Out in the Wild: Investigating the Impact of Imperfect Data on a Tabular Foundation Model

Vasileios Papastergios
Aristotle University of Thessaloniki
Thessaloniki, Greece
papster@csd.auth.gr

Anastasios Gounaris
Aristotle University of Thessaloniki
Thessaloniki, Greece
gounaria@csd.auth.gr

ABSTRACT

Data quality issues are pervasive in real-world applications, posing a critical concern for the increasing deployment of foundation models in high-stakes domains. In this paper, we present the first analysis of how real-world data errors impact the performance of TabPFN, a recent tabular foundation model achieving state-of-the-art performance on tabular tasks across domains. Pre-trained exclusively on synthetic and evaluated on curated, benchmark data, its capabilities remain unexplored when considering imperfect datasets; the prevalent species in the wild of real-world data science. To bridge this gap, we introduce a novel, extensible experimental framework, specially designed for assessing the impact of data errors on tabular foundation models. Building upon well-established benchmarking and data corruption techniques, our investigation offers actionable insights into how imperfect data affects nuanced capabilities beyond predictive performance: in-context learning and internal representations. Our rigorous experimental evaluation comprising more than 10K experiments shows that the presence of data errors affects the representations of not only the corrupted but also the clean samples. Also, targeted endeavors to clean the context data can be beneficial, especially for errors on categorical values.

VLDB Workshop Reference Format:
Vasileios Papastergios and Anastasios Gounaris. Out in the Wild: Investigating the Impact of Imperfect Data on a Tabular Foundation Model. VLDB 2025 Workshop: 14th International Workshop on Quality in Databases (QDB'25).

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/Bilpapster/QualiTab.

1 NEW AVENUES FOR DATA QUALITY

Tabular data, structured in rows and columns, forms the foundation for countless applications across domains, including healthcare [4, 24], finance [3, 30], cybersecurity [11], and scientific research [9, 55]. Despite its critical importance in real-world decision-making, machine learning approaches for tabular data had not experienced the transformative advances seen in other domains [53], such as computer vision [6, 20, 28] and natural language processing [12, 40, 57]. Tree-based ensemble methods had maintained their position as the go-to approach for tabular tasks for nearly two decades [13, 27, 41], highlighting the challenges posed by the significantly heterogeneous nature of tabular data [51]. The recent introduction of the Tabular Prior-data Fitted Network (TabPFN[1]) [22] marks a significant breakthrough in this landscape. As a foundation model for tabular data, TabPFN leverages in-context learning [10] to address classification and regression tasks. Unlike conventional models, which require individual training for each dataset, TabPFN undergoes a single pre-training phase on millions of synthetic datasets, enabling it to capture diverse data patterns and relationships. When evaluated on curated, benchmark datasets, TabPFN demonstrates unprecedented generalization capabilities, outperforming state-of-the-art methods while requiring substantially less computational resources [23, 56].

However, real-world tabular datasets are usually far from curated, benchmark ones. In fact, data quality issues are pervasive in machine learning workflows [26, 46–48], with research indicating that organizations lose an average of $12.9 million annually due to poor data quality [25]. The impact of these issues extends beyond financial figures, especially in high-stakes domains. Despite its recent emergence, TabPFN has already revolutionized tabular tasks across numerous such critical domains, including healthcare and medicine [5, 14, 15, 17–19, 38, 39, 52], industrial fault detection [32], and environmental monitoring [29]. However, the exploration of how data quality affects such state-of-the-art foundation models remains a largely uncharted territory. Our work is therefore motivated by this fact. Given the ubiquity of data errors in modern data science, our research aims to equip data practitioners with the knowledge needed to make informed decisions when deploying TabPFN in practical applications.

Distinction from existing work. Extensive research has been conducted to examine how data quality affects traditional machine learning models [1, 31, 33, 36], demonstrating the impact of various error types on algorithms like random forests and gradient boosting machines. However, the emergence of tabular foundation models like TabPFN, with their unprecedented performance, opens up new avenues for investigation. Distinct from prior work, our research goes beyond mere predictive performance, offering novel insights into how data quality issues influence the nuanced capabilities of these models, including their internal representations and in-context learning. To the best of our knowledge, this is the first systematic investigation at the intersection of data quality and foundation models that adopts a user-centric lens, mirroring the challenges of imperfect data faced in real-world deployments.

Proceedings of the VLDB Endowment. ISSN 2150-8097.

[1]We use *TabPFN* to refer to the 2[nd] version of the model [22]. Some works cite it as *TabPFN v2* to distinguish from [21]. We adopt the naming used by the original authors.

Contributions. Motivated by TabPFN's remarkable – yet unexplored in the presence of realistic data errors – performance, we make the following contributions:

- We raise awareness towards a new research avenue at the intersection of data quality and tabular foundation models by presenting preliminary information about TabPFN and the research questions targeted by this work (Section 2);
- We develop and open-source the first experimental framework designed to systematically evaluate the impact of data quality on tabular foundation models, supporting easy integration of new datasets, error types, and metrics (Section 3);
- We leverage this framework to conduct an extensive evaluation comprising more than 10 thousand experiments on 67 real-world datasets, 3 realistic error types, 3 corruption rates, and 3 practice-driven scenarios (Section 3); and
- We reveal how data errors at various rates influence TabPFN's embeddings, uncovering actionable insights into the benefits of informed data cleaning decisions (Section 4).

In Section 5 we review related work, before concluding with future research directions in Section 6. The source code of our evolving experimental framework is publicly available under an open source license at https://github.com/Bilpapster/QualiTab.

2 PRELIMINARIES AND RESEARCH QUESTIONS

This section provides background on TabPFN while motivating our investigation through targeted research questions (RQs) that guide our multi-level analysis. We first outline general information about TabPFN (Section 2.1) to render our work self-contained. We then focus on more nuanced capabilities: in-context learning for predictions (Section 2.2) and embeddings extraction through representation learning (Section 2.3). For comprehensive details on TabPFN, we refer readers to Hollmann et al. [22] and Ye et al. [56].

2.1 What is TabPFN?

TabPFN is a foundation model for tabular data that leverages in-context learning for classification and regression tasks, without explicit fine-tuning required for each new task. Data practitioners can easily use it on their own tabular datasets through a few lines of Python [42] or R [44] code. A service for free GPU inference is also provided by the authors to registered users [45].

TabPFN was pre-trained once on more than 100 million synthetic tabular datasets generated using causal models to capture diverse underlying relationships. This pre-training approach eliminates concerns about potential data leakage when evaluating on real-world datasets, like we do in our study. Extensive evaluations reveal TabPFN's exceptional performance not only for tables [56] but also, startlingly, for time series [23]. Architecturally, TabPFN is transformer-based with a two-way attention mechanism; one across features and one across samples. It is designed for small-to-medium scale tabular data, with operational limits of 10 000 samples, 500 features, and, in case of classification, 10 classes. To ensure these limitations are met, we apply a principled transformation methodology to larger datasets that preserves their essential characteristics, similar to the one described by Ye et al. [56], as detailed in Section 3.

2.2 In-Context Learning for Prediction

Formally, a tabular dataset is a set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ of N *samples* $(\boldsymbol{x}_i, y_i)$, where, $\boldsymbol{x}_i \in \mathbb{R}^d$ is represented by d *features* while the *label* y_i belongs to a set $C = \{C_1, \ldots, C_m\}$ of m classes (classification) or is a numerical value (regression). For the training samples $\boldsymbol{X}_{train}$, the labels $\boldsymbol{y}_{train}$ are known, while for the test (unseen) data $\boldsymbol{X}_{test}$, the labels $\boldsymbol{y}_{test}$ are unknown. While a typical machine learning model is trained on $\boldsymbol{X}_{train}$ and $\boldsymbol{y}_{train}$, producing predictions $\hat{\boldsymbol{y}}_{test}$ when presented with $\boldsymbol{X}_{test}$, TabPFN follows a different approach [37]. At its core, the model approximates Bayesian prediction by learning the posterior predictive distribution $p(\hat{\boldsymbol{y}}_{test} \mid \boldsymbol{X}_{test}, \boldsymbol{X}_{train}, \boldsymbol{y}_{train})$ for the prior defined by its synthetic training datasets. Its key innovation is the use of in-context learning (ICL) [10], a mechanism that has driven the success of large language models. At inference time, TabPFN leverages its transformer architecture to process an entire dataset ($\boldsymbol{X}_{train}$, $\boldsymbol{y}_{train}$, and $\boldsymbol{X}_{test}$) in a single forward pass. The model conditions on the provided training examples as context to infer the underlying relationships and directly predict $\hat{\boldsymbol{y}}_{test}$. Essentially, the model learns to learn from the provided context, mimicking a learned approximation of Bayesian inference over the distribution of synthetic datasets it was trained on.

2.3 Tabular Representation Learning

Beyond its predictive capabilities, TabPFN's transformer-based architecture inherently learns rich representations of tabular data that can be extracted as *embeddings*. As input data flows through TabPFN's multiple transformer layers, the architecture naturally generates embeddings at the output of the final layer. Thus, embeddings are inherently involved in all model's operations, including making predictions. Given a tabular dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ where each sample $\boldsymbol{x}_i \in \mathbb{R}^d$ has d features, TabPFN can be viewed as learning a mapping $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^k$ that transforms the original d-dimensional feature vectors into a k-dimensional embedding space. The resulting embeddings $\boldsymbol{e}_i = f_\theta(\boldsymbol{x}_i) \in \mathbb{R}^k$ encapsulate the complex relationships between features observed in the current input, synthesized with the accumulated knowledge acquired during TabPFN's pre-training on diverse synthetic data.

These learned embeddings can significantly benefit various downstream tasks, unlocking for tables a domain typically considered a privilege only of other modalities, such as images and text [50]. By projecting the data into a latent space, more meaningful semantic information can be captured compared to the raw features [56]. This can lead to improved performance in tasks such as clustering, dimensionality reduction, and as input features for other machine learning models. The rich contextual information encoded within these embeddings, potentially capturing intricate feature interactions through TabPFN's attention mechanisms, makes them a valuable asset for data analysis and modeling.

While extensive analyses on TabPFN's embeddings demonstrate highly promising results [22, 56], they only consider high-quality, curated datasets. This can be misleading for practitioners using TabPFN as a feature extractor on real-world, imperfect datasets, expecting to get state-of-the-art performance out of the box. Motivated by the embeddings' pivotal involvement in all model's operations and the pervasive nature of data errors in modern data science tasks, we pose the research questions of our study:

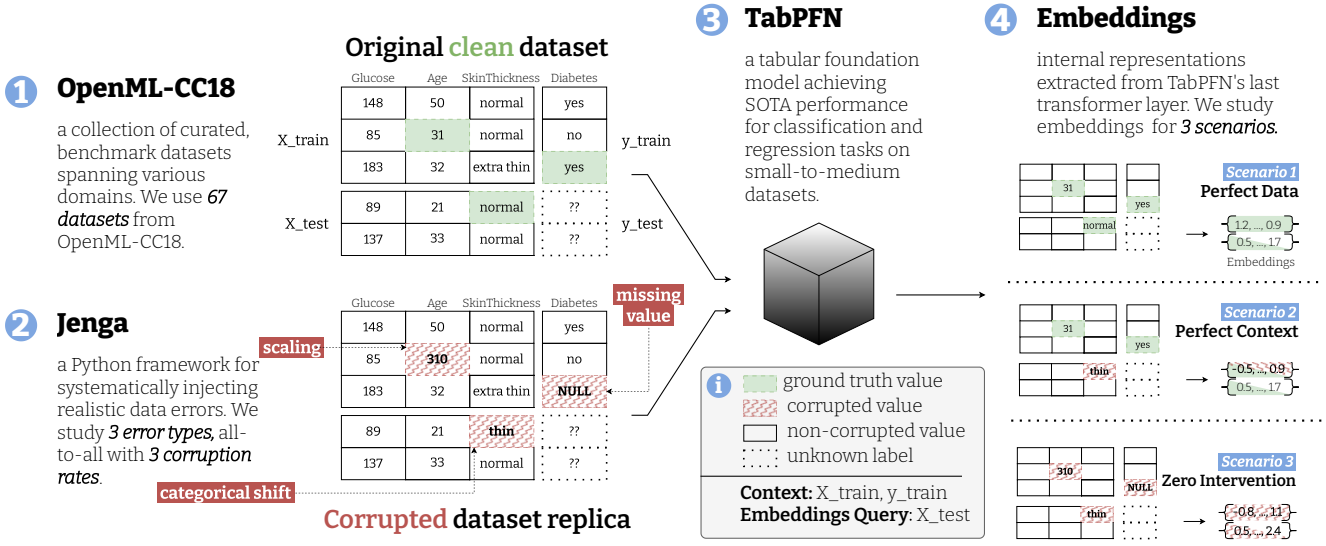

Figure 1: High-level overview of the experimental setup.

RQs of current study

RQ1 What is the impact of imperfect data on TabPFN's internal representations (i.e., embeddings)?

RQ2 What are the benefits of cleaned context data and how can they drive informed cleaning decisions?

3 EXPERIMENTAL FRAMEWORK DETAILS

In this section, we elaborate on our experimental framework, tailored to assessing the impact of data quality on tabular foundation models. While we instantiate it for TabPFN, it is carefully designed to be easily extensible with more datasets, error types, models, and evaluators. Figure 1 illustrates a high-level overview of the setup.

Hardware. All experiments were conducted utilizing Kaggle's free-tier computational resources: 2 Tesla T4 GPUs with 15GB of memory for a total of 45.7 GPU hours.

Datasets. We utilized the OpenML-CC18 benchmark [8], a widely adopted collection of 72 curated, real-world classification tabular datasets across diverse domains. Due to hardware limitations, our analysis includes the **67 datasets** that could be successfully loaded in memory at inference time. For the subset of datasets exceeding TabPFN's operational limits (10K samples, 500 features, 10 classes), we employed a principled transformation methodology detailed in Algorithm 1. Our methodology, inspired by the post-hoc divide-and-conquer technique proposed in [56], aims to preserve the dataset's essential characteristics, while transforming it to be directly compatible with TabPFN for fair comparisons.

Systematic Data Corruption. To study the impact of realistic data errors on TabPFN's internal representations, we systematically injected errors to the original (clean) datasets, creating corrupted replicas of them, as shown in Figure 1. For the corruption process, we utilized Jenga [48], a well-established tool for introducing realistic data errors in machine learning datasets. We focus on three distinct error types (E). **(E1) missing values:** a generic corruption simulating absent or unknown data; **(E2) scaling:** randomly scaling numerical values by 10, 100, or 1000, mimicking unit conversion errors or pre-processing bugs; and **(E3) categorical shift:** randomly swapping categorical values within a column, simulating misclassifications or data entry errors. Since data errors appear in varying frequency and localization levels, we independently combined all studied error types with several corruption rates: **10%, 20%, and 40%** for samples and $max(\{20\%, 1\})$ for columns.

Our selection of error types is in line with the recent contributions by Mohammed et al. [34, 36], which quantify the impact of data quality on traditional machine learning. Additionally, Schinninger et al. [49] include these errors in their data stream-specific polluter Icewafl, highlighting the community's consensus on the ubiquity of these error types for both static and streaming data. This choice also allows us to investigate TabPFN's resilience to both error types its pre-training has accounted for, such as missing values [22], and others common in real-world data.

Randomness and Reproducibility. To ensure robust and reproducible results, each experiment was repeated 10 times with 10 distinct random seeds. The same random seed was consistently used within a single experimental run for any random operations, such as feature and sample selection in Algorithm 1.

Practice-Driven Scenarios. To set the stage for addressing our RQs, we investigated three distinct scenarios (S) that mirror how data practitioners might leverage TabPFN for embedding extraction in data science tasks. In each scenario, the training data serves as *context* for the model to learn underlying patterns, while the unlabeled test data acts as the *query* for which embeddings are extracted. As depicted in Figure 1, there is a one-to-one correspondence between the samples in the query set and the resulting embeddings.

Based on potential data quality management strategies employed by practitioners, we defined the following scenarios. **(S1) Perfect Data:** Represents a baseline where both the context and the query

Algorithm 1 Principled Transformation for Large Datasets

Input: Dataset $\mathcal{D}$ with N samples, d features, and m classes. **Output:** Transformed training set $\mathcal{D}'_{train}$ and test set $\mathcal{D}'_{test}$ compatible with TabPFN's limits.

if $m > 10$ **then**
  Randomly sample 10 classes $C' \subset C$.
  $\mathcal{D} \leftarrow \{(\boldsymbol{x}_i, y_i) \in \mathcal{D} \mid y_i \in C'\}$.
if $d > 500$ **then**
  Randomly sample 500 features $F' \subset \{1, \ldots, d\}$.
  $\boldsymbol{x}_i \leftarrow \{\boldsymbol{x}_{i,j}$ for all $j \in F'\}$.
Split $\mathcal{D}$ into $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$; drop labels for $\mathcal{D}_{test}$. ▷ 70/30 train/test split
if $|\mathcal{D}_{train}| > 10K$ **then**
  Randomly sample 10K samples to keep in $\mathcal{D}_{train}$.
return $\mathcal{D}_{train}, \mathcal{D}_{test}$.

Table 1: Table of symbols.

Symbol	Description
ALL	evaluation performed on all samples
CC	label assignment based on clean/corrupted
CLE	evaluation performed on clean (non-corrupted) samples
COR	evaluation performed on corrupted samples
CosD	cosine distance (1− cosine similarity)
TL	label assignment based on true labels (classes)
zED	z-normalized Euclidean Distance

data are entirely clean, reflecting an unrealistic yet optimal condition. **(S2) Perfect Context:** Simulates a scenario where practitioners invest resources in cleaning a subset of their data to provide a high-quality context for TabPFN. This cleaned context, established once, can then be used repeatedly for extracting embeddings from new, potentially imperfect query samples arriving either offline or in a streaming manner. **(S3) Zero Intervention:** Reflects a situation where practitioners opt to forgo potentially costly or unsuccessful data cleaning operations [26] for both the context and the query data. In this case, embeddings can still be extracted offline or online, with the possibility of the context being dynamically updated (e.g., to include only recent, yet potentially imperfect, samples). This scenario essentially assesses the extent to which TabPFN's extensive pre-training can compensate for the presence of data errors. We deliberately omit the scenario of a corrupted context and a cleaned query, considering it an impractical approach as it would necessitate cleaning every new query while neglecting the (typically less volatile) context data.

4 FINDINGS AND DISCUSSION

In this section, we present and discuss the key findings from our experiments, providing answers for our RQs. To help readers better comprehend our findings, we adopt a three-stage presentation strategy. We start with how we evaluated the impact of imperfect data on the extracted embeddings (Section 4.1). We then present the evaluation results, focusing on patters that are consistently observed (Section 4.2). Finally, we synthesize these patterns into actionable key takeaway findings meant for data practitioners (Section 4.3).

4.1 Impact Evaluation Metrics

To evaluate the extracted embeddings, we have used both supervised and unsupervised metrics in different variations as follows.

Distance from perfect self. Quantifies the distance between the current embeddings and the respective ones extracted from the Perfect Data scenario. We compute two different distance measures for robust results: cosine and z-normalized Euclidean distance. This metric directly captures the impact of the presence of data errors, using the ideal scenario as reference. Higher distance means that the embeddings are more impacted. We detail our analysis on all, only clean, and only corrupted samples for fine-grained insights.

Linear Probing. Quantifies the separability of extracted embeddings in the latent space by training a linear classifier with default parameters on them. Due to its simplicity, linear probing is widely used for evaluating image embeddings [7]. We favor ROC AUC metric over accuracy to avoid misleading results for imbalanced datasets. Higher ROC AUC score means that the embeddings can be more easily (linearly) separated into classes in the latent space; thus they are potentially more suitable for downstream tasks. Beyond the true labels (TL, classes), we also use the labels of clean/corrupted (CC) samples. Intuitively, this provides insights into the separability of the latent space into clean and corrupted samples.

4.2 Evaluation Results

Here, we present our evaluation results, focusing on repetitive, consistent patterns across various experimental configurations. Table 1 summarizes the symbols used in this section, while Figures 2 – 4 present comparative evaluation results between the three identified scenarios. We elaborate on these figures in the rest of the section.

Figure 2 depicts the cosine (cosD) and z-normalized Euclidean (zED) distances of embeddings from their perfect self, comparing the perfect context (dashed) and zero intervention (solid) scenarios. When all (ALL) or only clean (CLE) samples are considered, the distance gets higher in the presence of more erroneous samples, which means that the magnitude of impact on the embeddings is analogous to the error rate. Interestingly, the distance of corrupted (COR) samples is relatively high even for low corruption rates, with slight, if any, increases in the presence of more errors. In the majority of configurations, having perfect context data leads to reduced distances compared to the zero intervention scenario.

Figure 3 focuses on the separability of the latent space, involving all scenarios. Here, the green dash-dotted line represents the baseline, perfect data scenario. The clean samples alone (subfig. a) tend to retain their separability levels in the presence of scaled or missing values but struggle for categorical shifts. However, when considering the corrupted samples alone (subfig. b) or all the samples as a whole (subfig. c), the separability regarding the true labels (TL) significantly diminishes as the number of errors increases. The presence of clean context data yields no significant advantages in such cases. Interestingly, for all three error types, the separability between clean and corrupted (CC) samples appears significantly improved compared to the baseline (subfig. d).

Figure 4 illustrates a fine-grained analysis on the size of clean context with respect to the separability of the latent space. When true labels are considered, up to 5K clean samples (dashed) appear to produce a slightly more separable space compared to 5K or more samples (solid) for missing and scaled values. However, corrupted samples seem to be more easily separated from clean ones when a larger clean context is available, especially for categorical shifts.

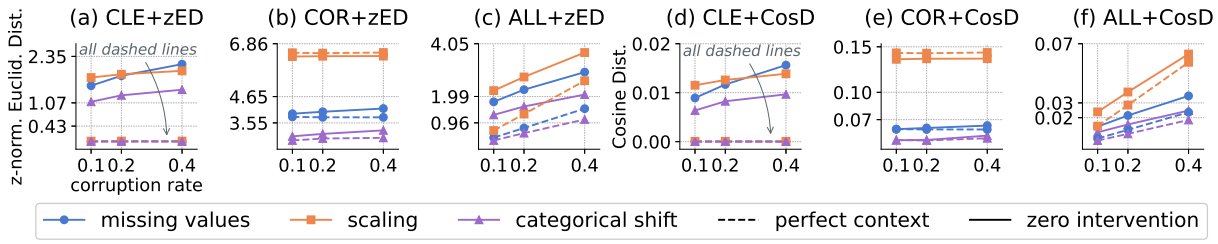

Figure 2: Perfect Context vs Zero Intervention: Distance of embeddings from the respective Perfect Data representation.

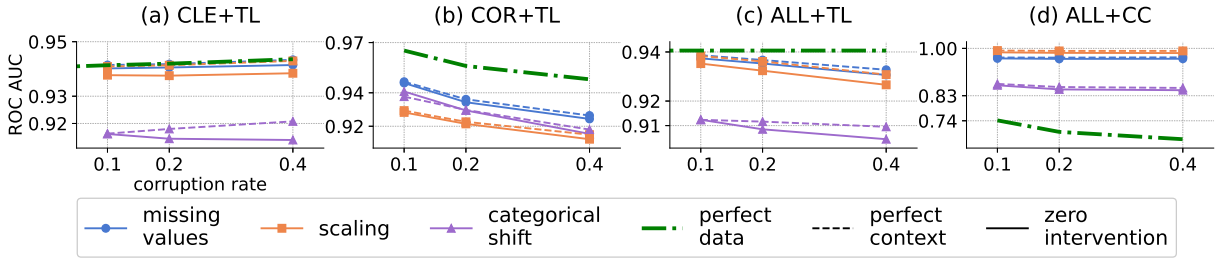

Figure 3: All scenarios: Performance of linear probing on the extracted embeddings.

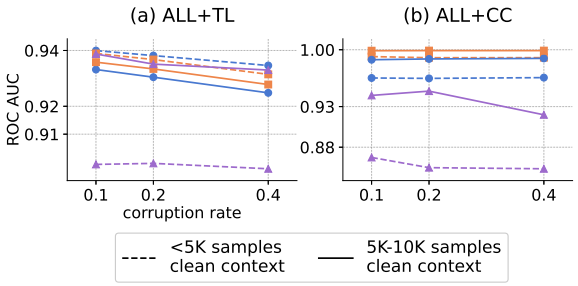

Figure 4: Perfect context: Less vs more clean context samples.

4.3 Synthesized Takeaway Findings

In this section, we synthesize our evaluation results in the form of actionable takeaway findings (F) for practitioners. Figure 5 illustrates indicative examples of extracted embeddings in the 2-dimensional space to provide a complementary, visual perspective and help readers better comprehend our findings. Due to the intertwined nature of our two RQs, we opt to jointly present our findings for them, starting from the erroneous samples in F1.

F1. Errors hit different [RQ1 & RQ2, Figures 2, 3]

The representations of erroneous samples tend to **get distorted, even with perfect context**. The distance of the erroneous embeddings from their perfect self is high, even for lower corruption rates (Figure 2b,e), while the separability diminishes significantly (Figure 3b). Cleaning the context tends to yield slight advantages (see solid and dashed same-colored lines), which should be considered against the cleaning costs.

While the positive effect of cleaning the context may be limited for corrupted samples, F2 reveals the other side of the coin: maintaining imperfect context comes with a hidden cost.

F2. Even clean data pays the price [RQ1, Figures 2, 3, 5]

Imperfect context impacts the model's embeddings, **even** for **non-corrupted samples**. The more the data errors, the less similar the non-corrupted samples become to their perfect self (Figure 2a, d), which is also reflected on the whole query set (Figure 2c, f). This makes the latent space less separable (Figure 3a, c) and the extracted embeddings potentially less informative when used as features for downstream tasks. Visually, this results to a gradual blending of clusters (see left-to-right transition across subfig. b-e in each row of Figure 5).

Thus, a clean context can yield benefits; yet, it comes with its own costs. Given that the context can be cleaned once and used for multiple inferences, F3 attempts to help practitioners quantify the balance between the costs and the potential benefits.

F3. If to clean, fewer may be enough [RQ2, Figure 4]

A **larger size of clean context samples may not necessarily improve** the quality of embeddings for unseen new samples. Contrasting common intuition, as shown in Figure 4, up to 5K of clean context samples (dashed lines) yield on par, or even slightly better, evaluation results on average compared to a larger context size of 5K or more samples (solid lines). Further investigation is required to reveal the connection between the dataset characteristics and cost-effective cleaning.

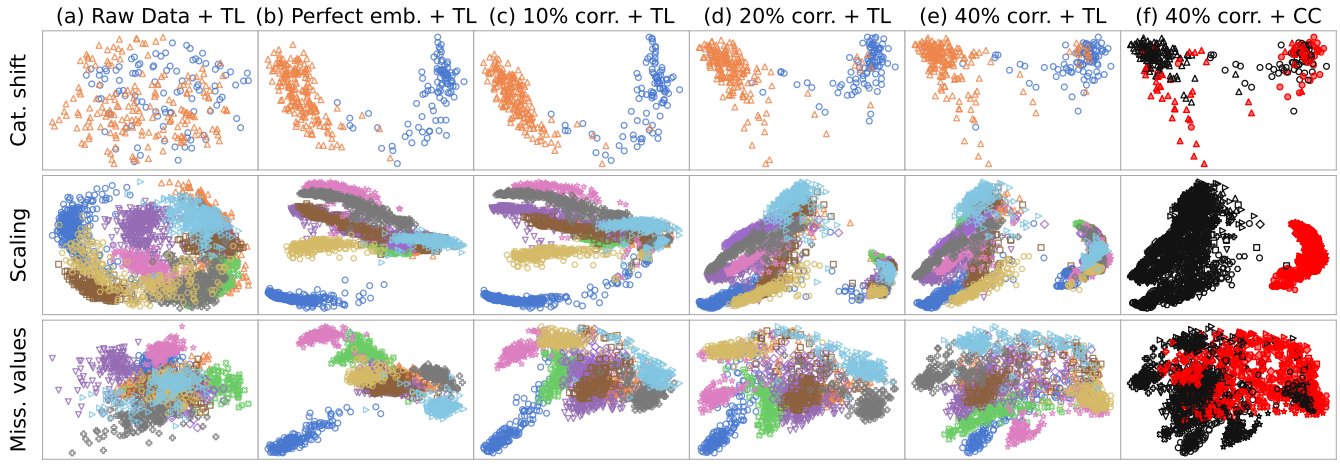

Figure 5: Indicative 2D visualizations of embeddings for scaling (top), missing values (bottom), and categorical shift (bottom).

Finally, assuming that cleaning is decided, F4 provides insights into *what* to clean. This knowledge could help to get the most out of cleaning, especially when budget constraints apply.

F4. If to clean, do it smartly [RQ1 & RQ2, All Figures]

Different error types have different impact on TabPFN's embeddings. While **categorical shifts** (top row of Figure 5, purple in the rest figures) seem to have the greatest impact on the internal representations (Figure 3), providing a (large enough) clean context leads to the highest performance increase (see solid and dashed purple lines in Figure 4). On the other hand, **abnormally scaled values** (middle row, orange) seem to trigger some inherent outlier detection mechanism (see the rightmost, heterogeneous cluster in the middle row of Figure 5f). This is also revealed by the nearly perfect separation of the latent space into clean and corrupted samples (see orange lines in Figure 3d); separability not observed in the baseline (green). This behavior, however, makes scaled values the hardest error to mitigate via clean context (Figure 4). Finally, **missing values** (bottom row, blue) is the error type that TabPFN presents to be the most robust against across the majority of experimental configurations, while still impacting the representations compared to the ideal, baseline scenario.

Based on the task at hand and data profiling results, informed cleaning endeavors could be applied to selectively address specific error types that can have the greatest positive effect on the performance of the downstream task. Further investigation on more error types and larger corpora of datasets is required to advance this research line towards model- and task-aware data cleaning.

5 RELATED WORK

Recent research has increasingly focused on the crucial interplay between data quality and the value it brings to downstream tasks [2], raising concerns about the complexity of assessing intangible benefits like knowledge gained from data [35]. Our work touches this from a practical standpoint by investigating what TabPFN actually learns from the input data, using extracted embeddings as a proxy.

The vast majority of existing work has primarily investigated the impact of data quality on the predictive performance of traditional machine learning models. Li et al. [31] provided the first systematic analysis on the impact of data quality issues on classification tasks. Abdelaal et al. [1] extended the scope towards data cleaning, investigating the impact of combinations of error detection and correction techniques on machine learning tasks in the presence of various data errors. Mohammed et al. [33, 36] further advanced this research line with an extensive evaluation and a practical tool targeting data quality for traditional machine learning applications in multi-error scenarios. In contrast, our study uniquely examines the effects of data errors on the more nuanced, internal representations learned by emerging tabular foundation models, offering insights into their less tangible, yet pivotal capabilities.

Our work also relates to the growing interest in learning from imperfect data, overviewed by Karlaš et al. [26], as a potentially more practical alternative to costly or infeasible exhaustive data cleaning. We contribute to this direction by systematically evaluating three distinct scenarios: an ideal baseline, a strategy of cleaning the context data only, and a scenario with zero intervention. This allows us to explore the trade-offs associated with different levels of cleaning efforts when leveraging tabular foundation models. Finally, contrasting with studies focused on ideal pre-training data, such as the one by Wen et al. [54] for time-series data, our research examines the practical, resource-efficient use of the readily available TabPFN by data practitioners facing real-world quality challenges in downstream tasks.

6 CONCLUSION AND FUTURE WORK

In this work, we presented the first systematic framework for assessing the impact of realistic data errors on tabular foundation models. Focusing on TabPFN, we studied how its learned representations are affected by errors of various types, at varying rates, and across different scenarios encountered in practice. Our extensive experimental evaluation demonstrates that, although errors significantly

distort the representations of corrupted values, clean data also pays the price, leading to a less separable latent space overall. Among the studied data errors, different error types appear to have different impact, potentially informing targeted data cleaning strategies.

This study opens up new avenues at the intersection of data quality and tabular foundation models, yet covers a limited spectrum of potential investigations. We have identified three orthogonal extension directions: models, datasets and error types. In particular, since we first conducted our evaluations, new tabular foundation models have emerged, such as TabICL [43], highlighting the vibrant evolution of this research area. As part of future work, we plan to incorporate such models in our experimental framework with the view to enrich and generalize our takeaway findings. Using a larger corpora of tabular datasets, such as the very recently introduced TabArena [16], could also advance this research line towards the same direction. Additionally, investigating more error types, such as constraint violations, and covering more data quality dimensions, such as consistency and timeliness, could be of great benefit to the broader data science community. Finally, an alternative promising research direction could be the application of our findings to the development of error detection and data cleaning techniques, potentially leveraging properties of the latent space produced by tabular foundation models.

REFERENCES

[1] Mohamed Abdelaal, Christian Hammacher, and Harald Schöning. 2023. REIN: A Comprehensive Benchmark Framework for Data Cleaning Methods in ML Pipelines. https://doi.org/10.48786/EDBT.2023.43

[2] Anastasia Ailamaki, Samuel Madden, Daniel Abadi, Gustavo Alonso, Sihem Amer-Yahia, Magdalena Balazinska, Philip A. Bernstein, Peter Boncz, Michael Cafarella, Surajit Chaudhuri, Susan Davidson, David DeWitt, Yanlei Diao, Xin Luna Dong, Michael Franklin, Juliana Freire, Johannes Gehrke, Alon Halevy, Joseph M. Hellerstein, Mark D. Hill, Stratos Idreos, Yannis Ioannidis, Christoph Koch, Donald Kossmann, Tim Kraska, Arun Kumar, Guoliang Li, Volker Markl, Renée Miller, C. Mohan, Thomas Neumann, Beng Chin Ooi, Fatma Ozcan, Aditya Parameswaran, Ippokratis Pandis, Jignesh M. Patel, Andrew Pavlo, Danica Porobic, Viktor Sanca, Michael Stonebraker, Julia Stoyanovich, Dan Suciu, Wang-Chiew Tan, Shiv Venkataraman, Matei Zaharia, and Stanley B. Zdonik. 2025. The Cambridge Report on Database Research. https://doi.org/10.48550/ARXIV.2504.11259

[3] Khaled Gubran Al-Hashedi and Pritheega Magalingam. 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review* 40 (2021), 100402. https://doi.org/10.1016/j.cosrev.2021.100402

[4] Salah S. Al-Zaiti, Christian Martin-Gill, Jessica K. Zègre-Hemsey, Zeineb Bouzid, Ziad Faramand, Mohammad O. Alrawashdeh, Richard E. Gregg, Stephanie Helman, Nathan T. Riek, Karina Kraevsky-Phillips, Gilles Clermont, Murat Akcakaya, Susan M. Sereika, Peter Van Dam, Stephen W. Smith, Yochai Birnbaum, Samir Saba, Ervin Sejdic, and Clifton W. Callaway. 2023. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine* 29, 7 (June 2023), 1804–1813. https://doi.org/10.1038/s41591-023-02396-3

[5] Sarah A. Alzakari, Asma Aldrees, Muhammad Umer, Lucia Cascone, Nisreen Innab, and Imran Ashraf. 2024. Artificial intelligence-driven predictive framework for early detection of still birth. *SLAS Technology* 29, 6 (2024), 100203. https://doi.org/10.1016/j.slast.2024.100203

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6836–6846.

[7] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15619–15629.

[8] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. 2019. OpenML Benchmarking Suites. *arXiv:1708.03731v2 [stat.ML]* (2019).

[9] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2024. Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 35, 6 (2024), 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[11] Anna L. Buczak and Erhan Guven. 2016. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials* 18, 2 (2016), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

[12] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 3, Article 39 (March 2024), 45 pages. https://doi.org/10.1145/3641289

[13] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[14] Daniiar Dyikanov, Aleksandr Zaitsev, Tatiana Vasileva, Iris Wang, Arseniy A. Sokolov, Evgenii S. Bolshakov, Alena Frank, Polina Turova, Olga Golubeva, Anna Gantseva, Anna Kamysheva, Polina Shpudeiko, Ilya Krauz, Mary Abdou, Madison Chasse, Tori Conroy, Nicholas R. Merriam, Julia E. Alesse, Noel English, Boris Shpak, Anna Shchetsova, Evgenii Tikhonov, Ivan Filatov, Anastasia Radko, Anastasiia Bolshakova, Anastasia Kachalova, Nika Lugovykh, Andrey Bulahov, Anastasiia Kilina, Syimyk Asanbekov, Irina Zheleznyak, Pavel Skoptsov, Evgenia Alekseeva, Jennifer M. Johnson, Joseph M. Curry, Alban J. Linnenbach, Andrew P. South, EnJun Yang, Kirill Morozov, Anastasiya Terenteva, Lira Nigmatullina, Dmitry Fastovetz, Anatoly Bobe, Linda Balabanian, Krystle Nomie, Sheila T. Yong, Christopher J.H. Davitt, Alexander Ryabykh, Olga Kudryashova, Cagdas Tazearslan, Alexander Bagaev, Nathan Fowler, Adam J. Luginbuhl, Ravshan I. Ataullakhanov, and Michael F. Goldberg. 2024. Comprehensive peripheral blood immunoprofiling reveals five immunotypes with immunotherapy response characteristics in patients with cancer. *Cancer Cell* 42, 5 (May 2024), 759–779.e12. https://doi.org/10.1016/j.ccell.2024.04.008

[15] Moumen El-Melegy, Ahmed Mamdouh, Samia Ali, Mohamed Badawy, Mohamed Abou El-Ghar, Norah Saleh Alghamdi, and Ayman El-Baz. 2024. Prostate Cancer Diagnosis via Visual Representation of Tabular Data and Deep Transfer Learning. *Bioengineering* 11, 7 (2024). https://doi.org/10.3390/bioengineering11070635

[16] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. 2025. TabArena: A Living Benchmark for Machine Learning on Tabular Data. https://doi.org/10.48550/ARXIV.2506.16791

[17] Mert Karabacak et al. 2024. Data-Driven Prognostication in Distal Medium Vessel Occlusions Using Explainable Machine Learning. *American Journal of Neuroradiology* 46, 4 (Oct. 2024), 725–732. https://doi.org/10.3174/ajnr.a8547

[18] Mert Karabacak et al. 2024. A machine learning-based approach for individualized prediction of short-term outcomes after anterior cervical corpectomy. *Asian Spine Journal* 18, 4 (Aug. 2024), 541–549. https://doi.org/10.31616/asj.2024.0048

[19] Perciballi Giulia et al. 2024. Adapting TabPFN for Zero-Inflated Metagenomic Data. In *NeurIPS 2024 Third Table Representation Learning Workshop*. https://openreview.net/forum?id=3I0bVvUj25

[20] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. 2023. A Survey on Vision Transformer. *IEEE Trans. on Pattern Analysis & Machine Intelligence* 45, 1 (2023), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

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

[22] Noah Hollmann, 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* (09 01 2025). https://doi.org/10.1038/s41586-024-08328-6

[23] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. 2025. The Tabular Foundation Model TabPFN Outperforms Specialized Time Series Forecasting Models Based on Simple Features. https://doi.org/10.48550/ARXIV.2501.02945

[24] Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. 2020. Early prediction of circulatory failure in the intensive

care unit using machine learning. *Nature Medicine* 26, 3 (March 2020), 364–373. https://doi.org/10.1038/s41591-020-0789-4

[25] Gartner Inc. 2024. *Data Quality: Best Practices for Accurate Insights*. Gartner Inc. Retrieved April 2025 from https://www.gartner.com/en/data-analytics/topics/data-quality

[26] Bojan Karlaš, Babak Salimi, and Sebastian Schelter. 2024. Navigating Data Errors in Machine Learning Pipelines: Identify, Debug, and Learn. In *Proceedings of SIGMOD'25*. Association for Computing Machinery, New York, NY, USA.

[27] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *NeurIPS*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[28] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. *ACM Comput. Surv.* 54, 10s, Article 200 (Sept. 2022), 41 pages. https://doi.org/10.1145/3505244

[29] Sadegh Khanmohammadi, Miguel G. Cruz, Daniel D.B. Perrakis, Martin E. Alexander, and Mehrdad Arashpour. 2024. Using AutoML and generative AI to predict the type of wildfire propagation in Canadian conifer forests. *Ecological Informatics* 82 (2024), 102711. https://doi.org/10.1016/j.ecoinf.2024.102711

[30] Boris Kovalerchuk and Evgenii Vityaev. 2005. *Data mining in finance: advances in relational and hybrid methods*. Springer Science & Business Media.

[31] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. 2021. CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. 13–24. https://doi.org/10.1109/ICDE51399.2021.00009

[32] L. Magadán, J. Roldán-Gómez, J. C. Granda, and F. J. Suárez. 2023. Early Fault Classification in Rotating Machinery With Limited Data Using TabPFN. *IEEE Sensors Journal* 23, 24 (2023), 30960–30970. https://doi.org/10.1109/JSEN.2023.3331100

[33] Sedir Mohammed, Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2025. The effects of data quality on machine learning performance on tabular data. *Information Systems* 132 (July 2025), 102549. https://doi.org/10.1016/j.is.2025.102549

[34] Sedir Mohammed, Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2025. The effects of data quality on machine learning performance on tabular data. *Information Systems* 132 (July 2025), 102549. https://doi.org/10.1016/j.is.2025.102549

[35] Sedir Mohammed, Lisa Ehrlinger, Hazar Harmouch, Felix Naumann, and Divesh Srivastava. 2025. The Five Facets of Data Quality Assessment. *SIGMOD Record* 54, 2 (2025), 1–10.

[36] Sedir Mohammed, Felix Naumann, and Hazar Harmouch. 2025. Step-by-Step Data Cleaning Recommendations to Improve ML Prediction Accuracy. (2025). https://doi.org/10.48786/EDBT.2025.43

[37] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. 2022. Transformers Can Do Bayesian Inference. In *International Conference on Learning Representations*. https://openreview.net/forum?id=KSugKcbNf9

[38] Ryunosuke Noda, Daisuke Ichikawa, and Yugo Shibagaki. 2024. Machine learning-based diagnostic prediction of minimal change disease: model development study. *Scientific Reports* 14, 1 (Oct. 2024). https://doi.org/10.1038/s41598-024-73898-4

[39] Fabian Offensperger, Gary Tin, Miquel Duran-Frigola, Elisa Hahn, Sarah Dobner, Christopher W. am Ende, Joseph W. Strohbach, Andrea Rukavina, Vincenth Brennsteiner, Kevin Ogilvie, Nara Marella, Katharina Kladnik, Rodolfo Ciuffa, Jaimeen D. Majmudar, S. Denise Field, Ariel Bensimon, Luca Ferrari, Evandro Ferrada, Amanda Ng, Zhechun Zhang, Gianluca Degliesposti, Andras Boeszoermenyi, Sascha Martens, Robert Stanton, André C. Müller, J. Thomas Hannich, David Hepworth, Giulio Superti-Furga, Stefan Kubicek, Monica Schenone, and Georg E. Winter. 2024. Large-scale chemoproteomics expedites ligand discovery and predicts ligand behavior in cells. *Science* 384, 6694 (2024), eadk5864. https://doi.org/10.1126/science.adk5864 arXiv:https://www.science.org/doi/pdf/10.1126/science.adk5864

[40] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 https://arxiv.org/abs/2303.08774

[41] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

[42] TabPFN GitHub Repository (Python). 2025. TabPFN GitHub Repository (Python). https://github.com/PriorLabs/TabPFN.

[43] Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. 2025. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=0VvD1PmNzM

[44] TabPFN GitHub Repository (R). 2025. TabPFN GitHub Repository (R). https://github.com/PriorLabs/R-tabpfn.

[45] TabPFN Client GitHub Repository. 2025. TabPFN Client GitHub Repository. https://github.com/PriorLabs/tabpfn-client.

[46] Sebastian Schelter, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert, and Gyuri Szarvas. 2015. On challenges in machine learning model management. *IEEE Data Engineering Bulletin* (2015). https://www.amazon.science/publications/on-challenges-in-machine-learning-model-management

[47] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating Large-Scale Data Quality Verification. *Proceedings of the VLDB Endowment* 11, 12 (2018), 1781–1794.

[48] Sebastian Schelter, Tammo Rukat, and Felix Biessmann. 2021. JENGA - A Framework to Study the Impact of Data Errors on the Predictions of Machine Learning Models. (2021). https://doi.org/10.5441/002/EDBT.2021.63

[49] Christoph Schinninger, Fabian Panse, Constantin Kühne, and Lisa Ehrlinger. 2025. Icewafl: A Configurable Data Stream Polluter. https://doi.org/10.48786/EDBT.2025.64

[50] Hassan Shahmohammadi, Maria Heitmeier, Elnaz Shafaei-Bajestan, Hendrik P. A. Lensch, and R. Harald Baayen. 2023. Language with vision: A study on grounded word and sentence embeddings. *Behavior Research Methods* 56, 6 (Dec. 2023), 5622–5646. https://doi.org/10.3758/s13428-023-02294-z

[51] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (May 2022), 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

[52] Vinh Quang Tran and Haewon Byeon. 2024. Predicting dementia in Parkinson's disease on a small tabular dataset using hybrid LightGBM–TabPFN and SHAP. *DIGITAL HEALTH* 10 (Jan. 2024). https://doi.org/10.1177/20552076241272585

[53] Boris Van Breugel and Mihaela Van Der Schaar. 2024. Position: Why Tabular Foundation Models Should Be a Research Priority. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 235. PMLR, 48976–48993. https://proceedings.mlr.press/v235/van-breugel24a.html

[54] Songkang Wen, Vasilii Feofanov, and Jianfeng Zhang. 2024. Measuring Pretraining Data Quality without Labels for Time Series Foundation Models. In *NeurIPS Workshop on Time Series in the Age of Large Models*. https://openreview.net/forum?id=nBtV1A3PrR

[55] Chao Ye, Guoshan Lu, Haobo Wang, Liyao Li, Sai Wu, Gang Chen, and Junbo Zhao. 2024. Towards Cross-Table Masked Pretraining for Web Data Mining. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 4449–4459. https://doi.org/10.1145/3589334.3645707

[56] Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. 2025. A Closer Look at TabPFN v2: Strength, Limitation, and Extension. https://doi.org/10.48550/ARXIV.2502.17361

[57] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. https://doi.org/10.48550/ARXIV.2303.18223