STRUGGLETAB: A CHALLENGING BENCHMARK FOR TABULAR SYNTHESIS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While recent tabular synthesis models have achieved remarkable performance, the standard framework for evaluating these models has remained too simple. In fact, the use of overly-sanitized evaluation datasets has created a misleading appearance that synthetic tabular data is nearly at parity with real-world data. This misguided impression quickly falls apart when models are evaluated on more challenging datasets. We introduce the *StruggleTab benchmark*, comprising seven datasets containing three challenging characteristics common in real-world data: (1) text, (2) feature imbalance and (3) strict invariants. Our audit of 5 leading tabular synthesis models (TabDLM, Tabby, TabDiff, ICL, CTGANP) on *StruggleTab evaluation metrics* reveals many failure modes including high dimensionality, sequence lengths and more. Concerning results include easy distinguishability from non-synthetics (average 85% score on C2ST) and all 5 models receiving a negative R^2 score in the standard downstream utility metric (Machine Learning Efficacy) on the financial Limit Order Book task. StruggleTab raises the empirical standards for tabular synthesis, highlighting the need for architectures compatible with real-world heterogeneity.

1 INTRODUCTION

Real-world tabular datasets are plagued by limited sample sizes, inherent biases, missing values and strict privacy limitations. To overcome these compounding bottlenecks, *tabular data synthesis* has emerged as a foundational research priority (van Breugel & van der Schaar, 2024), promising robust data augmentation, missing value imputation, bias adjustment and privacy preservation, as well as broader tabular data understanding.

Driven by these demands, a recent wave of tabular synthesis architectures has demonstrated remarkable progress. For example, TabDLM (Cai et al., 2026) achieves state-of-the-art fidelity through a hybrid transformer-diffusion approach, while Shi et al. (2025a) develop a method for joint diffusion modeling of numerical and categorical features. Further, Cromp et al. (2025) use a modified language model architecture to demonstrate parity on Machine Learning Efficacy (MLE) with non-synthetic data on 5/8 evaluated datasets.

These results contribute to a general impression that tabular data synthesis is nearly a solved problem. However, **this apparent success is driven by heavily sanitized evaluation tasks**, which overwhelmingly rely on simplified datasets restricted to continuous numerical values and low-cardinality categorical variables. When deployed on real-world data featuring unstructured text features, severe class imbalances or strict physical invariants, state-of-the-art models fail catastrophically. Evaluating architectures exclusively on simplified data creates a false ceiling of parity, masking fundamental flaws in current methods.

To draw attention towards this critical gap, we introduce **StruggleTab: A rigorous benchmark of seven challenging datasets sourced from diverse domains**, including financial limit order books, medical records and cybersecurity logs. Through comprehensive evaluations of five leading models, we demonstrate that current tabular synthesis approaches are ill-equipped to handle real-world heterogeneity, often failing to capture underlying invariants or produce usable data. StruggleTab is designed to draw the attention of the tabular synthesis community towards the large roadblocks that remain ahead. To this end, our contributions include:

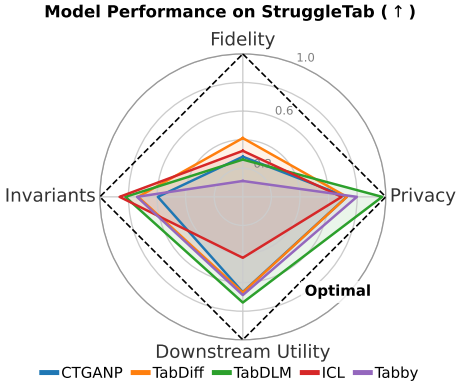

Figure 1: Model Evaluation Performance on StruggleTab *(Higher Better)*; Details App. D

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

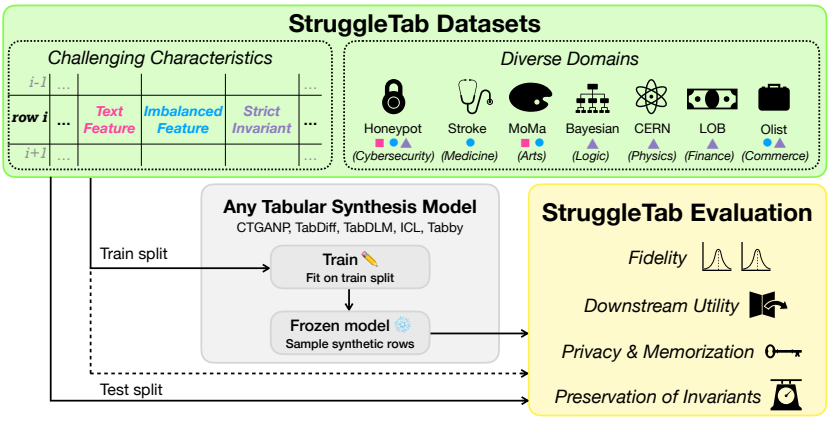

Figure 2: Overview of the Tabular Synthesis Evaluation Process, with our novel *StruggleTab benchmark*. Using the 7 StruggleTab benchmark datasets, selected for their challenging attributes (text, feature imbalance and strict invariants), a tabular synthesis model is trained on the train split. Resulting synthetic data is evaluated against the testset across Fidelity, Downstream Utility, Privacy and Invariant Preservation, using StruggleTab evaluation metrics compatible with these datasets.

- **StruggleTab Benchmark Datasets:** Seven complex datasets specifically curated to break current synthesis models using real-world attributes like freeform text and extreme long-tail distributions.
- **StruggleTab Benchmark Evaluation Metrics:** Formulations of tabular evaluation metrics (e.g., Machine Learning Efficacy, Distance to Closest Record) adapted to natively handle complex data modalities, plus a collection of dataset-specific metrics to measure synthetics' adherence to physical and logical constraints.
- **A Unified Evaluation Harness:** A standardized, plug-and-play data handler enabling seamless and reproducible preprocessing, ingestion and evaluation across model pipelines.
- **Audit of SOTA Model Deficits:** Extensive evaluations of five state-of-the-art models, including TabDLM (Cai et al., 2026) and Tabby (Cromp et al., 2025), exposing critical failure modes, privacy leakages and inability to respect physical invariants.

Section 2 provides background on current standard evaluation methodologies and their limitations, while Section 3 details the curation of the StruggleTab benchmark. Section 4 presents our comprehensive audit of the failure modes present in current SOTA models. Section 5 discusses findings and outlines necessary shifts for the tabular synthesis community, followed by prior work in Section 6 and concluding remarks in Section 7.

2 Background: Preexisting Tabular Synthesis Evaluations

Typical tabular synthesis evaluation involves splitting a real tabular dataset into train and test sets, training the synthesis model on the trainset, then comparing the held-out testset to synthetic data produced by the model across a variety of metrics. We now discuss common datasets and metrics.

Typical Datasets: Over time, a small handful of tabular datasets have become the accepted standards for synthesis evaluations. Appendix A lists these commonly-accepted datasets, along with the set of data types represented within each dataset. These datasets are heavily biased in favor of numerical and low-cardinality categorical features and do not allow models to demonstrate their capabilities on the messier, highly multimodal datasets common in the real world.

Standard Metrics: The reliance on simple evaluation datasets has also enabled the use of evaluation metrics that assume clean, limited-modality datasets. A standard tabular evaluation pipeline typically employs a suite of metrics to assess generative quality across three priorities:

- *Fidelity:* Measures the capability of a discriminative model to distinguish between synthetic and real samples. The *Classifier Two-Sample Test* (Lopez-Paz & Oquab, 2016) trains a discriminator to distinguish between real and synthetic rows. An AUC of 0.5 indicates indistinguishable data, while 1.0 indicates complete distributional separation.
- *Downstream Utility:* The metric of *Machine Learning Efficacy (MLE)* (Xu et al., 2019) measures whether the synthesis model preserves underlying predictive signals. A standard ML model (e.g., random forest) is trained entirely on synthetic data, then evaluated on a held-out real test set, reporting accuracy or RMSE.
- *Privacy and Memorization:* Models must generate novel data, rather than leak training examples. *Distance to Closest Record (DCR)* (Zhao et al., 2021) calculates the distance between a synthetic

Table 1: Summary of Datasets and Associated Challenges in the StruggleTab Benchmark. Details in Appendix B.
*Features represented: *Categorical (C), Date/Time (D), Numerical (N), String (S), JSON (J).*

	Honeypot	**Stroke**	**MoMA**	**Bayesian**	**CERN**	**LOB**	**Olist**
Text Feat. (■)	✓		✓				
Imbalance (●)	✓	✓	✓				✓
Invariant (▲)	✓			✓	✓	✓	✓
Feat. Types*	S, J, N, C	C, N	S, D, N, C	N, C	C, N	N	D, C, N
# Features	5	11	25	6	17	21	10
# Rows	7199	5110	12962	100000	99915	17108	113314
Source	Sconzo (2014)	Palacios (2020)	MoMA (2026)	Synthetic	McCauley (2014)	Nielsen (2021)	Olist & Sionek (2018)

row and its nearest trainset point, but lacks clear thresholds to separate between low-score memorization, middle-score healthy generalization and high-score spurious noise. *Membership Inference Attacks (MIAs)* (Shokri et al., 2017) test whether an adversary model can infer if a specific record was a member of the trainset.

While these metrics provide a solid baseline for numerical and low-cardinality categorical data, several of the standard metrics do not immediately translate to multimodal tabular data. For instance, tree-based MLE evaluators (Scikit-Learn models (Pedregosa et al., 2011) are the typical evaluators) ignore zero-variance features resulting from a model's failure to generate text features, while string embeddings add sparsity to distance-based metrics like DCR, resulting in outsized influence for string features on the metric compared to low-dimensional numerical and categorical features.

3 The StruggleTab Benchmark

We now describe StruggleTab in greater detail. We introduce the benchmark's datasets in Section 3.1 and metrics in Section 3.2. Lastly, we discuss the convenience and reproducibility of StruggleTab for those developing novel tabular synthesis models in Section 3.3.

3.1 StruggleTab Benchmark Datasets

Summarized in Table 1, StruggleTab contains seven datasets challenging to contemporary tabular synthesis approaches. Three criteria were employed in the selection of these datasets:

1. **Presence of Three Challenging Attributes:** We select three challenging aspects to target with our benchmark; these are the presence of *(■) a freeform text feature, (●) a heavily imbalanced feature and (▲) strict, latent logical, mathematical or physical invariants*. Each dataset must contain at least one challenging attribute.
2. **Variety:** To ensure diversity in the benchmark, we prioritize datasets from various domains, of varying sizes and containing a mix of data types. Six datasets are from real-world domains, while one (Bayesian) is synthetic, to control for the possibility of pretraining leakage.
3. **Open Access:** We ensure that all datasets in StruggleTab are free and convenient to obtain.

Details on each dataset, its reason for inclusion and sample datapoints are available in Appendix B.

3.2 StruggleTab Benchmark Evaluation Metrics

To assess the capabilities of tabular synthesis approaches, we modify the standard metrics introduced in Section 2 for compatibility with the multimodal StruggleTab data, evaluating across four axes:

- **Fidelity:** To measure structural and semantic realism simultaneously, we employ the Classifier Two-Sample Test (C2ST). We project all features into a unified vector space (Appendix D.2) and train a random forest discriminator to distinguish real from synthetic rows, reporting AUC.
- **Downstream Utility:** We evaluate Machine Learning Efficacy (MLE) to determine whether the synthesis model preserves predictive signals. Using the standard Train-on-Synthetic-Test-on-Real (TSTR) process (Esteban et al., 2017), we report R^2 for regression tasks and Macro F1-score for classification tasks. Macro F1 penalizes models that collapse to the most common target class on imbalanced datasets (e.g., Stroke, MoMA). We lower-bound R^2 at 0.00 (equivalent to naive mean prediction); unbounded results in Appendix E. For downstream model design, see Appendix D.2.

- **Privacy and Memorization:** Because heterogeneous tables contain both continuous manifolds and freeform strings, a single metric leaves blind spots. We opt for three metrics. First, to detect leakage involving text features, we use *Membership Inference Attacks (MIAs)*, reporting the worst-case AUC across a battery of likelihood-based attacks. Second, we compute the Euclidean *Distance to Closest Record (DCR)* in a unified vector space. To address DCR's interpretability issues, we report the DCR of the held-out real test set as the optimal target for ideal generalization. To accompany DCR, Appendix E also provides a *Match Rate* metric measuring complete duplication of trainset rows (rows with DCR of 0.0).
- **Domain Constraints and Invariants:** We introduce dataset-specific validation functions to report the *Constraint Violation Rate*, measuring how often a model breaks deterministic rules such as temporal causality (Olist), non-negative bounds (LOB) and mass-energy conservation (CERN).

Further details and implementation considerations for these metrics are available in Appendix D.2.

3.3 Reproducibility and Ease of Use

To facilitate adoption, the StruggleTab benchmark provides a lightweight, plug-and-play evaluation harness. Requiring only $\sim$4 lines of Python code to integrate into standard training and sampling pipelines as demonstrated in Appendix C, design was inspired by the SynthCity harness of legacy datasets and metrics (Qian et al., 2023). All seven StruggleTab datasets are openly accessible under permissive licenses (See Appendix B), enabling fast and programmatic retrieval and configuration. Our codebase will be open-sourced upon publication.

4 Evaluations

We now evaluate tabular synthesis models on StruggleTab benchmark. The models are introduced in Section 4.1; then, we evaluate the models along the four criteria outlined in Section 3.2:

§4.2 **Fidelity**	Evaluating divergence via Classifier Two-Sample Tests.
§4.3 **Downstream Utility**	Preserving predictive signal via Machine Learning Efficacy.
§4.4 **Privacy & Memorization**	Exposing memorization via MIA and Distance to Closest Record.
§4.5 **Domain Constraints**	Analyzing failure modes across physical and logical constraints.

4.1 Setup

We evaluate five popular and state-of-the-art models on StruggleTab, summarized in Table 2. These models represent the most common tabular synthesis architectures: namely, diffusion models, transformers and generative adversarial networks (GANs). More details on each model are provided in Appendix D. For each model, we conduct a hyperparameter search, also discussed in Appendix D. All results are averaged across three random seeds; tables report average score $\pm$ standard deviation.

The "String Compatible" row of Table 2 indicates whether a model is compatible with string features. If a model is not string-compatible, all string features in its synthetics are set to empty (""). We represent JSON objects as strings. Date/time features are treated as strings for string-compatible models and numerical Linux epochs otherwise. We note that because 3/5 features in the Honeypot dataset must be represented as strings, we do not evaluate models incompatible with strings (i.e., CTGANP and TabDiff) on Honeypot.

Performance Profiles: Sections 4.2 and 4.3 contain Performance Profiles (Dolan & Moré, 2002) as Figures 3 and 4. These plots visualize cumulative model performance across all datasets. The y-axis, $\rho_m(\tau)$, denotes the fraction of datasets for which model m performs within a factor τ of the best-scoring model for that dataset. The x-axis, plotted as $\log(\tau)$, represents the performance ratio of a given model relative to that best-performing model; $\log(\tau) = 0$ indicates that the model achieved the absolute best score. When reading these profiles, **curves that are higher and further to the left represent superior performance**. Specifically, a model's y-intercept shows the fraction of datasets where it is the definitive top performer, while its maximum height on the right side of the plot indicates its overall robustness (the total fraction of datasets it successfully processed). Details in Appendix D.

Table 2: Models evaluated in Section 4. See Appendix D.1 for more details on model selection.

	Tabby	CTGANP	ICL	TabDiff	TabDLM
Architecture	LLM	GAN	LLM	Diffusion	Diffusion
String Compatible	✓	×	✓	×	✓
Source	Cromp et al. (2025)	Zhao et al. (2024)	Brown et al. (2020)	Shi et al. (2025b)	Cai et al. (2026)

Table 3: Data Fidelity Evaluated using AUC for the Classifier 2-Sample Test (C2ST; see Sec. 3.2 for definition). **Closest to 0.50 is best.** Text ■, Invariant ▲, Imbalance ●. * *Incompatible with string features; synthetic strings set to "".*

	Honeypot ■●▲	Stroke ●	MoMA ■●	Bayesian ▲	CERN ▲	LOB ▲	Olist ●▲
CTGANP*	-	0.74±0.03	1.00±0.00	0.58±0.01	0.90±0.02	1.00±0.00	0.82±0.01
TabDiff*	-	**0.59±0.03**	1.00±0.00	**0.48±0.03**	1.00±0.00	0.99±0.00	**0.54±0.00**
TabDLM	0.98±0.01	0.74±0.00	0.98±0.02	0.91±0.04	1.00±0.00	1.00±0.00	0.56±0.02
ICL	**0.85±0.13**	0.70±0.02	1.00±0.00	1.00±0.00	**0.89±0.03**	**0.90±0.01**	0.65±0.01
Tabby	0.94±0.00	1.00±0.00	**0.94±0.02**	0.99±0.00	1.00±0.00	1.00±0.00	0.92±0.01

4.2 Results: Poor Fidelity

Setup: To assess the fidelity of synthetic data across the StruggleTab datasets, we perform the Classifier Two-Sample Test (C2ST) (details in Sec. 3.2 and App. D.2). A random forest discriminator is trained to distinguish between real and synthetic rows. An Area Under the ROC Curve (AUC) of 0.50 indicates the synthetic data is indistinguishable from non-synthetic data; 1.00 indicates complete distinguishability. We report the worst-case C2ST AUC in Table 3 (raw accuracy in App. E).

Results: Table 3 and Figure 3 demonstrate severe architectural trade-offs; no single model universally achieves high fidelity. Optimized for continuous latent spaces, *diffusion models (TabDiff, TabDLM) dominate datasets containing standard numerical and categorical features*, leading on Stroke (0.59) and Olist (0.54). However, this capability breaks down on text features (■). TabDiff completely drops strings, while TabDLM's text-serialization yields high distributional separation (0.98 on Honeypot and MoMA).

On Bayesian, TabDiff achieves a C2ST AUC close to chance (0.48 ± 0.03), indicating limited distinguishability between real and synthetic samples under the evaluated discriminator. Further, *while the ICL model performs relatively strongly in Table 3, privacy metrics indicate that this is at least partially due to leakage of its prompt examples*, particularly on Honeypot and MoMA (see Table 5), Stroke and Olist (Table 6) and CERN (Table 25).

Excluding architectures reliant on memorization, *Tabby is the best baseline for text-heavy datasets*. Its AUC of 0.94 on Honeypot and MoMA remains highly distinguishable, while other models omit strings entirely (e.g., TabDiff) or rely on memorization (e.g., ICL), indicating the importance of futher progress on text-heavy tabular synthesis. Conversely, Tabby struggles on strict numerical relationships (1.00 on CERN, LOB).

Key Takeaways:

- No one model dominates on data fidelity: **a model equally capable on both textual and numerical fronts does not currently exist**.
- Transformer-based models are currently the most reliable for text-heavy datasets, while Diffusion-based models remain the best option for complex numerical manifolds.
- High fidelity scores should be carefully interrogated for signs of trainset or prompt leakage.

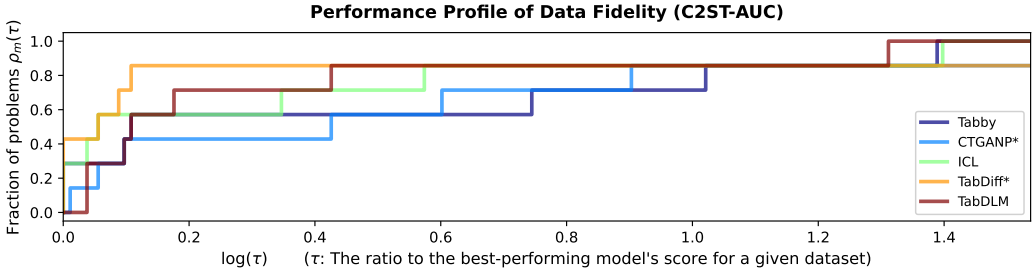

Figure 3: Performance Profile (Definition in Sec. 4.1) of C2ST AUC. Higher and left-shifted curves denote greater fidelity. No single model is universally reliable; details and caveats in Sec. 4.2.

Table 4: Downstream ML Efficacy. We report Macro F1-score for classification tasks (Honeypot, Stroke, MoMA, Bayesian) and R^2 for regression tasks (CERN, LOB, Olist). **Higher is better.** Text ■, Invariant ▲, Imbalance ●. ** Incompatible with string features; synthetic strings are set to "" in all rows.*

	Honeypot ■●▲	**Stroke** ●	**MoMA** ■●	**Bayesian** ▲	**CERN** ▲	**LOB** ▲	**Olist** ●▲
CTGANP*	-	0.49±0.01	0.16±0.01	0.88±0.00	0.67±0.02	**0.00±0.00**	0.08±0.02
TabDiff*	-	0.49±0.00	**0.56±0.05**	**0.89±0.00**	**0.75±0.02**	**0.00±0.00**	0.07±0.01
TabDLM	**0.23±0.09**	0.49±0.00	0.28±0.09	0.78±0.09	0.10±0.02	**0.00±0.00**	0.06±0.01
ICL	0.12±0.02	**0.50±0.02**	0.01±0.00	0.00±0.00	0.40±0.04	**0.00±0.00**	**0.11±0.02**
Tabby	0.11±0.00	0.49±0.00	0.51±0.00	0.87±0.01	0.04±0.07	**0.00±0.00**	0.02±0.02
Baseline	*0.72*	*0.55*	*0.71*	*0.89*	*0.75*	*9.23e-2*	*0.11*

4.3 Results: Limited Downstream Utility

Setup: To determine whether synthetic datapoints preserve the predictive signals of the non-synthetic data, we evaluate Machine Learning Efficacy (MLE) using the Train-on-Synthetic-Test-on-Real (TSTR) process. A standard ML algorithm (e.g., Random Forest) is trained entirely on the synthetic dataset and evaluated on a held-out real test set to predict a specified target column, as listed in Table 21. We report Macro-F1 for classification tasks (Honeypot, Stroke, MoMA, Bayesian) and the Coefficient of Determination (R^2) for regression tasks (CERN, LOB, Olist) in Table 4. We lower-bound R^2 at 0.00 (equivalent to a naive mean-prediction baseline); unbounded R^2 in App. E. *Baseline* is the performance of a downstream model trained on the non-synthetic trainset.

Results: Table 4 and Figure 4 reveal two phenomena: universal collapse on high-precision continuous data and the contrast between utility and data fidelity (Section 4.2).

First, all models score $\sim$ 0.00 on LOB. While predicting limit order outcomes is inherently noisy (the *Baseline* achieves only a 9.23e-2 R^2), synthetic models completely collapse on these microscopic signals. Continuous architectures (TabDiff, CTGANP) smooth away minute predictive details, while text-native architectures suffer from numerical token hallucination. *This failure on high-precision continuous data highlights an unsolved frontier for tabular synthesis.*

Second, downstream utility does not imply high data fidelity. TabDiff achieves 0.56 on MoMA (against a 0.71 baseline) using only numeric features, despite its worst-possible 1.00 C2ST fidelity score. Only Tabby competitively synthesizes MoMA's strings (scoring 0.51) while retaining signal, validating language models for text-heavy tables. Finally, legacy MLE evaluations frequently report unweighted Accuracy, which actively hides majority-class collapse. On the highly imbalanced Stroke dataset, models can trivially score $\sim$ 0.95 Accuracy simply by predicting the majority class, a failure exposed here by their $\sim$ 0.49 Macro-F1 scores.

Key Takeaways:
- Models struggle to maintain predictive signal in synthetic data, particularly in **higher-dimensional numerical domains** like LOB.
- Assessments of synthetic data quality require the **comparison of multiple metrics**. Standard unweighted metrics (like classification accuracy) can mask model collapse on highly imbalanced datasets.

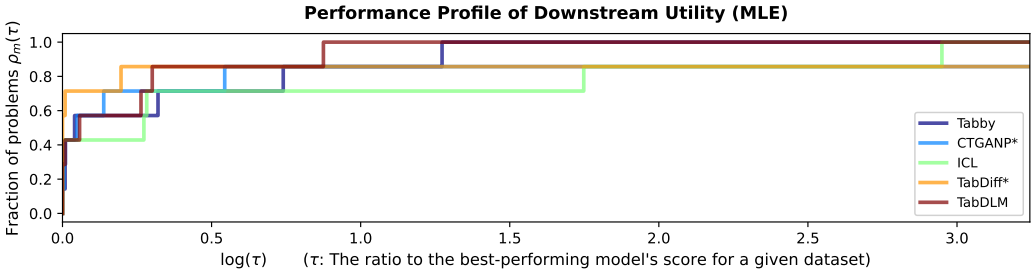

Figure 4: Performance Profile (Definition in Sec. 4.1) of MLE. Higher and left-shifted curves denote greater utility. Some datasets pose a struggle for all model types. Details and caveats in Sec. 4.3.

Table 5: Worst-Case Membership Inference Attack (MIA) AUC. **Values closer to 50.0% indicate optimal privacy preservation** (random adversary guessing). Text ■, Invariant ▲, Imbalance ●. ** Incompatible with string features; synthetic strings are set to "" in all rows.*

	Honeypot ■●▲	Stroke ●	MoMA ■●	Bayesian ▲	CERN ▲	LOB ▲	Olist ●▲
CTGANP*	-	53.1±0.7	53.7±1.1	53.1±0.7	52.7±0.8	51.6±0.8	53.1±0.6
TabDiff*	-	**52.4±1.6**	52.2±0.9	52.7±0.4	**52.6±0.4**	52.1±0.8	52.3±0.9
TabDLM	**52.3±0.3**	52.9±0.2	53.2±0.7	54.9±0.9	**52.6±0.7**	**50.9±0.3**	**52.0±0.5**
ICL	68.1±7.8	53.2±0.6	88.0±20.8	**51.0±0.5**	53.9±0.9	52.7±1.6	52.5±0.4
Tabby	52.8±1.9	53.5±1.4	**51.7±0.3**	52.1±0.4	53.7±0.8	51.9±0.7	57.8±7.8

Table 6: Distance to Closest Record (DCR). *Baseline* row represents the DCR of the held-out test set, representing ideal generalization without memorization. **Closest to *Baseline* is best.** Text ■, Invariant ▲, Imbalance ●. ** Incompatible with string features; synthetic strings set to "" in all rows.*

	Honeypot ■●▲	Stroke ●	MoMA ■●	Bayesian ▲	CERN ▲	LOB ▲	Olist ●▲
CTGANP*	-	0.65±0.03	0.14±0.09	0.13±0.01	**1.07±0.01**	1.53±0.07	**0.75±0.05**
TabDiff*	-	**0.62±0.02**	0.65±0.07	**0.11±0.00**	1.56±0.01	**1.59±0.06**	1.36±0.02
TabDLM	**0.31±0.07**	0.74±0.01	2.13±0.54	0.59±0.40	2.76±0.22	1.53±0.04	1.54±0.01
ICL	0.63±0.26	0.19±0.01	1.95±0.08	1.17±0.00	1.22±0.01	5.86±1.27	0.16±0.01
Tabby	0.09±0.01	1.47±0.05	**0.96±0.01**	0.07±0.00	1.51±0.01	11.67±1.73	0.24±0.07
Baseline	*0.44*	*0.55*	*0.93*	*0.11*	*0.75*	*2.13*	*1.02*

4.4 Results: Memorization and Privacy Leakage

Setup: Because StruggleTab contains a mix of datatypes, we rely on a two-pronged privacy audit. First, to detect unstructured text memorization, we conduct Membership Inference Attacks (MIAs), reporting the worst-case adversarial AUC in Table 5. A score of 50.0% represents random adversarial guessing (ideal privacy preservation). Second, to detect exact-match geometric memorization, we calculate the Distance to Closest Record (DCR) in a unified vector space, baselining the synthetic distances against the held-out real test set (results in Table 6). Metric details in Appendix D.2.

Results: Each metric exposes a distinct failure mode. Namely, the MIAs reveal memorization in language based models and text leakage in all text-compatible models, while DCR indicates structural copy-pasting in other architectures and datatypes.

Textual Leakage (MIAs): Treating tabular generation as a language modeling task introduces unique privacy vulnerabilities (Table 5). By serializing rows into text tokens, LLM-based generators become highly susceptible to token-level Membership Inference Attacks (MIAs). Consequently, the ICL baseline suffers severe prompt leakage on text-heavy datasets like MoMA (88.0%) and Honeypot (68.1%). This vulnerability extends beyond freeform text: Tabby exhibits significant leakage on Olist (57.8%) by memorizing the textualized representations of numerical and categorical features. In contrast, TabDiff and CTGANP obtain MIA AUCs close to the 50.0% random-guessing baseline, indicating limited membership distinguishability under the evaluated attacks.

Geometric Memorization (DCR): While MIA detects text leakage, DCR exposes geometric memorization in numerical and categorical features. On LOB, all continuous models (CTGANP, TabDiff, TabDLM) score between 1.53 and 1.59 (Table 6), falling significantly below the 2.13 testset baseline. These results indicate greater proximity to training records, although the reported exact-match rates on LOB are zero for these models (Table 25). Conversely, LLMs (ICL, Tabby) score anomalously high on LOB, again indicating their inability to generate realistic continuous LOB datapoints. Interestingly, low DCR does not always imply memorization. CTGANP's low MoMA score (0.14) results from modality incompatibility: inability to generate text yields empty strings, artificially minimizing distances to sparse training records.

We highlight the importance of including the "Baseline" of the held-out test set DCR. Because natural record distances vary drastically by dataset (e.g., 0.11 for Bayesian versus 2.13 for LOB), the practice of applying a universal "lower is better" or "higher is better" DCR threshold without a non-synthetically determined optimum is flawed.

Key Takeaways:
- **Depending on datatype and model architecture, trainset leakage manifests differently**.
- String-based metrics (MIAs) are most effective for detecting leakage in transformer models and/or for string features, while DCR is most effective for other features and architectures.
- DCR evaluation requires comparison to a non-synthetic baseline for interpretation.

Table 7: Domain Constraint Violation Rate. **Lower is better.** Text ■, Invariant ▲, Imbalance ●.
* *Incompatible with string features; synthetic strings are set to "" in all rows.*

	Honeypot ■●▲	**Bayesian** ▲	**CERN** ▲	**LOB** ▲	**Olist** ●▲
CTGANP*	-	0.06±0.05	0.99±0.00	**0.00±0.00**	0.00±0.00
TabDiff*	-	0.00±0.00	0.65±0.07	**0.00±0.00**	0.02±0.00
TabDLM	0.01±0.01	0.57±0.49	1.00±0.00	**0.00±0.00**	0.04±0.01
ICL	0.00±0.00	0.02±0.00	**0.01±0.00**	0.81±0.01	0.02±0.00
Tabby	**0.00±0.00**	**0.00±0.00**	0.99±0.00	0.99±0.00	**0.00±0.00**

4.5 Results: Unreliability for Physical Invariants

Setup: We evaluate the Constraint Violation Rate, measuring how often a model breaks deterministic invariants (▲). Examples include temporal causality (e.g., packages delivered before they are purchased in Olist), non-negative transaction volumes in Limit Order Books (LOB), or mass-energy conservation bounds in particle collisions (CERN). See Appendix B.2 for more information.

Results: Surprisingly, fidelity and downstream utility do not guarantee that a model has learned the strict physical or logical rules governing a dataset. In fact, despite our hyperparameter optimization pipeline explicitly penalizing constraint violations during validation (giving models the best possible chance to succeed; see Appendix D), Table 7 reveals that *constraint adherence is largely dictated by architectural compatibility with the constraint's modality.*

Token-based LLMs (Tabby, ICL) successfully respect textual and structural logic, achieving perfect 0.00 violation rates on Honeypot's JSON payloads. However, these text-native architectures fail on continuous numerical boundaries, with Tabby and ICL scoring 0.99 and 0.81 respectively on LOB. Continuous architectures (TabDiff, CTGANP) enforce LOB's strict limits to score 0.00, though Table 6 indicates this is partly aided by memorization. Meanwhile, these models struggle with exact mass-energy physics (CERN), where continuous variance smoothing destroys exact conservation bounds. Interestingly, the ICL baseline maintains adherence on CERN (0.01); because ICL is prompted with 5 mathematically valid real records, it partly bypasses actual constraint learning by simply blending together the prompt examples (indicated in Table 25).

While most models are relatively successful with the logical constraints present in the Bayesian dataset, scoring below 0.10, TabDLM struggles considerably (0.57). We hypothesize that TabDLM's hybrid architecture—which applies a continuous diffusion process over textually serialized numerical features—introduces destructive reconstruction noise. While LLMs (Tabby, ICL) also textualize numbers, they predict discrete tokens directly, allowing them to learn exact symbolic mappings. TabDLM, however, adds continuous Gaussian noise to discrete token embeddings. During denoising, a minor continuous error can snap to an entirely different digit character, destroying the precise algebraic relationships required to satisfy the causal collider's addition bounds.

Ultimately, we highlight a critical gap: *safe tabular synthesis will require architectures capable of enforcing deterministic mathematical boundaries across all modalities while authentically modeling real-world noise, all without resorting to exact-row memorization*.

Key Takeaways:
- While most models handle basic logic, **nearly all state-of-the-art architectures fail to enforce strict physics and financial invariants without demonstrable memorization**.
- Any critical synthetics application must still rely on external, deterministic verification.

5 Discussion

The evaluations across the StruggleTab benchmark reveal a bifurcation in current tabular synthesis capabilities: *architectures force a trade-off between generative fidelity and downstream utility*. As demonstrated in Section 4.2 and Section 4.3, models optimized for statistical fidelity often fail to preserve actionable predictive signals across multimodal data. Conversely, models can achieve seemingly competitive performance on some metrics while completely failing others, often by sidestepping complex modalities entirely (e.g., omitting text features) or relying on structural memorization. This decoupling highlights that optimizing for a single performance axis provides an incomplete, often misleading, picture of a model's true capabilities.

The failures throughout Section 4 expose a critical blind spot: *we cannot assume that generative capabilities on simple numerical datasets naturally extend to complex multimodal tables.* To ensure safe and actionable tabular synthesis, we offer two core recommendations:

- **For Practitioners:** Never assume a generative model has implicitly learned strict constraints such as mathematical or logical rules. Synthetic data must be actively screened through deterministic constraint filters prior to any downstream use.
- **For Researchers:** Single-metric evaluations leave dangerous blind spots. Diagnosing multimodal architectures requires pairing orthogonal metrics—such as coupling MIAs with geometric DCR—to simultaneously capture textual leakage, numerical mode collapse and exact-row memorization.

5.1 Limitations

While StruggleTab aims to rigorously evaluate tabular synthesis, we constrain its scale to ensure practical usability. Excessively large or complex benchmarks often inadvertently invite cherry-picking, where model developers evaluate only on convenient subsets of the tasks. Determining the benchmark's scope therefore required deliberate compromises. Accordingly, we choose to leave additional challenging attributes such as time series or relational databases, or evaluating a broader suite of models (as discussed in Appendix D.1) to future work. We believe the current scope of StruggleTab strikes a necessary balance: sufficiently challenging to expose the shortcomings of modern architectures, yet tractable enough that developers can easily adopt and run the entire benchmark.

6 Prior Work

Tabular Synthesis Models: Early tabular generators relied on GANs and VAEs like CTGAN and TVAE (Xu et al., 2019). Recently, the field has split into two paradigms. Diffusion models like TabDiff (Shi et al., 2025a) treat rows continuously, achieving high numerical fidelity but lacking native text support. Conversely, sequence models like GReaT and Tabby (Borisov et al., 2022; Cromp et al., 2025) serialize rows into text, leveraging transformers for semantic modeling at the expense of continuous precision. However, no current architecture balances multimodal distributions with strict physical constraints.

Evaluation and Benchmarking: Synthesis architectures have outpaced rigorous evaluation methods, with models historically benchmarked on sanitized, low-dimensional datasets (Asuncion et al., 2007, for example). Frameworks like the Synthetic Data Vault (SDV) (Patki et al., 2016) provide standardized metrics but heavily prioritize 1D marginal matching and tree-based MLE on simple numeric and categorical data. Newer frameworks like SynthCity (Qian et al., 2023) offer unified evaluation harnesses—inspiring StruggleTab's plug-and-play design—yet still rely on simplified datasets. While recent benchmarks target specific structural or modal properties, such as Jiang et al. (2026) evaluating causal independence via multi-target prediction or Zhang et al. (2026) testing text semantic alignment, they remain confined to statistical associations on sanitized, low-modality tables. StruggleTab evaluates multi-modal heterogeneity under deterministic real-world physical and logical invariants, exposing critical trade-offs between exact-row memorization and distributional over-sanitization that purely statistical or causal evaluations overlook.

Table Reasoning and Non-Synthesis Models: Parallel to synthesis research, table understanding and fact verification have advanced significantly (Chen et al., 2019). Models like TaPas (Herzig et al., 2020) adapt language models to extract information and perform aggregations over structured tables. However, discriminative table understanding remains fundamentally distinct from generative synthesis. Whereas reasoning tasks benefit from the bounded context of an established table, synthesis faces the unconstrained challenge of constructing valid joint distributions from scratch.

7 Conclusion

The advancement of tabular synthesis has been bottlenecked by sanitized, low-complexity evaluations. We introduce StruggleTab, a rigorous benchmark of 7 challenging datasets and set of accompanying evaluation metrics. Our audit of state-of-the-art models reveals that no current model totally solves tabular synthesis. StruggleTab offers rigorous evaluations to test the next generation of performant tabular synthesis models.

AI USE STATEMENT

In this work, we used generative AI tools to receive feedback on the paper, copyedit the paper for readability, to create starter code for generating plots and to find related works. We did not use generative AI tools for the required disclosure tasks of generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis or interpreting results. We have reviewed all AI-assisted work. For example, we proofread all LLM outputs for accuracy, verified and tested all LLM-generated code, performed a manual literature search and reviewed all cited papers ourselves. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure reproducibility. Our benchmark is designed for reproducibility, as stated in Section 3.3, through easy integration with model codebases and the selection of easy-to-access datasets. We provide exact URLs to access datasets in Table 9 and model implementations in Table 19. Our metrics implementations and the hyperparameter process for all model/dataset pairs are explained in Section D. Anonymized source code for StruggleTab is included with our submission and open source code will be released alongside the publication of the paper.

REFERENCES

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language Models are Realistic Tabular Data Generators. September 2022. URL `https://openreview.net/forum?id=cEygmQNOeI`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Donghong Cai, Jiarui Feng, Yanbo Wang, Da Zheng, Yixin Chen, and Muhan Zhang. Tabdlm: Free-form tabular data generation via joint numerical-language diffusion. *arXiv preprint arXiv:2602.22586*, 2026.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

Sonia Cromp, Satya Sai Srinath Namburi GNVV, Mohammed Alkhudhayri, Catherine Cao, Samuel Guo, Nicholas Roberts, and Frederic Sala. Tabby: A language model architecture for tabular and structured data synthesis. *arXiv preprint arXiv:2503.02152*, 2025.

Elizabeth D Dolan and Jorge J Moré. Benchmarking optimization software with performance profiles. *Mathematical programming*, 91(2):201–213, 2002.

Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4320–4333, 2020.

Xiangjian Jiang, Nikola Simidjievski, and Mateja Jamnik. Tabstruct: Measuring structural fidelity of tabular data, 2026. URL https://arxiv.org/abs/2509.11950.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database. *PhysioNet*, September 2016. doi: 10.13026/C2XW26. URL https://doi.org/10.13026/C2XW26. Version 1.4.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *International conference on artificial intelligence and statistics*, pp. 1288–1296. PMLR, 2024.

Michael Kahn. Diabetes. UCI Machine Learning Repository, 1994. URL https://www.openml.org/search?type=data&sort=runs&id=37&status=active. DOI: https://doi.org/10.24432/C5T59G.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pp. 17564–17579. PMLR, 2023.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Thomas McCauley. Events with two electrons from 2010, 2014. URL https://opendata.cern.ch/record/304.

Open Data Robot MoMA. Moma collection - automatic update, June 2026. URL https://zenodo.org/doi/10.5281/zenodo.20506252.

Martin Søgaard Nielsen. High frequency crypto limit order book data, 2021. URL https://www.kaggle.com/datasets/martinsn/high-frequency-crypto-limit-order-book-data.

Olist and André Sionek. Brazilian e-commerce public dataset by olist, 2018. URL https://www.kaggle.com/dsv/195341.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, 1997.

Federico Soriano Palacios. Stroke prediction dataset, 2020. URL https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pp. 399–410. IEEE, 2016.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023. URL https://arxiv.org/abs/2301.07573.

C. Sakar and Yomi Kastro. Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5F88Q.

Mike Sconzo. Security repo - honeypot dataset, 2014. URL https://secrepo.com/honeypot/honeypot.json.zip.

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. Tabdiff: a mixed-type diffusion model for tabular data generation. In *International Conference on Learning Representations*, volume 2025, pp. 37353–37375, 2025a.

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation, February 2025b. URL http://arxiv.org/abs/2410.20626. arXiv:2410.20626 [cs].

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Tejashvi. Tour & Travels Customer Churn Prediction, 2023. URL https://www.kaggle.com/datasets/tejashvi14/tour-travels-customer-churn-prediction.

Boris van Breugel and Mihaela van der Schaar. Why Tabular Foundation Models Should Be a Research Priority, May 2024. URL https://arxiv.org/abs/2405.01147v2.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *International Conference on Learning Representations*, volume 2024, pp. 52829–52857, 2024.

Tianping Zhang, Shaowen Wang, Shuicheng Yan, Li Jian, and Qian Liu. Generative table pre-training empowers models for tabular prediction. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 14836–14854, 2023.

Yuhao Zhang, Liang Yan, Shaoming Duan, Xinyu Zha, Jinhang Su, Peiyi Han, and Chuanyi Liu. AFT-tab: Adversarial fine-tuning for tabular data synthesis with long text columns. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4581–4594, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.209. URL https://aclanthology.org/2026.acl-long.209/.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian conference on machine learning*, pp. 97–112. PMLR, 2021.

Zilong Zhao, Robert Birke, and Lydia Chen. TabuLa: Harnessing Language Models for Tabular Data Synthesis, October 2023. URL http://arxiv.org/abs/2310.12746. arXiv:2310.12746 [cs] version: 1.

Zilong Zhao, Aditya Kunar, Robert Birke, Hiek Van der Scheer, and Lydia Y Chen. Ctab-gan+: Enhancing tabular data synthesis. *Frontiers in big Data*, 6:1296508, 2024.

Table 8: Links to prior datasets popular in tabular synthesis evaluations.

Dataset	Link
Diabetes	https://www.openml.org/search?type=data&sort=runs&id=37&status=active
Travel	kaggle.com/datasets/tejashvi14/tour-travels-customer-churn-prediction/data
Adult	https://archive.ics.uci.edu/dataset/2/adult
Shoppers	https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset
House	scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html

Table 9: URLs to access each dataset in the StruggleTab benchmark.

Dataset	URL
Honeypot	secrepo.com/honeypot/honeypot.json.zip
Stroke	kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
MoMA	zenodo.org/doi/10.5281/zenodo.20506252
Bayesian	N/A
CERN	opendata.cern.ch/record/304
LOB	kaggle.com/datasets/martinsn/high-frequency-crypto-limit-order-book-data
Olist	kaggle.com/dsv/195341

A Prior Common Datasets

Diabetes (Kahn, 1994) contains medical information on female hospital patients, including age, number of pregnancies and skin thickness. Downstream models learn to predict whether a given patient suffers from diabetes. Apart from the label, all dataset columns are numerical, with some columns taking only integer values, while others are floats.

Travel (Tejashvi, 2023) was collected by a travel agency wishing to predict customer churn. With the binary variable churn as the target, features include whether the traveler booked a hotel, frequent flyer status and traveler age. While most features are categorical, there are two numerical columns: traveler age and the number of times that the customer has used the travel agency in the past.

Adult (Becker & Kohavi, 1996) is a dataset commonly used to benchmark tabular classification algorithms. Each row contains basic information on one American adult, such as their age, years of education and marital status. For each adult, the downstream task is to predict whether their annual income is above or below $\$50,000$. The features are a mix of categorical and numerical columns, with each numerical column taking only integer values.

The *Shoppers* (Sakar & Kastro, 2018) dataset contains session-level data from an e-commerce website over a one-year period. Each record describes a single visit using 18 behavioral and contextual features, including page visit counts and durations, bounce and exit rates, visitor type, month and weekend indicator. The target variable is binary, indicating whether the session resulted in a purchase, making the dataset suitable for evaluating classification methods on imbalanced user-behavior data.

House (Pace & Barry, 1997) is a standard regression dataset. Each row represents a block of houses in California during the 1990 census. The dataset records the number of households residing in the block, the block's median building age, average number of bedrooms and other basic information. The dataset's target column is the block's median house value, which is numerical and allows us to assess Tabby's synthetic data in a regression task.

B The StruggleTab Datasets

B.1 Access Information

Table 9 provides the URLs to access each dataset. A major factor in the choice of these datasets was data licensing and permissiveness, ensuring that all datasets can be accessed programatically, are free to use for research purposes and do not require users to apply for access. While this decision prevents the selection of some potentially interesting datasets, such as the clinical MIMIC-III dataset Johnson et al. (2016), it allows us to keep StruggleTab accessible to all members of the tabular research community. For instance, if the source of one dataset prevents hosting the data at some time in the future, the licensing permissiveness would allow the dataset to be hosted through a new medium rather than the dataset becoming lost.

We generated the Bayesian dataset ourselves to ensure that pretrained models did not experience any training data leakage. To generate Bayesian, we construct a 100000-sample synthetic Directed Acyclic Graph (DAG) designed to evaluate a model's ability to capture causal colliders, non-linear manifolds and strict logical boundaries amidst uninformative noise. The data generation process uses a fixed random seed to ensure exact reproducibility and defines the following structural relationships:

- **Independent Roots:** Two base features are sampled from a standard normal distribution, $X_1, X_2 \sim \mathcal{N}(0, 1)$.
- **Collider:** A third feature $X_3 = X_1 + X_2 + \mathcal{N}(0, 0.1)$ introduces a causal collider, rendering the independent roots X_1 and X_2 highly correlated when conditioning on X_3.
- **Non-Linear Proxy:** We sample a hidden confounder $U \sim \mathcal{U}(-3, 3)$. Because U is explicitly excluded from the final dataset, models must infer its signal through a proxy feature defined by a sine wave manifold: $X_5 = \sin(U) + \mathcal{N}(0, 0.05)$.
- **Categorical Distractor:** A uniformly random categorical feature $X_6 \in \{\text{Alpha, Beta, Gamma, Delta}\}$ is included to test model robustness against uninformative noise.
- **Target:** A binary classification target defined by a strict logical boundary between the collider and the hidden confounder: $Y = \mathbb{I}(\sigma(X_3 + U) > 0.5)$, where σ is the standard logistic sigmoid function.

The final published dataset consists exclusively of $\{X_1, X_2, X_3, X_5, X_6\}$ and the target Y. The Bayesian dataset will be released along with the StruggleTab codebase.

B.2 Challenging Aspects

For each of the three challenging dataset selection criteria, we now elaborate on each dataset's ability to meet that criteria.

Text Features (■): We consider a feature to be text (as opposed to categorical) if there are at least 200 unique categories, or if the number of unique feature values is greater than 10% of rows in the dataset. StruggleTab datasets with features meeting this criterion are:

- *Honeypot:* 3/5 features are strings (source_ip, request_url and request_raw). Specifically, "request_raw" is a JSON object, which we represent as a string to all synthesis models.
- *MoMA:* 10/25 features are strings (Title, ConstituentID, ArtistBio, BeginDate, EndDate, Date, Medium, Dimensions, CreditLine, AccessionNumber).

Feature Imbalance (●):

- *Honeypot:* The "pattern" feature contains 13 unique values across the dataset's rows. In this feature, 80% of rows have the value of "unknown", 11% "head", 6% "tomcat_manager" and $< 1\%$ for each value of "login","robots","style_css","phpmyadmin","lfi","phpinfo","rfi", "comments", "file_server", "sqli".
- *Stroke:* In the binary "stroke" feature, indicating whether the patient represented by a given row experienced a stroke, less than 5% of entries are positive.
- *MoMA:* In the "OnView" feature, indicating whether the artwork represented by a given row is on display at the museum, over 98.98% of values are empty (indicating the piece is in storage). The remaining 47 possible OnView values are specific rooms of the museum, each represented by $< 1\%$ of total rows.
- *Olist:* In the "order_status" feature, 98% of rows are "delivered", 1% are "shipped" and each status of "cancelled", "invoiced", "processing", "unavailable" and "approved" are represented by $< 1\%$ of all rows.

Strict Invariants (▲):

- *Honeypot:* The "request_raw" feature can be a JSON object or array. If this feature is a JSON, it must be syntactically valid and successfully parse without decoding errors.
- *Bayesian:* The causal structure relies on deterministic addition. The "collider" feature must equal the direct sum of "root_1" and "root_2", bounded by a noise tolerance of 0.35.
- *CERN:* The generated invariant mass ("M") must equal $\sqrt{2 \cdot p_{T1} \cdot p_{T2} \cdot (\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2))}$, evaluated within a tolerance margin of 0.05.
- *LOB:* Financial order book mechanics dictate strict boundary limits relative to the current mid-price. All generated notional (volume) features must be strictly non-negative. Furthermore, limit order price distances must respect their order side: ask prices are placed above the mid-price (requiring positive distance values), while bid prices are placed below the mid-price (requiring negative distance values).
- *Olist:* Temporal causality governs chronological events. A package's delivery timestamp ("order_delivered_customer_date") cannot mathematically precede its purchase timestamp ("order_purchase_timestamp").

B.3 Data Splits

Each dataset is split into train/validation/test sets following proportions of 75/10/15. For Honeypot, LOB and Olist, the split is chronological to prevent leakage. In the case of LOB, we leave 2-hour gaps between each split, to further prevent leakage. For MoMA, Stroke, CERN and Bayesian, the split is random.

B.4 Example Dataset Rows

Below, we provide three example rows of each dataset as examples of each dataset's structure and contents: Table 10 for Honeypot, Table 11 for Stroke, Table 16 for CERN, Table 17 for LOB, Table 12 for MoMA, Table 18 for Olist and Table 15 for Bayesian.

Table 10: Three example rows from the Honeypot dataset.

pattern	source_ip	source_port	request_url	request_raw
style_css	162.197.24.67	60871	/style.css	GET /style.css HTTP/1.1 Accept: text/css,*/*;q=0.1 Accept-Encoding: gzip,deflate,sdch Accept-Language: en-US,en;q=0.8 Connection: keep-alive Dnt: 1 Host: ec2-54-68-96-53.us-west-2.compute.amazonaws.com Referer: http://ec2-54-68-96-53.us-west-2.compute... User-Agent: Mozilla/5.0 (Macintosh; Intel Mac...
unknown	173.174.71.56	62918	/favicon.ico	GET /favicon.ico HTTP/1.1 Accept: */* Accept-Encoding: gzip,deflate,sdch Accept-Language: en-US,en;q=0.8 Connection: keep-alive Host: ec2-54-68-96-53.us-west-2.compute.amazonaws.com User-Agent: Mozilla/5.0 (Macintosh; Intel Mac...
tomcat_manager	123.249.19.23	3423	/manager/html	GET /manager/html HTTP/1.1 Accept: text/html, */* Authorization: Basic Og== Content-Type: text/html Host: 54.68.96.53 User-Agent: Mozilla/3.0 (compatible; Indy Library)

Table 11: Three example rows from the Stroke dataset.

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	31	0	0	No	Self-employed	Rural	64.85	23	Unknown	0
Male	40	0	0	Yes	Self-employed	Rural	65.29	28.3	never smoked	0
Female	8	0	0	No	children	Urban	74.42	22.5	Unknown	0

Table 12: Three example rows from the MoMA dataset, features 1-10 (Table 1 of 3).

Title	ConstituentID	ArtistBio	Nationality	BeginDate	EndDate	Gender	Date	Medium	Dimensions
”I WILL NOT...	6300	(Austrian, 1886)	(Austrian)	-1886	-1981	(female)	-1948	Lithograph	Comp. 29 3/4 x 24 1/4”...
Untitled from... & Not	3228	(American, born 1945)	(American)	-1945	0	(male)	1990	Screenprint	
Saucepot and Cover	9491		(American)	0	0	()	1938,1942	Glaze...	Overall: h. 3 3/8”...

Table 13: Three example rows from the MoMA dataset, features 11-17 (Table 2 of 3).

CreditLine	AccessionNumber	Department	DateAcquired	Cataloged	ObjectID	OnView
Gift of the artist	280.195	Drawings & Prints	10/11/50	N	65910	
John B. Turner Fund	232.1996	Drawings & Prints	5/21/96	N	64881	
Gift of Lewis & Conger	54.1943.a-b	Architecture & Design	2/18/43	Y	1865	

Table 14: Three example rows from the MoMA dataset, features 18-25 (Table 3 of 3).

Circumference (cm)	Depth (cm)	Diameter (cm)	Height (cm)	Length (cm)	Weight (kg)	Width (cm)	Duration (sec.)
			75.5			61.5	
		16.2	8.6				

Table 15: Three example rows from the Bayesian dataset.

root_1	root_2	collider	proxy_sin	distractor_cat	target_class
0.496714153	1.030594588	1.683492812	0.624255334	Delta	1
-0.138264301	-1.155354767	-1.303041899	-1.002071084	Gamma	0
0.647688538	0.57543719	1.090172169	-0.449602327	Delta	0

Table 16: Three example rows from the CERN dataset.

E1	px1	py1	pz1	pt1	eta1	phi1	Q1	E2	px2	py2	pz2	pt2	eta2	phi2	Q2	M
28.5869	23.3933	14.677	7.3859	27.6163	0.264357	0.560324	-1	77.623	-19.8573	-34.8481	-66.4577	40.1087	-1.27878	-2.08873	1	85.8586
93.9132	-11.5125	16.4228	-91.7466	20.056	-2.22539	2.1822	1	14.6013	-4.48348	0.453562	-13.8885	4.50636	-1.84406	3.04077	-1	8.7139
9.37408	-0.47143	-3.62015	8.63398	3.65072	1.5959	-1.70029	-1	98.234	-22.0727	14.2707	-94.6523	26.2842	-1.99313	2.56766	1	59.6546

Table 17: Three example rows from the LOB dataset (Middle Columns Omitted).

bids_distance_0	asks_distance_0	bids_notional_0	asks_notional_0	...	bids_distance_4	asks_distance_4	bids_notional_4	asks_notional_4	future_return
-8.95E-08	8.95E-08	714.91	773.45	...	-0.000162175	8.04E-05	5190.99	99068.04	0.0036
-1.28E-05	1.28E-05	2797.40	6982.20	...	-0.00025854	2.12E-05	497.98	2803.09	0.0027
-8.93E-08	8.93E-08	95037.17	3.04	...	-0.000173082	0.000485506	2998.19	8406.15	0.0018

Table 18: Three example rows from the Olist dataset.

review_score	price	freight_value	product_category_name	product_name_length	product_description_length	product_photos_qty	order_status	order_purchase_timestamp	order_delivered_customer_date
1	32.9	31.67	moveis_decoracao	41	754	3	shipped	9/4/16 21:15	
1	59.5	15.56	telefonia	42	381	1	canceled	9/5/16 00:15	
1	44.99	2.83	beleza_saude	34	1036	1	delivered	9/15/16 12:16	11/9/16 07:47

C STRUGGLETAB BENCHMARK IMPLEMENTATION

The StruggleTab benchmark is easy to integrate into preexisting tabular synthesis model workflows. The process of using StruggleTab is summarized in the following psuedocode:

Listing 1: Example integration of the StruggleTab evaluation harness.

```
import sys
sys.path.append("path_to_struggletab_codebase")
from handler import UnifiedDataLoader; from eval import evaluate

# 1. Load the benchmark dataset
train_df = UnifiedDataLoader(dataset_name).get_train_data()

# 2. Train and sample from your generative model
synthetic_data_frame = your_full_model_pipeline(train_df)

# 3. Evaluate the generated synthetic data
result = evaluate(dataset_name, synthetic_data_frame)
```

The result of the `evaluate` method is a dictionary containing metric name keys and corresponding result values.

D EXPERIMENTAL DETAILS

D.1 MODELS (SELECTION, COMPATIBILITY AND HYPERPARAMETERS)

Selection: We aim to represent the state of the art within two primary classes of modern tabular synthesis models, LLMs and diffusion models, as well as the legacy class of GANs.

LLMs: Tabby Cromp et al. (2025) and In-Context Learning (ICL) Brown et al. (2020) are our LLM models. While ICL was chosen as a few-shot baseline, Tabby represents a fine-tuned LLM. Specifically, Tabby replaces the language modeling head of a pretrained LLM with a mixture of experts layer, so that each column is assigned a dedicated expert within the language modeling head. While we also considered the inclusion of older legacy models such as GReaT Borisov et al. (2022), TapTap Zhang et al. (2023) and Tabula Zhao et al. (2023), these models were not designed for compatibility with features such as freeform text, and Tabby has already demonstrated higher performance than these older models on standard tabular evaluations.

Diffusion models: We evaluate TabDiff Shi et al. (2025a) and TabDLM Cai et al. (2026). TabDiff constructs a joint continuous diffusion space for both numerical and categorical features but lacks the architectural capacity to process strings. Conversely, TabDLM employs a hybrid transformer-diffusion architecture that textually serializes all features prior to diffusion, granting it full string compatibility. Similar to the LLMs, we considered a broader class of legacy models including TabSyn Zhang et al. (2024) (a similar, older model from the authors of TabDiff), TabDDPM Kotelnikov et al. (2023) and ForestDiffusion Jolicoeur-Martineau et al. (2024). However, these models demonstrate poorer performance than the newer models on standard tabular synthesis evaluations and were also not easily compatible with the StruggleTab features, including text and extreme feature imbalance.

Pre-Deep Learning Models: Before the uptake of deep learning-based tabular synthesis models, GAN and VAE architectures were standard. We include the CTAB-GAN+ Zhao et al. (2024) model to represent this group (Shortened to CTGANP in this work). This model continues to be widely included as a baseline within many papers' evaluations (e.g., Cromp et al. (2025), Kotelnikov et al. (2023), Zhao et al. (2024)) and has widely demonstrated higher performance over other older models, including CTAB-GAN and TVAE Xu et al. (2019).

Scope of Evaluated Models: The primary contribution of this work is the introduction of the StruggleTab benchmark itself. To validate the benchmark's utility and difficulty, we selected a representative suite of five state-of-the-art models spanning all major contemporary architectural paradigms (GANs, continuous diffusion, hybrid diffusion, fine-tuned MoE LLMs, graph-augmented LLMs and prompt-based ICL). Because our objective is to demonstrate the benchmark's capacity to highlight

Table 19: URLs of codebases for publicly-available models evaluated in Section 4

Model	URL
CTGANP	https://github.com/Team-TUD/CTAB-GAN-Plus
Tabby	https://github.com/soCromp/tabby/tree/main
TabDiff	https://github.com/MinkaiXu/TabDiff
TabDLM	https://github.com/ilikevegetable/TabDLM

Table 20: Hyperparameters selected per model/dataset pair, according to hyperparameter search detailed in Appendix D.

Model	ICL Temperature	Tabby LR	TabDiff LR	TabDLM Batch Size	TabDLM Batch Accum.
Honeypot	0.6	0.0001	0.0001	2	8
Stroke	0.2	0.0001	0.0001	32	8
MoMA	0.2	0.0001	0.001	2	4
Bayesian	0.2	0.00005	0.001	32	4
CERN	0.2	0.0001	0.001	32	8
LOB	1.0	0.0005	0.0001	16	16
Olist	0.2	0.00001	0.001	32	16

the limitations of current leading methods on heterogeneous data, an exhaustive survey of legacy or strictly numeric tabular models falls outside the scope of this evaluation. The systemic challenges observed across this diverse, cutting-edge sample illustrate the gap between current generative capabilities and real-world tabular complexity, confirming StruggleTab's value as a rigorous evaluation tool for tabular synthesis research.

Implementations: Table 19 lists the URLs for the implementations we use of publicly-available models.

We implement the In-Context Learning (ICL) baseline ourselves. The ICL baseline evaluates the zero-shot and few-shot tabular synthesis capabilities of pre-trained large language models (by default, Llama-3-8B). For each synthetic row, the pipeline constructs a prompt containing k randomly sampled real records (default k=5) serialized into a strict JSON format. The model is instructed to act as a tabular data synthesizer and output a single novel record that adheres to the demonstrated schema and statistical distribution. To encourage valid output formatting, the generation is artificially seeded with an opening JSON brace. The resulting raw text is parsed using regular expressions to extract the generated key-value pairs; hallucinated keys are discarded and unparseable outputs default to missing values to ensure the final synthetic table matches the dimensionality of the training set.

Compatibility: Dates are given to string-compatible models as strings and to models incompatible with strings as numerical Linux epochs. For models incompatible with strings, we augment their synthetic data with blank ("") string features in each row prior to evaluations.

Hyperparameters: For each compatible model-dataset pair, we perform a careful hyperparameter optimization in the following search spaces:

- *ICL:* Temperature in $\{0.2, 0.6, 1.0\}$ for Meta Llama 3 8B (Grattafiori et al., 2024) with 5 shot sampling.
- *Tabby:* Learning rate in $\{1e-5, 5e-5, 1e-4, 5e-4, 1e-3\}$ for base model Meta Llama 3 1B (Grattafiori et al., 2024). We use the codebase's PEFT default configuration and run for 2 epochs for CERN, 4 epochs for Olist, 7 epochs for Bayesian and 10 epochs for all other datasets.
- *TabDiff:* Learning rate in $\{1e-5, 1e-4, 1e-3\}$.
- *TabDLM:* Batch size in $\{2, 8, 16, 32\}$ with batch accumulation in $\{4, 8, 16\}$, with LoRA $r = 4$ and $\alpha = 128$, 15 training epochs and sampling temperature 1.0.

CTGANP is run with default hyperparameters, as the public implementation of this model does not expose hyperparameters.

For each hyperparameter setting, the model is trained on the train split, then sampled 200 times and graded according to a score function. The score function is calculated as the sum of the average Kolmogorov-Smirnov (KS) distance (Massey Jr, 1951) and a dataset-specific invariant penalty. The KS component ensures baseline 1D marginal distributional fidelity across all numerical features, explicitly assigning a maximum penalty of 1.0 if a model fails to generate a required column. The invariant penalty acts as a targeted regularizer to penalize severe domain-specific failures. Depending on the dataset, this penalty detects violations such as breaking temporal causality (Olist), hallucinating negative financial or biological bounds (LOB, Stroke), violating non-linear physics equations (CERN) or failing to learn cross-feature relationships (Honeypot).

By selecting hyperparameters that minimize constraint violations on the validation set, the models are given an advantage to adhere to the dataset rules. Therefore, the constraint violation rates ultimately observed on the test set (Section 4.5) cannot be dismissed as an artifact of suboptimal hyperparameter tuning; rather, they expose current limitations in modeling high-precision mathematical boundaries.

Chosen hyperparameters for each model-dataset pair are specified in Table 20.

D.2 Adapting Metrics for the StruggleTab Datasets

Unified Vector Space: To calculate the Section 4.2 Classifier Two-Sample Test (C2ST) and Section 4.4 Distance to Closest Record (DCR) metrics and to run the downstream classifier/regressor models for Section 4.3 MLE across heterogeneous datasets, all tabular rows are projected into a unified, high-dimensional continuous vector space. This ensures that metrics and downstream models can natively process unstructured text and discrete categories alongside continuous numerical values. The projection pipeline applies modality-specific transformations before horizontal concatenation:

- *Continuous Features:* Missing continuous values are imputed with 0.0 and all numerical columns are normalized using a Scikit-Learn standard scaler fit to the real training distribution. This ensures that features with large magnitudes do not disproportionately dominate distance calculations.
- *Categorical Features:* Missing values are cast to a distinct "Missing" category and all low-cardinality discrete variables are mapped into the continuous space via one-hot encoding (Scikit-Learn OneHotEncoder), ignoring unknown categories encountered during evaluation.
- *Unstructured Text and JSON:* High-cardinality strings and freeform payloads (e.g., "Title" in MoMA, "request_raw" in Honeypot) are standardized to empty strings if missing, then embedded using the all-MiniLM-L6-v2 sentence transformer. This maps variable-length sequences and JSON objects into 384-dimensional vectors.

Following these transformations, the outputs are horizontally concatenated into a single dense vector per row. For DCR, we compute the exact L_2 (Euclidean) distance between synthetic and real vectors in this unified space using a 1-Nearest Neighbor search. For C2ST, the unified vectors are fed directly as inputs into the random forest discriminator.

Downstream Utility MLE Tasks: For MLE, we define datasets' tasks and targets as specified in Table 21. We choose downstream models from the Scikit-Learn library that optimize the MLE of the "Baseline" (i.e., the non-synthetic trainset tested on the non-synthetic testset). For all Classification tasks, we use a RandomForestClassifier. For LOB, we use RidgeCV with robust scaling, for Olist, we use HistGradientBoostingRegressor with maximum depth of 10 and log-scale the long-tailed price/freight targets. For CERN, we opt for the RandomForestRegressor.

Performance Profiles: For a given method $m \in M$ evaluated over a set of compatible tasks T_m, the performance profile curve $\rho_m(\tau)$ represents the fraction of tasks where the method's performance ratio $\tau_{t,m}$ is within a factor $\tau \geq 1$ of the best observed model:

$$\rho_m(\tau) := \frac{1}{|T_m|} \left| \{ t \in T_m : \tau_{t,m} \leq \tau \} \right|.$$

The calculation of the performance ratio $\tau_{t,m}$ for a raw metric score $M_{t,m}$ is slightly different per metric:

For C2ST AUC (Section 4.2 and Figure 3), where 0.50 indicates real and synthetic data are perfectly indistinguishable, we first compute the absolute error: $E_{t,m} = |M_{t,m} - 0.50|$. The ratio is calculated as $\tau_{t,m} = \frac{E_{t,m}}{\min_{m'}(E_{t,m'})+\epsilon}$. We add a small constant ($\epsilon = 10^{-5}$) to the denominator to prevent undefined division if a model achieves a mathematically perfect C2ST score.

For MLE (Section 4.3 and Figure 4), where higher scores indicate better performance, the ratio is calculated as $\tau_{t,m} = \frac{\max_{m'}(M_{t,m'})}{M_{t,m}}$. To handle models that score exactly 0.0, their scores are replaced with a small constant ($\epsilon = 10^{-3}$) to prevent infinite ratios. When the absolute best score achieved by *any* model on a dataset is exactly 0.0 (e.g., the LOB dataset), standard ratio scaling breaks down. In these edge cases, all compatible models evaluated on that dataset are assigned a baseline ratio of $\tau_{t,m} = 1.0$, correctly reflecting an equal multi-way tie where no architecture outperformed a naive baseline.

Finally, the x-axis is plotted on a logarithmic scale ($\log_{10} \tau$) to appropriately visualize relative performance decay. For each method, we summarize overall robustness by calculating the Area Under the Profile (AUP) score:

$$\text{AUP}_m = \int_0^{\log_{10}(\tau^*)} \rho_m(10^x)\, dx$$

where $x = \log_{10}(\tau)$ and τ^* is the smallest τ such that $\rho_m(\tau) = 1$ for all methods $m \in M$. A higher AUP score indicates strictly better performance across the benchmark. Note that models fundamentally incompatible with a dataset's modalities (e.g., TabDiff on text-heavy datasets) are excluded from their respective $|T_m|$ denominator.

Table 21: Prediction targets and task types for downstream Machine Learning Efficacy (MLE) evaluations.

	Task	**Target**
Honeypot	Classification	pattern
Stroke	Classification	stroke
MoMA	Classification	Department
Bayesian	Classification	target_class
CERN	Regression	M
LOB	Regression	future_return
Olist	Regression	freight_value

Membership Inference Attacks (MIAs): Our privacy evaluations in Section 4.4 use a battery of likelihood-based Membership Inference Attacks (MIAs)[1]. Because implicit generative models, such as Generative Adversarial Networks (e.g., CTGANP), do not natively expose tractable exact likelihoods for generated samples, we evaluate all architectures under a realistic black box threat model using the Tab-MIA framework. Rather than relying on white-box model weights or native exact likelihoods, the attack assumes the adversary only has access to the released synthetic dataset. For each evaluated model, a surrogate language model is fine-tuned directly on the synthetic outputs to act as a continuous density estimator. The exact likelihood-based thresholding and entropy calculations are subsequently executed against this surrogate. This methodology standardizes the MIA process across diverse architectures (LLMs, Diffusion and GANs), ensuring that privacy leakage is measured based on the memorization embedded within the synthetic data artifacts themselves, rather than relying on architecture-specific likelihood tractability.

Radar Figure 1: To generate the performance summary in Figure 1, we normalize the raw evaluation metrics such that all axes follow a "higher is better" formulation bounded between 0.0 and 1.0. Evaluations were then filtered through gating mechanisms, including a memorization penalty triggered by excessive deviations from empirical baseline match rates, as well as a penalty for models lacking string support on text-heavy datasets. These filtered metrics were then transformed into performance ratios (τ) and we report the Area Under the Performance Profile (AUP) for each metric against its absolute theoretical optimum (the dotted black line).

[1] We use the Tab-MIA codebase, `https://github.com/eyalgerman/Tab-MIA`

E ADDITIONAL METRICS AND RESULTS

Table 22: Summary of Evaluation Metrics in the StruggleTab Benchmark

Metric	Definition	Results	Criterion Assessed	Target
C2ST AUC	§3.2, §D.2	§4.2	Fidelity	Closest to 0.50
MLE (Macro F1)	§3.2, §D.2	§4.3	Downstream Utility	Higher
MLE (R^2)	§3.2, §D.2	§4.3	Downstream Utility	Higher
Distance to Closest Record (DCR)	§3.2, §D.2	§4.4	Privacy & Memorization	Closest to Baseline
Membership Inference Attack (MIA)	§3.2, §D.2	§4.4	Privacy & Memorization	Closest to 50.0%
Constraint Violation Rate	§3.2, §B.2	§4.5	Domain Constraints	Lower
Joint Fréchet Distance (FID)	§E	§E	Fidelity	Lower
Exact Match Rate	§E	§E	Privacy & Memorization	Closest to Baseline
Synthetic Row Count	§E	§E	Domain Constraints	Higher

We now share supplemental metrics for fidelity, downstream utility and privacy. For an overview of all metrics used in the paper, refer to Table 22.

Fidelity: Fréchet Inception Distance (FID) results are in Table 24. While FID is standard in image synthesis, our experiments revealed it is ill-suited for assessing the fidelity of heterogeneous tabular data. When models hallucinate out-of-bounds continuous values (e.g., LLMs generating 10^{22} for financial prices), the covariance trace in the FID calculation explodes. We instead opt for the Classifier Two-Sample Test (C2ST) AUC to detect cross-modal failures without suffering from numerical overflow.

Utility: In main paper MLE results (Section 4.3, Table 4) we lower-bound the R^2 scores of regression tasks (CERN, LOB, Olist) at 0.00. In Table 23, we provide the full, unbounded R^2 scores of these regression task datasets.

Privacy: To supplement the geometric DCR evaluations, we evaluate exact whole-row matches. For every synthetic row, we check whether it occurs in the training set. The match rate is calculated as the number of synthetic rows appearing in the training data divided by the total number of synthetic rows. To execute this, we standardize harmless representation differences, such as column order and numeric serialization, but do not round numbers or change text features. Additionally, we measure the rate of natural repetition within the real dataset by finding the number of testset rows that co-occur in the trainset. This metric largely supports the findings indicated by DCR. For instance, match rate reveals vulnerabilities in prompt-based ICL, reproducing exact training records in 85% of its Olist generations despite the natural baseline repetition rate being 0.00.

Synthetic Sample Sizes: Table 27 reports the average number of synthetic rows successfully generated by each model. We note that the In-Context Learning (ICL) baseline produced relatively few valid samples for the text-heavy Honeypot (17 ± 1) and MoMA (10 ± 0) datasets. This reduction is a natural artifact of prompting LLMs with dense, multimodal records, which frequently exhaust context windows or result in truncated, unparseable JSON outputs. While these smaller sample cohorts limit large-scale distributional comparisons, they remain informative for exposing exact-row memorization (Table 25) and demonstrating the structural vulnerabilities inherent to zero/few-shot tabular synthesis.

Table 23: Downstream Utility via ML Efficacy (MLE). We report unbounded R^2 for regression tasks (Table 4 provides MLE with R^2 lower-bounded to 0.0, as well as MLE for classification tasks). **Higher is better.**

	CERN ▲	**LOB** ▲	**Olist** ●▲
CTGANP	0.67±0.02	**−0.04±0.00**	0.08±0.02
TabDiff	**0.75±0.02**	−0.05±0.01	0.07±0.01
TabDLM	0.10±0.02	−0.05±0.01	0.06±0.01
ICL	0.40±0.04	−0.05±0.00	**0.11±0.02**
Tabby	0.04±0.07	−0.05±0.00	0.02±0.02
Baseline	*0.75*	*9.23e-2*	*0.11*

Table 24: Joint Fréchet Distance (FID) across multimodal embeddings.

	Honeypot ■●▲	**Stroke** ●	**MoMA** ■●	**Bayesian** ▲	**CERN** ▲	**LOB** ▲	**Olist** ●▲
CTGANP	-	0.07±0.05	**7.04±1.91**	0.03±0.01	**0.22±0.07**	**5.90±1.97**	**1.41±0.39**
TabDiff	-	**0.02±0.01**	61.35±46.89	**0.02±0.01**	2.17±0.07	28.46±13.78	59.38±77.52
TabDLM	8.20±3.39	0.18±0.04	1869.07±296.22	0.72±0.60	3.00±1.10	7.43±3.96	193.99±104.64
ICL	2.72±1.40	0.46±0.01	44.66±12.19	12.44±2.00	6.23±0.77	$(1.33\pm2.25)\times10^{22}$	200.63±140.81
Tabby	**5.67±0.01**	4.25±1.25	2982.71±967.50	1.07±0.06	7.97±1.35	$(1.46\pm1.94)\times10^{11}$	1315.98±517.80

Table 25: Exact Match Rate (including string features). Measures the proportion of synthetic rows that perfectly replicate a record in the training set. The *Baseline* row evaluates the natural repetition rate in the held-out test set.

	Honeypot ■●▲	**Stroke** ●	**MoMA** ■●	**Bayesian** ▲	**CERN** ▲	**LOB** ▲	**Olist** ●▲
CTGANP	-	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
TabDiff	-	0.00±0.00	0.06±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
TabDLM	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
ICL	0.56±0.17	0.65±0.02	0.10±0.00	0.00±0.00	0.21±0.01	0.05±0.00	0.85±0.01
Tabby	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.63±0.07
Baseline (Test)	*0.09*	*0.00*	*0.05*	*0.00*	*0.00*	*0.00*	*0.00*

Table 26: Exact Match Rate (excluding string features). Measures the proportion of synthetic rows that perfectly replicate the categorical and continuous features of a training record.

	Honeypot ■●▲	**Stroke** ●	**MoMA** ■●	**Bayesian** ▲	**CERN** ▲	**LOB** ▲	**Olist** ●▲
CTGANP	-	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
TabDiff	-	0.00±0.00	0.06±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
TabDLM	0.99±0.01	0.00±0.00	0.45±0.02	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
ICL	0.64±0.17	0.65±0.02	0.10±0.00	0.00±0.00	0.21±0.01	0.05±0.00	0.85±0.01
Tabby	1.00±0.00	0.00±0.00	0.11±0.02	0.00±0.00	0.00±0.00	0.00±0.00	0.63±0.07
Baseline (Test)	*1.00*	*0.00*	*0.12*	*0.00*	*0.00*	*0.00*	*0.00*

Table 27: Average number of generated synthetic samples across trials.

	Honeypot ■●▲	**Stroke** ●	**MoMA** ■●	**Bayesian** ▲	**CERN** ▲	**LOB** ▲	**Olist** ●▲
CTGANP	-	2000±0	2000±0	2000±0	2000±0	2000±0	2000±0
TabDiff	-	2000±0	2000±0	2000±0	2000±0	2000±0	2000±0
TabDLM	2000±0	2000±0	2000±0	1093±785	2000±0	2000±0	2000±0
ICL	17±1	2000±0	10±0	2000±0	2000±0	2000±0	2000±0
Tabby	2000±0	2000±0	1822±291	2000±0	1999±0	1981±1	1974±7