{-8} are excluded and counted. A crossed random-effects ANOVA

is fitted over the 48 configurations. The model estimates main-effect components for context draw (C), column order (P), label coding (L) and model seed (E); their interactions and unmodelled variation remain in the residual. Negative method-of-moments component estimates are truncated at zero before shares are formed. For nuisance factor $F \in \{C, P, L, E\}$, its variance share is

$$\pi_F = \frac{\sigma_F^2}{\sigma_C^2 + \sigma_P^2 + \sigma_L^2 + \sigma_E^2 + \sigma_{\text{res}}^2}.$$

Shares are summarised by the median over instances and features, with percentile cluster-bootstrap intervals over instances (Efron and Tibshirani 1993). Mixed models were fitted with statsmodels 0.15.0 (Seabold and Perktold 2010); numerical analysis used NumPy 2.4.6, SciPy 1.17.1 and pandas 3.0.5.

Outcome access was withheld until all planned units had completed, separating the prespecified analysis choices from the observed results.

4. Results

All 632 planned units completed; none was invalid, failed, retried or reseeded, and every prespecified model-by-condition comparison was complete.

4.1 Operator choice materially changes correctness, and the best operator depends on the estimand

Table 1 reports estimated marginal mean normalised attribution error, averaged over the 12 regimes. Lower values indicate closer agreement with the target attribution.

Table 1. Estimated marginal mean normalised attribution error by model, reference game and removal operator.

Model	Reference game	O1	O2	O3	O4	O5
TabICL v2	Interventional	0.577	0.498	0.498	0.662	0.678
TabPFN v2	Interventional	0.597	0.531	0.543	0.643	0.680
TabICL v2	Conditional	0.892	0.936	0.936	0.442	0.408
TabPFN v2	Conditional	0.988	1.015	0.945	0.453	0.442

Note. Estimates are averaged over 60 regime–replicate units per model and reference game. O1, marginal imputation; O2, mean imputation; O3, missing-value encoding; O4, recontextualisation; O5, conditional estimation through in-context learning. O2 and O3 are tied for TabICL v2 under the interventional game.

The operator main effect is evident in every cell (Holm-adjusted $p \leq 1.8 \times 10^{-33}$). The materiality criterion is met for the conditional target: the largest pairwise

difference has a 95% CI lower bound of 0.486 (TabICL) and 0.502 (TabPFN), exceeding $\delta_E = 0.10$. For the interventional target, the corresponding lower bounds are 0.057 and 0.026 and therefore do not meet the prespecified margin. The confirmatory materiality conclusion consequently rests on the conditional cells rather than on a pooled contrast.

The orderings are close to reversed between targets, consistent with the distinction between interventional and conditional removal games (Covert, Lundberg, and Lee 2021; Olsen et al. 2024): imputation-style operators approximate the interventional game, while operators that re-fit or re-estimate in context approximate the conditional game. The ranking also changes with dependence: in TabICL · conditional, $\rho = 0.0$ gives O2, O3, O4, O5, O1 and $\rho = 0.9$ gives O5, O4, O1, O2, O3 (Figure 1).

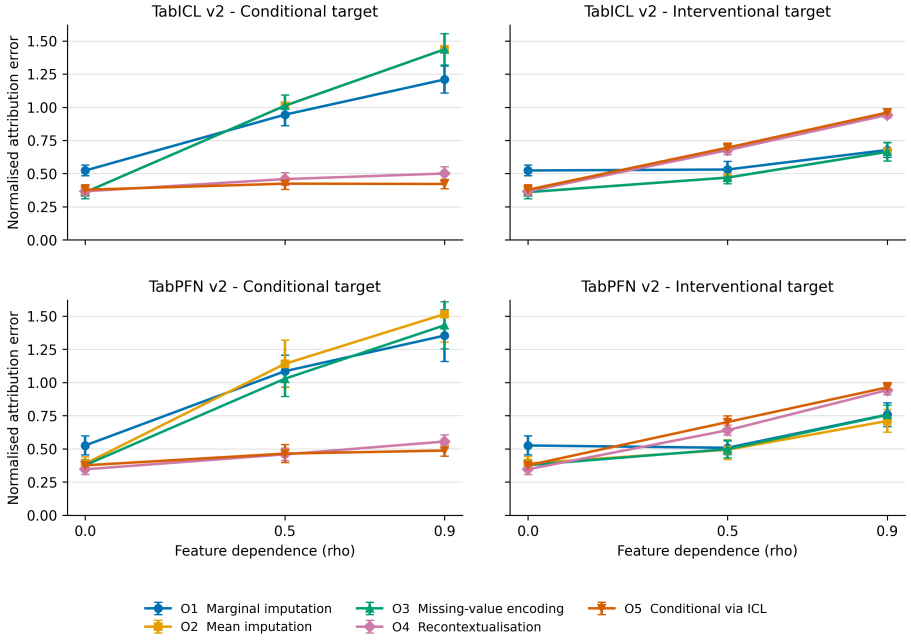


Figure 1. Normalised attribution error as a function of feature dependence for five removal operators. Results are shown separately for TabICL v2 and TabPFN v2 and for conditional and interventional reference games. Points show mean unit-level error; error bars show descriptive 95% normal-approximation intervals. Confirmatory inference uses the cluster-bootstrap intervals reported in the text and supplementary tables. Lower values indicate closer agreement with the target attribution.

For TabICL v2, O2 and O3 are numerically identical in all 316 units ($\max\|\boldsymbol{\phi}^{O2} - \boldsymbol{\phi}^{O3}\|_\infty = 0$). Thus, these operators are operationally indistinguishable for the evaluated TabICL checkpoint and inputs; this result does not establish a general identity between its NaN handling and mean imputation. For TabPFN, by

contrast, the maximum absolute difference is 12.80.

4.2 Attribution instability is dominated by the context draw

Figure 2 summarises the factor-specific variance shares. Table 2 gives the corresponding model-level medians over the 48 crossed configurations for the synthetic conditions, pooled over O1 and O4.

Table 2. Median share of attribution variance associated with each nuisance factor.

Model	π_C (context)	π_P (column order)	π_L (label coding)	π_E (seed)
TabICL v2	0.978 [0.974, 0.982]	0.0002	0.0027	0.0000
TabPFN v2	0.908 [0.900, 0.914]	0.0009	0.0003	0.0000

Note. Bracketed values are 95% cluster-bootstrap intervals for the context share. Full intervals and operator-specific estimates are reported in the Supplementary Material. C, context draw; P, column order; L, label coding; E, model seed.

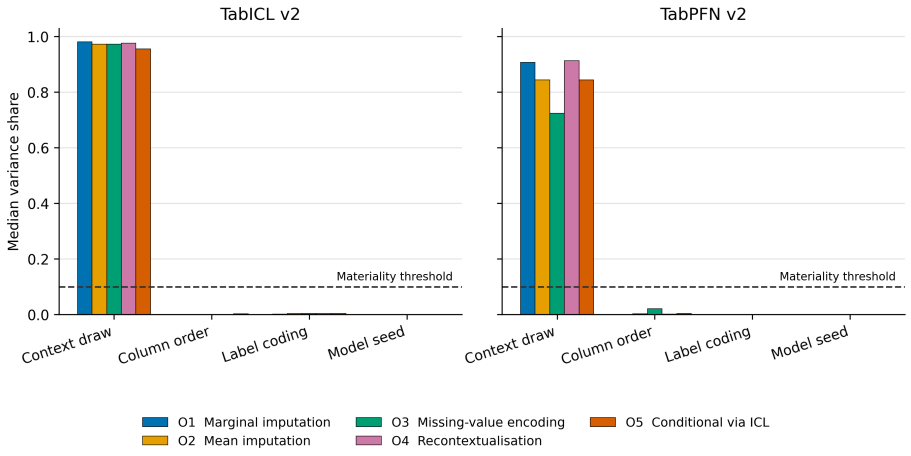


Figure 2. Median share of attribution variance associated with each nuisance factor across the fully crossed design. Results are shown by model and removal operator. The dashed line marks the prespecified materiality threshold of 0.10. Context draw dominates in both models; confirmatory cluster-bootstrap intervals and operator contrasts are reported in Table 2 and the Supplementary Material.

We report this by its interval rather than by its p-value, which sits at the resolution floor of the bootstrap. The result is not an artefact of explainer

sampling: σ_C (0.107, 0.118) exceeds the explainer-noise standard deviation σ_U (0.053, 0.067) in both models. This is a variance decomposition — the variability is *associated with* the context draw, not shown to be caused by it.

The prespecified operator-dependence hypothesis was rejected because all eight confidence intervals for the O1−O4 contrasts lie within the ±0.05 equivalence region. This evidence is nevertheless only informative for context sensitivity: three of the four factor shares are near-degenerate (π_E is exactly zero; π_P and π_L are approximately 10^{-3}). For the context factor, the TabPFN interval [−0.025, +0.018] spans most of the equivalence region, whereas the remaining contrasts are concentrated close to zero.

4.3 A null-feature control that fails

Each synthetic and real unit carries planted features with no effect on the label, and the protocol verifies model invariance by permuting such a feature jointly in context and query, requiring max $|\Delta \text{ logit}| \leq 0.05$.

Of 842 invariance checks, 836 failed (99.3%; median 0.411, 90th percentile 0.885, maximum 2.040), precluding evaluation of the null-feature attribution-mass metric. This is itself informative: permuting a feature constructed to carry no label information nevertheless changed the TFM prediction. The other two controls behaved as expected: dose-response was strongly monotone (Spearman 0.86–0.91), and label randomisation disrupted the attribution pattern (rank correlations −0.18 to +0.15).

4.4 Secondary observations

Per-instance instability correlates positively with per-instance error in every cell (Spearman 0.50–0.83). Operator cost spans more than an order of magnitude — O2/O3 ≈ 2–8 s per unit, O1 ≈ 31–114 s, O5 ≈ 53 s, O4 ≈ 94–103 s — so the operators aligned with the conditional estimand are the expensive ones; the full runtime comparison is shown in Supplementary Figure S1. On real data, operators disagree substantially (normalised divergence 0.59–1.79), with the largest disagreements between the imputation and recontextualisation families.

4.5 Robustness

All six preregistered robustness analyses are reported. Restricting H1 to units with below-median model approximation error leaves every verdict unchanged; the material pairwise lower bounds remain 0.440 for TabICL and 0.545 for TabPFN. Thus, the principal conditional-estimand result is not attributable solely to the poorest-fitting TFM regimes, although this stratification does not eliminate model misfit. Direction-consistency checks based on rank correlation, top-three agreement and sign agreement match the primary ordering in all 12 cells (Spearman 0.6–1.0). Extending the H2 pooling to all five operators leaves π_C at 0.972 and 0.855. The $d = 4$ versus $d = 8$ comparison is non-monotone: interventional errors are lower at $d = 4$ in all ten cells, whereas conditional errors

are higher at $d = 4$ for O1/O2/O3 and lower for O4/O5. This sign reversal limits extrapolation across dimensionality.

5. Discussion

The main implication is that the removal operator is part of the scientific question, not a technical setting chosen after the fact. Shapley values allocate the value function supplied to them, and different removal operators supply different value functions (Covert, Lundberg, and Lee 2021; Olsen et al. 2024). The reversal between interventional and conditional rankings therefore does not identify one universally correct implementation. It shows that an attribution can be judged only after its estimand has been stated. Recontextualisation is especially important in this respect: removing a column from the context changes the information available to the in-context learner, whereas imputing a query leaves the fitted context intact. Although both procedures return Shapley values, they explain different counterfactual prediction problems.

This distinction changes how an operator should be selected in practice. An interventional analysis, in which absent features are replaced without preserving their observed dependence structure, is most closely served here by mean imputation or missing-value encoding. A conditional analysis, in which coalitions are evaluated over feature combinations compatible with the observed distribution, is better served by recontextualisation or conditional estimation through in-context learning. These recommendations are semantic rather than universal performance rankings: interventional attributions should not be described as conditionally valid merely because they are inexpensive, and conditional attributions should not be preferred when the intended question is explicitly interventional. Computational cost enters only after that choice. In our setting, the operators best aligned with the conditional estimand were substantially more expensive, so faster explanations may require accepting a different estimand rather than a less accurate estimate of the same one.

The stability results also change what constitutes the object being explained. For an in-context predictor, a prediction is determined jointly by the pretrained parameters, the labelled context and the query. An attribution computed for one context therefore describes that realised prediction environment, not an invariant property of the pretrained model. The dominance of context composition extends earlier evidence of context-sensitive TFM predictions (Abdullah 2026) to their feature attributions and provides a TFM-specific source of explanation multiplicity (Hwang et al. 2026). The positive association between attribution error and instability further suggests that instability is diagnostically relevant, although the present design does not establish that one causes the other. A deployment report should therefore describe the context-selection policy and, when contexts can vary, summarise attribution variation over plausible context draws rather than presenting a single explanation without uncertainty.

By contrast, the small variance shares for column order, label coding and model

seed should not be interpreted as general invariance guarantees. They show only that these factors contributed little under the evaluated models, configurations and perturbation ranges. Prior work establishes that TFM architectures can violate column- or target-permutation invariance (Arbel, Salinas, and Hutter 2025; Biloš et al. 2026), and the current experiment was not designed to prove equivalence for every factor. The absence of a detectable difference between O1 and O4 is similarly narrow: three factor shares were near-degenerate, and the informative context contrast retained appreciable uncertainty. The defensible conclusion is that changing the operator did not remove context sensitivity, not that all operators have equivalent stability.

The failed null-feature control clarifies another boundary between predictor behaviour and explainer behaviour. Null-feature attribution metrics assume that the predictor is invariant to the planted feature before assessing whether an explainer assigns negligible mass to it. That assumption did not hold here. Consequently, non-zero attribution to the planted feature could reflect a genuine change in the TFM prediction induced by its context-dependent processing, rather than an attribution error. Randomisation and benchmark-based sanity checks remain valuable (Adebayo et al. 2018; Agarwal et al. 2022), but model invariance must be verified before their scores are interpreted as properties of the explainer. For TFMs, this diagnostic should precede null-feature faithfulness evaluation rather than appear only after an anomalous score is observed.

These implications are bounded by the evaluated setting. The non-monotone comparison between four and eight features rules out a simple extrapolation with dimension, and the findings do not establish operator rankings for multiclass tasks, larger feature spaces, ensembles or other TFM families. They instead establish two requirements that remain valid within the evidence: the estimand and removal operator must be reported together, and explanation reliability must account for context variation. Extending exact or controlled-reference evaluation to higher dimensions, alternative context-selection policies and ensembled inference is the next step needed to determine how broadly the observed rankings and instability mechanisms persist.

6. Limitations

Attribution error combines removal-operator error with model approximation error. The prespecified misfit-stratified analysis reduces, but cannot causally separate, these components. Moreover, normalised error measures agreement with a reference game of the data-generating function rather than faithfulness to the model's own predictive computation, a distinction recognised in explanation-evaluation frameworks (Agarwal et al. 2022; Liu et al. 2021). O5 also uses a TabPFN regressor as the common conditional-expectation estimator for both classifiers. This controls operator implementation across models, but the estimator's inductive bias may affect O5's apparent advantage and should be tested with alternative conditional estimators. The H2 design contains only four context levels, three column orders and two levels each for label coding

and model seed. Its main-effects decomposition assigns factor interactions to the residual; consequently, the reported shares describe relative variation under these perturbations rather than population-wide invariance or causal effects. The empirical scope is limited to two pretrained TFMs, binary classification, exact Shapley attribution for $d \leq 8$, and CPU inference with one estimator configuration. Because the dimensionality comparison changes direction across estimands and operators, extrapolation to larger feature spaces is unwarranted. The real-data analyses lack ground-truth attributions and therefore contribute evidence only about predictive faithfulness, operator divergence and stability. Secondary analyses are descriptive, with Benjamini–Hochberg control applied within each family.

Three departures from the prespecified analysis warrant explicit qualification. First, the C2 materiality criterion was evaluated using a regime-pooled interval rather than at a single value of ρ. Second, the unused rejection rule for C4 was implemented with pooled rather than operator-specific medians; this branch was not activated and therefore did not affect the reported conclusion. Third, pairwise differences were computed for the label-randomisation control but not for the dose-response control. These departures constrain the corresponding interpretations but do not alter the primary estimates.

7. Reproducibility

The supplementary package provides the protocol, study matrix, records for all 632 experimental units, software and environment identifiers, and source data for the reported results. Each unit can be reproduced from its recorded seed, and repeated execution at a fixed thread count produced bit-identical output. The confirmatory analysis specifications and decision thresholds were fixed before outcome access. Subsequent corrections affected rendering and audited implementation details but did not change the confirmatory analysis path or numerical results. Independent regeneration from the unit records reproduced all 30 result tables and every reported statistical conclusion. The code repository and raw per-unit attribution records are not publicly available at the time of submission.

8. Conclusion

The removal operator is a first-class methodological choice for explaining tabular foundation models, not an implementation detail. Its effect on agreement with reference attributions can be large, is estimand-relative, and interacts with feature dependence. Attribution instability, meanwhile, is dominated by the in-context example set and was equivalent between O1 and O4 within the prespecified margin in this design. We recommend that TFM attribution results name the removal operator and target game, report context-draw sensitivity alongside the attribution, and verify model invariance before applying any null-feature faithfulness metric.

Statements and Declarations

Funding

No funding was received for conducting this study.

Competing interests

The author has no relevant financial or non-financial interests to disclose.

Author contributions

Emin Talip Demirkiran: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing, and Project administration. Emin Talip Demirkiran is the sole author of this work and takes responsibility for the integrity of the study, interpretation and verification of the results, preparation and critical revision of the manuscript, and approval of the final version.

Data availability

Synthetic conditions are reproducible from the locked specifications and master seed. The real-data conditions use public OpenML datasets 1462, 40983, 1489, and 1046. Derived analysis tables, figure data, claim objects, and the 632-unit study index are supplied with the supplementary materials. Raw per-unit attribution records are not publicly released.

Code availability

The analysis code, protocol, and analysis constants were frozen and checksummed. The code repository is not publicly available at the time of submission.

Ethics approval

Not applicable. The study used synthetic data and public benchmark datasets and involved no human participants, animals, or personal data.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Use of generative AI

Generative AI tools were used during manuscript preparation for literature exploration, manuscript organization and restructuring, language refinement,

and reviewer-style critique. The author independently checked the cited sources, archived data, analysis code, numerical results, methodological decisions, and interpretations and takes full responsibility for the final manuscript. Generative AI was not treated as an author and was not used to create the primary study data or to replace execution of the reported analyses.

References

Aas, Kjersti, Martin Jullum, and Anders Løland. 2021. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *Artificial Intelligence* 298: 103502. https://doi.org/10.1016/j.artint.2021.103502.

Abdullah, Mohammed. 2026. "Understanding Context Sampling in TabPFN on Small Tabular Datasets." *arXiv Preprint arXiv:2607.26628*. https://doi.org/10.48550/arXiv.2607.26628.

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. "Sanity Checks for Saliency Maps." In *Advances in Neural Information Processing Systems*. Vol. 31. https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Agarwal, Chirag, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. "OpenXAI: Towards a Transparent Evaluation of Model Explanations." In *Advances in Neural Information Processing Systems*, 35:15784–99. https://doi.org/10.52202/068431-1148.

Alkhatib, Amr, Roman Bresson, Henrik Boström, and Michalis Vazirgiannis. 2025. "Prediction via Shapley Value Regression." In *ICML 2025 (PMLR 267)*, 267:1056–1101. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v267/alkhatib25a.html.

Arbel, Michael, David Salinas, and Frank Hutter. 2025. "EquiTabPFN: A Target-Permutation Equivariant Prior Fitted Network." In *Advances in Neural Information Processing Systems*. Vol. 38.

Barr, Brian, Ke Xu, Claudio Silva, Enrico Bertini, Robert Reilly, C. Bayan Bruss, and Jason D. Wittenbach. 2020. "Towards Ground Truth Explainability on Tabular Data." In *ICML Workshop on Human Interpretability in Machine Learning*. https://arxiv.org/abs/2007.10532.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Biloš, Marin, James T. Wilson, Anderson Schneider, and Yuriy Nevmyvaka. 2026. "A Mechanistic Study of Tabular Foundation Models." *arXiv Preprint*. https://doi.org/10.48550/arXiv.2605.21288.

Chen, Hugh, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. 2020. "True to the Model or True to the Data?" *arXiv Preprint arXiv:2006.16234*. https://arxiv.org/abs/2006.16234.

Covert, Ian C., Scott Lundberg, and Su-In Lee. 2021. "Explaining by Removing:

A Unified Framework for Model Explanation." *Journal of Machine Learning Research* 22 (209): 1–90. https://jmlr.org/papers/v22/20-1316.html.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." *arXiv Preprint arXiv:1702.08608*. https://arxiv.org/abs/1702.08608.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman; Hall/CRC. https://doi.org/10.1201/9780429246593.

Fonseca, Joao, and Julia Stoyanovich. 2026. "ExplainerPFN: Towards Tabular Foundation Models for Model-Free Zero-Shot Feature Importance Estimations." *arXiv Preprint arXiv:2601.23068*. https://doi.org/10.48550/arXiv.2601.23068.

Frye, Christopher, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. 2021. "Shapley Explainability on the Data Manifold." In *International Conference on Learning Representations*. https://arxiv.org/abs/2006.01272.

Grinsztajn, Léo, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, et al. 2025. "TabPFN-2.5: Advancing the State of the Art in Tabular Foundation Models." *arXiv Preprint arXiv:2511.08667*. https://doi.org/10.48550/arXiv.2511.08667.

Grinsztajn, Léo, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, et al. 2026. "TabPFN-3: Technical Report." *arXiv Preprint arXiv:2605.13986*. https://arxiv.org/abs/2605.13986.

Gupta, Atharva, Dhruv Kumar, Murari Mandal, and Saurabh Deshpande. 2026. "Where Computation Lives Inside TabPFN: Causal Localisation of Attention Head Function." In *ICML Workshop on Foundation Models for Structured Data*. https://arxiv.org/abs/2606.12917.

Gupta, Aviral, Armaan Sethi, and Dhruv Kumar. 2026. "TabPFN Through the Looking Glass: An Interpretability Study of TabPFN and Its Internal Representations." *arXiv Preprint arXiv:2601.08181*. https://arxiv.org/abs/2601.08181.

Haug, Johannes, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. 2021. "On Baselines for Local Feature Attributions." In *AAAI Workshop on Explainable Agency in Artificial Intelligence*. https://arxiv.org/abs/2101.00905.

Hollmann, Noah, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. "Accurate Predictions on Small Data with a Tabular Foundation Model." *Nature* 637: 319–26. https://doi.org/10.1038/s41586-024-08328-6.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70. https://www.jstor.org/stable/4615733.

Hwang, Hyunseung, Seungeun Lee, Lucas Rosenblatt, Julia Stoyanovich, and Steven Euijong Whang. 2026. "Explanation Multiplicity in SHAP: Characterization and Assessment." *arXiv Preprint arXiv:2601.12654*. https://doi.org/10.48550/arXiv.2601.12654.

Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. 2020. "Feature Relevance Quantification in Explainable AI: A Causal Problem." In *Pro-*

ceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 108:2907–16. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v108/janzing20a.html.

Jethani, Neil, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. 2022. "FastSHAP: Real-Time Shapley Value Estimation." In *International Conference on Learning Representations*. https://arxiv.org/abs/2107.07436.

Jia, Jingyun, Chandan Singh, Rich Caruana, and Ben Lengerich. 2026. "Selecting Feature Interactions for Generalized Additive Models by Distilling Foundation Models." *arXiv Preprint arXiv:2604.13332*. https://arxiv.org/abs/2604.13332.

Knauer, Ricardo, and Erik Rodner. 2025. "In Search of Grandmother Cells: Tracing Interpretable Neurons in Tabular Representations." In *EurIPS Workshop on AI for Tabular Data*. https://arxiv.org/abs/2601.03657.

Koshil, Mykhailo, Matthias Feurer, and Katharina Eggensperger. 2025. "In-Context Learning of Soft Nearest Neighbor Classifiers for Intelligible Tabular Machine Learning." In *Proceedings of the 4th Table Representation Learning Workshop*, 182–91. https://doi.org/10.18653/v1/2025.trl-1.15.

Krishna, Satyapriya, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2024. "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective." *Transactions on Machine Learning Research*. https://arxiv.org/abs/2202.01602.

Lee, Jun Rui, Sadegh Emami, Michael David Hollins, Timothy C. H. Wong, Carlos Ignacio Villalobos Sánchez, Francesca Toni, Dekai Zhang, and Adam Dejl. 2025. "XAI-Units: Benchmarking Explainability Methods with Unit Tests." In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2892–2905. https://doi.org/10.1145/3715275.3732186.

Liu, Yang, Sujay Khandagale, Colin White, and Willie Neiswanger. 2021. "Synthetic Benchmarks for Scientific Research in Explainable Machine Learning." In *NeurIPS 2021 (Datasets and Benchmarks Track)*. https://doi.org/10.48550/arXiv.2106.12543.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*. Vol. 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Miftachov, Ratmir, Bruno Charron, and Simon Valentin. 2026. "Interpretable Tabular Foundation Models via in-Context Kernel Regression." *arXiv Preprint arXiv:2602.02162*. https://arxiv.org/abs/2602.02162.

Mueller, Andreas, Julien Siems, Harsha Nori, David Salinas, Arber Zela, Rich Caruana, and Frank Hutter. 2024. "GAMformer: Bridging Tabular Foundation Models and Interpretable Machine Learning." *arXiv Preprint arXiv:2410.04560*. https://arxiv.org/abs/2410.04560.

Müller, Samuel, Noah Hollmann, Sebastian Pineda, Josif Grabocka, and Frank Hutter. 2022. "Transformers Can Do Bayesian Inference." In *International Conference on Learning Representations*. https://arxiv.org/abs/2112.10510.

Muschalik, Maximilian, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. 2024. "Shapiq: Shapley Interac-

tions for Machine Learning." In *Advances in Neural Information Processing Systems*, 37:130324–57. https://doi.org/10.52202/079017-4141.

Olsen, Lars Henry Berge, and Dennis Christensen. 2026. "Computing Conditional Shapley Values Using Tabular Foundation Models." *arXiv Preprint arXiv:2602.09489*. https://doi.org/10.48550/arXiv.2602.09489.

Olsen, Lars Henry Berge, Ingrid Kristine Glad, Martin Jullum, and Kjersti Aas. 2024. "A Comparative Study of Methods for Estimating Model-Agnostic Shapley Value Explanations." *Data Mining and Knowledge Discovery* 38 (4): 1782–1829. https://doi.org/10.1007/s10618-024-01016-z.

Qu, Jingang, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. 2026. "TabICLv2: A Better, Faster, Scalable, and Open Tabular Foundation Model." *arXiv Preprint arXiv:2602.11139*. https://arxiv.org/abs/2602.11139.

Qu, Jingang, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. 2025. "TabICL: A Tabular Foundation Model for in-Context Learning on Large Data." In *Proceedings of the 42nd International Conference on Machine Learning*, 267:50817–47. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v267/qu25d.html.

Rezaei Balef, Amir, Mykhailo Koshil, and Katharina Eggensperger. 2025. "Towards Understanding Layer Contributions in Tabular in-Context Learning Models." In *EurIPS Workshop on AI for Tabular Data*. https://arxiv.org/abs/2511.15432.

———. 2026. "Is One Layer Enough? Understanding Inference Dynamics in Tabular Foundation Models." *arXiv Preprint arXiv:2605.06510*. https://doi.org/10.48550/arXiv.2605.06510.

Rundel, David, Fabian Fumagalli, Maximilian Muschalik, Bernd Bischl, and Matthias Feurer. 2026. "ShaplEIG: Bayesian Experimental Design for Shapley Value Estimation." *arXiv Preprint arXiv:2606.02247*. https://doi.org/10.48550/arXiv.2606.02247.

Rundel, David, Julius Kobialka, Constantin von Crailsheim, Matthias Feurer, Thomas Nagler, and David Rügamer. 2024. "Interpretable Machine Learning for TabPFN." In *Explainable Artificial Intelligence*, 465–76. Communications in Computer and Information Science. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63797-1_23.

Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *Proceedings of the 9th Python in Science Conference*, 92–96. https://doi.org/10.25080/Majora-92bf1922-011.

Sena, Luan Borges Teodoro Reis, and Francisco Galuppo Azevedo. 2026. "Real-Time Explanations for Tabular Foundation Models." In *2nd Foundation Models for Science Workshop at ICLR 2026*. https://doi.org/10.48550/arXiv.2603.29946.

Shapley, Lloyd S. 1953. "A Value for n-Person Games." In *Contributions to the Theory of Games II*, edited by Harold W. Kuhn and Albert W. Tucker, 307–18. Princeton University Press. https://doi.org/10.1515/9781400881970-018.

Swelam, Omar, Lennart Purucker, Jake Robertson, Hanne Raum, Joschka Boedecker, and Frank Hutter. 2025. "Does TabPFN Understand Causal Structures?" In *EurIPS Workshop on AI for Tabular Data*. https://arxiv.or

g/abs/2511.07236.

Vanschoren, Joaquin, Jan N. van Rijn, Bernd Bischl, and Luís Torgo. 2014. "OpenML: Networked Science in Machine Learning." *Journal of Machine Learning Research* 15 (90): 2537–53. https://www.jmlr.org/papers/v15/vanschoren14a.html.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf