ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

High Performance, Low Reliability: Uncertainty Benchmarking for Tabular Foundation Models

José Lucas De Melo Costa, Fabrice Popineau, Arpad Rimmel, Bich-liên Doan *

CentraleSupélec, Université Paris-Saclay
Gif-sur-Yvette - France

Abstract. Recent Tabular Foundation Models (TFMs) have demonstrated state-of-the-art predictive performance, often surpassing Gradient-Boosted Decision Trees (GBDTs). However, the trustworthiness of these models, particularly their uncertainty quantification, has been largely overlooked. We investigate this gap through an extensive study comparing TFMs, GBDTs, and classical baselines on the 112 datasets of the TALENT benchmark. Our results reveal a performance–uncertainty trade-off: although TFMs achieve the highest predictive performance (AUC), they exhibit lower conditional coverage under conformal prediction (SSCS) compared to GBDTs. Complementary experiments on synthetic datasets further characterize the regimes in which this effect intensifies. We conclude that while TFMs advance predictive frontiers, achieving well-calibrated uncertainty remains a major open challenge for their reliable adoption. Code is available at: https://github.com/jose-melo/high-performance-low-reliability

1 Introduction

Tabular data remain commonly found across industrial and scientific domains, where reliable predictive models are crucial for decision making in areas such as finance [1] and healthcare [2]. Historically, deep learning methods have struggled to match the performance of well-established Gradient Boosted Decision Trees (GBDTs) when applied to tabular datasets [3].

Recently, Tabular Foundation Models (TFMs) have emerged as a promising paradigm, leveraging large-scale synthetic pretraining combined with carefully designed inductive biases [4, 5, 6]. Models like TabPFN and TabICL have demonstrated the potential to not only narrow the gap but in some cases surpass classical methods, delivering superior performance on small and medium-scale tabular datasets in terms of accuracy and speed. While recent benchmarks and open leaderboards like *TabArena* [7] assess the performance of TFMs, they provide little insight into model uncertainty. However, measuring uncertainty is as crucial as raw predictive performance, given that tabular data are often embedded in safety-critical applications.

*This work used HPC resources from the Mésocentre of CentraleSupélec and ENS Paris-Saclay, supported by CNRS and Région Île-de-France, and also access to IDRIS under the GENCI allocation AD011011828R5. It was further supported by the Chair "Artificial intelligence applied to credit card fraud detection and automated trading," led by CentraleSupélec and sponsored by LUSIS.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

This lack of systematic analysis leaves open fundamental questions regarding the reliability and deployment of TFMs in critical scenarios. In this work, we address this gap by proposing a comprehensive benchmark of TFMs that evaluates not only predictive performance, but also uncertainty quantification, by the means of conformal prediction. Perpendicular to predictive performance, we introduce a new axis of investigation, providing new insights into the strengths and weakness of Tabular Foundation Models.

Specifically, we benchmark four Tabular Foundation Models against classical and deep learning-based alternatives across the TALENT benchmark [8] on 112 small and medium tabular datasets, complemented by synthetic datasets designed to isolate the conditions under which uncertainty amplifies. Our empirical results reveal that, although TFMs achieve higher performance, they exhibit greater predictive uncertainty, notably under high-noise and low-separability conditions. **Main contributions:**

1. We propose an evaluation protocol combining AUC, calibration, and conformal uncertainty metrics;
2. We conduct a large-scale benchmark comparing TFMs with tree-based, classical and deep learning baselines on tabular data;
3. We provide empirical evidence that TFMs systematically yield lower size-stratified coverage scores than traditional models, further characterizing exactly when this validity gap widens

2 Related Work

Tabular Foundation Models (TFMs) leverage transformer architectures and large-scale synthetic pretraining to approximate Bayesian inference in a single forward pass, achieving state-of-the-art accuracy and efficiency on small and medium tabular datasets [4, 5, 6]. Despite their strong performance, prior work has focused almost exclusively on accuracy, offering limited insight into model reliability or calibration.

Uncertainty quantification (UQ) has become a central topic for evaluating the trustworthiness of modern foundation models. Inspired by recent work applying conformal prediction to benchmark uncertainty in large language models [9], we adopt the same model-agnostic framework to evaluate TFMs. While conformal methods have proven effective for diagnosing overconfidence in LLMs, their application to tabular foundation models remains unexplored. Our work fills this gap by providing the first systematic assessment of TFM uncertainty and its trade-off with predictive performance.

3 Methodology

We evaluated a diverse set of tabular models under a conformal prediction framework, following the protocol illustrated in Figure 1. This setup is inspired by the benchmarking methodology introduced in [9] for Large Language Models.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

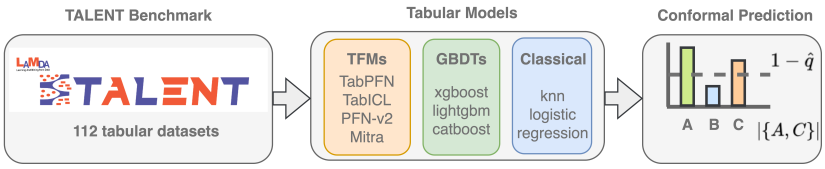

Fig. 1: Overview of the proposed methodology. Models from different families (TFMs, GBDTs, and classical baselines) are trained on the TALENT benchmark and evaluated through conformal prediction to derive calibrated prediction sets.

3.1 Conformal Prediction Framework

To rigorously evaluate the uncertainty of Tabular Foundation Models (TFMs), we adopt the framework of *conformal prediction* (CP). Conformal prediction provides distribution-free guarantees on predictive uncertainty by transforming point predictions into *prediction sets*. These sets are designed to contain the true label with a user-specified probability α, irrespective of the underlying data distribution [10].

Let $\hat{p}(y \mid x)$ denote the class probabilities estimated by a model (e.g., a TFM). Conformal prediction relies on a calibration set I_{cal} disjoint from the training data. For each $(x, y) \in I_{\text{cal}}$, a *nonconformity score* $s(x, y)$ is computed, reflecting how unusual the label y appears given the prediction $\hat{p}(\cdot \mid x)$. Given the collection of nonconformity scores on I_{cal}, we compute the empirical $(1 - \alpha)(1 + 1/|I_{\text{cal}}|)$-quantile q_α. For a new instance x, the conformal prediction set is then defined as $\hat{C}_\alpha(x) = \{y \in \mathcal{Y} : s(x, y) \leq q_\alpha\}$. By construction, this procedure guarantees that $\Pr\{y \in \hat{C}_\alpha(x)\} \geq 1 - \alpha$, ensuring valid marginal coverage without distributional assumptions. We use the *Least Ambiguous Class* (LAC) score, $s(x, y) = 1 - \hat{p}(y \mid x)$ at target coverage $1 - \alpha = 0.90$.

3.2 Metrics

Once prediction sets are generated, their quality is assessed through metrics that jointly capture reliability and informativeness. We first consider the **Coverage Rate (CR)**, defined as the proportion of test samples for which the true label lies inside the prediction set: $\text{CR} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{y_i \in \hat{C}_\alpha(x_i)\}$. A well-calibrated model should achieve coverage close to the target level $1 - \alpha$.

To quantify how informative these sets are, we use the **average set size (SS)** $\text{SS} = \frac{1}{n} \sum_{i=1}^{n} |\hat{C}_\alpha(x_i)|$, which measures the typical cardinality of prediction sets. Smaller sets correspond to more confident predictions, whereas larger sets indicate greater uncertainty. However, informativeness alone is insufficient: a model that produces small but unreliable sets is untrustworthy. To study reliability *conditionally* on uncertainty, we rely on **Size-Stratified Coverage (SSC)**, which measures empirical coverage across groups of samples with similar set sizes: $\text{SSC}(k) = \frac{1}{|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} \mathbf{1}\{y_i \in \hat{C}_\alpha(x_i)\}$, where $\mathcal{G}_k = \{\, i : |\hat{C}_\alpha(x_i)| = k \,\}$. The overall score is $\text{SSCS} = \min_k \text{SSC}(k)$.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Why conditional coverage matters? Consider two groups:

- **Group A:** Small prediction sets ($|\hat{C}_\alpha(x_i)| = 1$) cover the true label only 60% of the time.
- **Group B:** Large prediction sets ($|\hat{C}_\alpha(x_i)| = 5$) cover it 100% of the time.

The overall coverage may still average to the target (e.g., 90%), yet the model is poorly calibrated: **confident predictions (small sets) are unreliable**, while uncertain ones are overly conservative.

3.3 Implementation details

Four families of models were considered: (i) Tabular Foundation Models (TFMs): TabPFN [4], TabICL [6], PFN-v2[5], and Mitra[11]; (ii) GBDTs: xgboost, lightgbm, and catboost; (iii) classical baselines: k-nearest neighbors and logistic regression and (iv) deep learning methods: MLP and TabM [12]. We used a total of 112 small datasets, limited to 10 classes and up to 10000 samples, including binary and multiclass classification problems. The training data is split into 80% training and 20% calibration; hyperparameters are tuned on a held-out validation set, and final results are reported on the test set. For multiclass tasks, we report the weighted one-vs-one AUC. To stress-test model performance, we generate 20 synthetic datasets with high noise and weak class separation.

4 Results

Results are averaged over 15 seeds, with metrics reported in Table 1 and Figure 2.

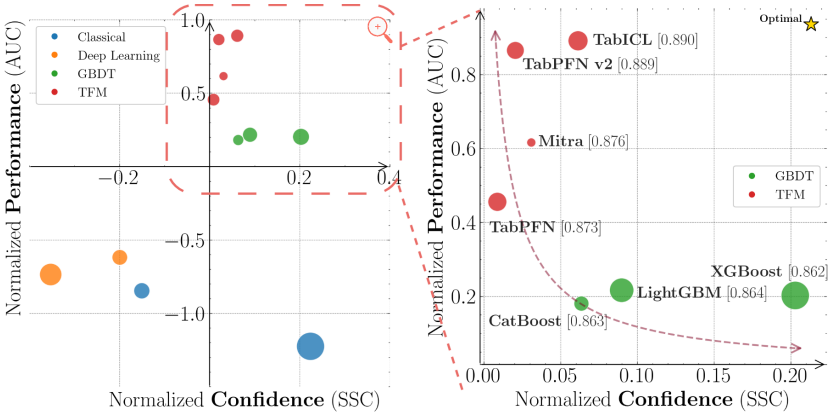

Fig. 2: Performance–confidence trade-off (AUC vs. SSCS). Metrics are normalized per dataset and averaged; marker size denotes ECE. The dashed curve highlights the trade-off, and bracketed values report min–max normalized AUC. Non-normalized metrics are reported in Table 1.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

While Tabular Foundation Models (TFMs) achieve the highest AUC scores, their normalized Size-Stratified Coverage (SSC) remains significantly lower than that of GBDTs, revealing an imbalance between accuracy and reliability. This phenomenon is problematic as models with **poor conditional coverage tend to make overconfident errors**, producing narrow prediction sets that fail to include the true label. As highlighted by the blue versus red cell coloring, there is an **inverted trend between performance (AUC) and uncertainty quality (SSCS)**. Also, although logistic regression achieves good marginal calibration, its sharper prediction sets hurt conditional SSCS near non-linear boundaries.

Model	AUC (↑)	SSCS (↑)
Classical		
knn	0.815 ± 0.022	0.637 ± 0.064
LogReg	0.823 ± 0.024	0.578 ± 0.070
Deep Learning		
tabm	0.832 ± 0.027	0.569 ± 0.072
mlp	0.833 ± 0.026	0.508 ± 0.071
Foundation		
PFN-v2	0.889 ± 0.019	0.496 ± 0.076
tabicl	0.890 ± 0.019	0.494 ± 0.076
mitra	0.877 ± 0.021	0.500 ± 0.075
tabpfn	0.874 ± 0.021	0.517 ± 0.074
GBDT		
lightgbm	0.864 ± 0.022	0.544 ± 0.072
catboost	0.864 ± 0.022	0.532 ± 0.075
xgboost	0.862 ± 0.023	0.540 ± 0.070

Table 1: Non-normalized AUC and SSCS for tabular models, averaged over 112 datasets and 15 seeds. The $\pm$ values denote standard deviation across datasets.

A closer analysis of dataset characteristics sheds light on when this trade-off becomes more pronounced. Under **extreme noise** or **low class separability**, TFMs produce high-variance, overconfident predictions. With non-Gaussian or skewed features, tree-based models remain robust, while TFMs lose calibration.

Group	AUC (↑)	SSCS (↑)
Foundation	$\mathbf{0.924} \pm 0.005$	0.614 ± 0.081
GBDT	0.889 ± 0.007	$\mathbf{0.840} \pm 0.020$

Table 2: Synthetic datasets

Table 2 present the results on the 20 synthetic datasets, with high noise and low class separability, averaged by different runs. As so, the performance-confidence contrast emerges clearly.

Foundation models achieve the highest AUC, but their SSC remains comparatively low, indicating that they produce confident yet poorly calibrated predictions. These controlled results provide empirical evidence for a pattern that was also reinforced in the TALENT benchmark: while TFMs excel in accuracy, GBDTs consistently offer stronger and more trustworthy uncertainty estimates.

5 Conclusion

Our study exposes a fundamental trade-off between predictive accuracy and uncertainty reliability in tabular learning. While Tabular Foundation Models (TFMs) consistently outperform traditional models in terms of AUC, they exhibit weaker conditional coverage, indicating that their confidence does not faithfully reflect true predictive uncertainty. GBDTs, by contrast, maintain more

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

stable uncertainty estimates, even at the cost of slightly lower accuracy.

Overall, **TFMs are best suited for well-curated, low-noise datasets** or few-shot scenarios where synthetic pretraining and in-context adaptation can leverage prior structure for higher accuracy. On small-to-medium datasets, GBDTs remain the most reliable choice for high-noise, heterogeneous, or complex feature distributions, offering better calibration and robustness.

Our analysis is limited to small and medium-scale datasets and to a model-agnostic conformal framework. Future work should evaluate TFMs under distribution shifts, explore model-specific uncertainty mechanisms (e.g., ensembling or Bayesian heads), and investigate how alternative pretraining distributions could improve calibration and reduce overconfidence.

References

[1] Prince Grover, Julia Xu, Justin Tittelfitz, Anqi Cheng, Zheng Li, Jakub Zablocki, Jianbo Liu, and Hao Zhou. Fraud Dataset Benchmark and Applications, September 2023. arXiv:2208.14417 version: 3.

[2] Flavio Di Martino and Franca Delmastro. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review*, 56(6):5261–5315, June 2023.

[3] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? 2022.

[4] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. September 2022.

[5] 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, January 2025. Publisher: Nature Publishing Group.

[6] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data, February 2025. arXiv:2502.05564 [cs].

[7] 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, June 2025. arXiv:2506.16791 [cs].

[8] Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and Han-Jia Ye. TALENT: A Tabular Analytics and Learning Toolbox, July 2024. arXiv:2407.04057 [cs].

[9] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking LLMs via Uncertainty Quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385, December 2024.

[10] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, December 2022. arXiv:2107.07511 [cs].

[11] Xiyuan Zhang, Danielle C. Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W. Mahoney, Cuixiong Hu, Huzefa Rangwala, George Karypis, and Bernie Wang. Mitra: Mixed Synthetic Priors for Enhancing Tabular Foundation Models, October 2025. arXiv:2510.21204 [cs].

[12] Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling, February 2025. arXiv:2410.24210 [cs].