Novelty Assessment Report

Paper: Using maximal information auxiliary variables to improve synthetic data generation based on TabPFN foundation models
PDF URL: https://openreview.net/pdf?id=6PkiUAcTWF
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Synthetic data generation for tabular datasets is shifting toward the use of large, general-purpose foundation models. TabPFN, a state-of-the-art example, uses in-context learning to generate probabilistic predictions conditioned on observed examples in a single forward pass. However, when variables are only weakly associated with others, the model's ability to generate realistic synthetic data deteriorates, as the context examples provide little predictive signal. To address this, we introduce the maximal information auxiliary variable (MIAV) strategy, which increases context information with auxiliary variables constructed by rank-matching random noise variables to real data. We establish theoretical properties of the approach which explain its good performance for weakly associated variables. Additional practical advantages of the MIAV approach include improved computational efficiency and invariance to variable order during the synthetic data generation process. Empirical evaluations, on simulated and real datasets, illustrate how the MIAV strategy improves data generation when compared to direct application of TabPFN, and is competitive against other baselines. To illustrate the generality of the MIAV approach we also present an implementation based on the TabICL model (a more scalable tabular foundation model restricted to classification tasks) for performing synthetic data generation on categorical datasets. Overall, MIAV offers an effective foundation model–based alternative to bespoke synthetic data generators.

Core Task Landscape

This paper addresses: **synthetic data generation for tabular datasets using foundation models**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Tabular Foundation Model Architectures and Pretraining**
- **Synthetic Data Generation Methodologies**
- **Privacy-Preserving Synthetic Data Generation**
- **Domain-Specific Applications and Specialized Contexts**
- **Evaluation, Benchmarking, and Methodological Surveys**

Complete Taxonomy Tree

- synthetic data generation for tabular datasets using foundation models Survey Taxonomy
- Tabular Foundation Model Architectures and Pretraining
  - In-Context Learning Foundation Models (3 papers)
  - [1] Accurate predictions on small data with a tabular foundation model (Noah Hollmann, 2025) View paper
  - [4] TabICL: A Tabular Foundation Model for In-Context Learning on Large Data (Qu, 2025) View paper
  - [23] In-context learning of evolving data streams with tabular foundational models (Gama, 2025) View paper
  - Pretraining Data and Synthetic Priors (3 papers)
  - [8] Real-tabpfn: Improving tabular foundation models via continued pre-training with real-world data (Garg Anurag, 2025) View paper
  - [12] Generating Synthetic Relational Tabular Data via Structural Causal Models (Hoppe, 2025) View paper
  - [14] Mitra: Mixed Synthetic Priors for Enhancing Tabular Foundation Models (Zhang Xiyuan, 2025) View paper
  - Cross-Table and Multi-Dataset Foundation Models (4 papers)
  - [6] Foundation Models for Tabular Data within Systemic Contexts Need Grounding (Klein, 2025) View paper
  - [21] Ctsyn: A foundational model for cross tabular data generation (Lin Xiao-feng, 2024) View paper
  - [42] Tabdpt: Scaling tabular foundation models on real data (Ma, 2025) View paper
  - Conceptual Frameworks and Research Agendas (2 papers)
  - [16] Why tabular foundation models should be a research priority (van Breugel, 2024) View paper
  - [26] Advancing web science through foundation model for tabular data (Inwon Kang, 2024) View paper
- Synthetic Data Generation Methodologies
  - Large Language Model-Based Generation
  - Prompt Engineering and In-Context Example Selection ★ (6 papers)
    - [0] Using maximal information auxiliary variables to improve synthetic data generation based on TabPFN foundation models (Anon et al., 2026) View paper
    - [10] TABGEN-RAG: iterative retrieval for tabular data generation with large language models (L Fang, 2024) View paper
    - [13] HARMONIC: Harnessing LLMs for tabular data synthesis and privacy protection (Sophia Ananiadou, 2024) View paper
    - [41] EPIC: Effective Prompting for Imbalanced-Class Data Synthesis in Tabular Data Classification via Large Language Models (Jaegul Choo, 2024) View paper

        - [44] Not All Features Deserve Attention: Graph-Guided Dependency Learning for Tabular Data Generation with Language Models (Zheyu Zhang, 2025) View paper
        - [47] TabGen-ICL: Residual-Aware In-Context Example Selection for Tabular Data Generation (Fang Lian-cheng, 2025) View paper
    - LLM Fine-Tuning and Adaptation (4 papers)
        - [15] Tabby: Tabular Data Synthesis with Language Models (Cromp, 2025) View paper
        - [18] Language Models are Realistic Tabular Data Generators (Borisov, 2022) View paper
        - [25] Tabula: Harnessing language models for tabular data synthesis (Zilong Zhao, 2025) View paper
        - [28] Generating Realistic Tabular Data with Large Language Models (Dang Nguyen, 2024) View paper
    - Zero-Shot and Text-to-Tabular Generation (2 papers)
        - [7] A text-to-tabular approach to generate synthetic patient data using LLMs (Margaux TÃ¶rnqvist, 2025) View paper
        - [45] Generative adversarial networks vs large language models: a comparative study on synthetic tabular data generation (Austin A. Barr, 2025) View paper
    - Diffusion Model-Based Generation (3 papers)
    - [3] Diffusion Models for Tabular Data Imputation and Synthetic Data Generation (Mario VillaizÃ¡n Vallelado, 2025) View paper
    - [33] TabDDPM: Modelling Tabular Data with Diffusion Models (Kotelnikov, 2022) View paper
    - [34] FinDiff: Diffusion Models for Financial Tabular Data Generation (Sattarov, 2023) View paper
    - GAN and VAE-Based Generation (3 papers)
    - [5] TTVAE: Transformer-based generative modeling for tabular data generation (Alex X. Wang, 2025) View paper
    - [22] Modeling Tabular data using Conditional GAN (Xu Lei, 2022) View paper
    - [35] Creating artificial students that never existed: Leveraging large language models and CTGANs for synthetic data generation (Mohammad Khalil, 2025) View paper
    - Energy-Based and Hybrid Approaches (2 papers)
    - [31] TabPFGen - Tabular Data Generation with TabPFN (Ma, 2024) View paper
    - [38] Differentially Private Normalizing Flows for Synthetic Tabular Data Generation (Jeong, 2022) View paper
- Privacy-Preserving Synthetic Data Generation
    - Differentially Private Generation with Foundation Models (3 papers)
    - [2] Is API Access to LLMs Useful for Generating Private Synthetic Tabular Data? (Swanberg, 2025) View paper
    - [46] DP-Tabula: Differentially Private Synthetic Tabular Data Generation with Large Language Models (Weijie Niu, 2025) View paper
    - [48] Differentially Private Tabular Data Synthesis using Large Language Models (Xiong Li, 2024) View paper
    - Privacy Risk Assessment and Benchmarking (2 papers)
    - [17] Risk In Context: Benchmarking Privacy Leakage of Foundation Models in Synthetic Tabular Data Generation (Lin Xiao-feng, 2025) View paper
    - [20] Mitigating and Assessing Bias and Fairness in Large Language Model-Generated Synthetic Tabular Data (Subah, 2024) View paper
- Domain-Specific Applications and Specialized Contexts
    - Healthcare and Medical Data Synthesis (1 papers)
    - [30] SynLLM: A Comparative Analysis of Large Language Models for Medical Tabular Synthetic Data Generation via Prompt Engineering (Shirazi, 2025) View paper
    - Low-Data Regime and Data Augmentation (2 papers)
    - [43] Curated LLM: Synergy of LLMs and data curation for tabular augmentation in low-data regimes (Seedat, 2023) View paper
    - [50] Causal Data Augmentation for Robust Fine-Tuning of Tabular Foundation Models (M BÃ¼hler, 2025) View paper
    - Specialized Application Domains (2 papers)
    - [37] Generating Synthetic Tabular Data for DDoS Detection Using Generative Models (Samed Saka, 2023) View paper
    - [49] CoddLLM: Empowering Large Language Models for Data Analytics (Zhang Jia-ni, 2025) View paper
    - Multi-Modal and Cross-Domain Tasks (1 papers)
    - [39] OmniTab: Pretraining with natural and synthetic data for few-shot table-based question answering (Z Jiang, 2022) View paper
- Evaluation, Benchmarking, and Methodological Surveys
    - Comprehensive Surveys and Literature Reviews (4 papers)
    - [11] Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding--A Survey (Fang Xi, 2024) View paper
    - [24] A comprehensive survey of synthetic tabular data generation (Wang Yili, 2025) View paper
    - [29] Generative AI for Tabular Data Synthesis (Alex X. Wang, 2025) View paper
    - [40] Deep Learning within Tabular Data: Foundations, Challenges, Advances and Future Directions (Ren, 2025) View paper
    - Benchmark Frameworks and Comparative Studies (4 papers)
    - [9] A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data (Daniele Panfilo, 2023) View paper
    - [19] Synthetic Data (Korbmacher, 2025) View paper
    - [27] Synthcity: a benchmark framework for diverse use cases of tabular synthetic data (Z Qian, 2023) View paper
    - [32] Data-centric AI: tabular data synthesis with deep generative models (Wang, 2024) View paper

Narrative

Core task: synthetic data generation for tabular datasets using foundation models. The field has evolved into several distinct branches that reflect different modeling philosophies and application priorities. Tabular Foundation Model Architectures and Pretraining focuses on building specialized pretrained models for tabular data, exemplified by works like TabPFN[1] and Real TabPFN[8], which adapt transformer-based approaches to handle heterogeneous table structures. Synthetic Data Generation Methodologies encompasses a broad spectrum of techniques, from classical GANs and diffusion models (e.g., TabDDPM[33], Diffusion Tabular Imputation[3]) to newer LLM-based approaches that leverage prompt engineering and in-context learning (e.g., TabICL[4], TABGEN RAG[10]). Privacy-Preserving Synthetic Data Generation addresses differential privacy and secure data sharing, with methods like LLM API Private Synthetic[2] and Differentially Private Flows[38]. Domain-Specific Applications target specialized contexts such as finance, healthcare, and cybersecurity, while Evaluation, Benchmarking, and Methodological Surveys provide critical assessments of generation quality and utility across diverse settings.

Recent activity has concentrated on harnessing large language models for tabular synthesis, where a key tension emerges between prompt-based methods that rely on careful example selection versus end-to-end learned representations. Works like HARMONIC[13] and EPIC[41] explore sophisticated prompt engineering and retrieval-augmented strategies to improve LLM-generated table quality, while Graph Guided Dependency[44] and TabGen ICL[47] investigate how to encode column dependencies and relational structure within the

in-context learning paradigm. Maximal Information Auxiliary[0] sits within this LLM-based generation cluster, emphasizing prompt engineering and in-context example selection to maximize information content in synthetic outputs. Compared to HARMONIC[13], which focuses on harmonizing retrieval with generation, and EPIC[41], which prioritizes example diversity, Maximal Information Auxiliary[0] appears to prioritize the informativeness of selected examples, addressing the challenge of balancing representativeness with privacy and utility in foundation model-driven tabular synthesis.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. TABGEN-RAG: iterative retrieval for tabular data generation with large language models

Authors: L Fang, A Liu, H Zhang, HP Zou, W Zhang | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ Large Language models (LLMs) have achieved encouraging results on tabular data generationâ¦ in-context learning ability of LLMs for tabular data generation. TABGEN-RAG operates â¦

Relationship Analysis

Both papers belong to the Prompt Engineering and In-Context Example Selection category, focusing on improving LLM-based tabular data generation through strategic selection of in-context examples. The original paper addresses weak variable associations by constructing maximal information auxiliary variables (MIAVs) through rank-matching, while the candidate paper (TABGEN-RAG) tackles the same challenge through iterative retrieval of residual samples that represent the gap between generated and real data distributions. The key difference is that the original paper uses pre-constructed auxiliary variables with theoretical guarantees about conditional independence, whereas TABGEN-RAG dynamically selects in-context examples at each iteration based on distributional distance metrics.

---

2. HARMONIC: Harnessing LLMs for tabular data synthesis and privacy protection

Authors: Sophia Ananiadou, Zhengyu Chen, Duanyu Feng, Jimin Huang, Hao Wang, et al. (7 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ The experiments find that our tabular data generation achieves â¦ metric DLT for LLM synthetic data generation, which quantifies â¦ We also assess a synthetic data generation method with â¦

Relationship Analysis

Both papers belong to the Prompt Engineering and In-Context Example Selection category, focusing on leveraging LLMs for tabular synthetic data generation through strategic prompt design and example selection. The original paper uses TabPFN with maximal information auxiliary variables (MIAVs) constructed via rank-matching to improve in-context learning for weakly associated variables, while the candidate paper (HARMONIC) employs instruction fine-tuning with k-nearest neighbor example selection on larger LLMs like LLaMA-2 to enhance privacy protection. The key difference is that the original paper addresses weak associations through auxiliary variable construction for direct TabPFN inference, whereas HARMONIC focuses on fine-tuning strategies with neighbor-based prompting to balance effectiveness and privacy in LLM-based generation.

---

3. EPIC: Effective Prompting for Imbalanced-Class Data Synthesis in Tabular Data Classification via Large Language Models

Authors: Jaegul Choo, Jinhee Kim, Tae-Sung Kim, Taesung Kim, J. Choo | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Large language models (LLMs) have demonstrated remarkable in-context learning capabilities across diverse applications. In this work, we explore the effectiveness of LLMs for generating realistic synthetic tabular data, identifying key prompt design elements to optimize performance. We introduce EPIC, a novel approach that leverages balanced, grouped data samples and consistent formatting with unique variable mapping to guide LLMs in generating accurate synthetic data across all classes, even fo...

Relationship Analysis

Both papers belong to the 'Prompt Engineering and In-Context Example Selection' category, focusing on improving LLM-based tabular data generation through strategic prompt design and example selection. While the original paper introduces maximal information auxiliary variables (MIAV) to enhance TabPFN's synthetic data generation by addressing weak variable associations through rank-matched noise variables, EPIC focuses on addressing class imbalance in tabular datasets through balanced, grouped data samples and consistent formatting strategies for general LLMs. The key difference is that the original paper targets foundation models like TabPFN with a novel auxiliary variable construction approach, whereas EPIC emphasizes prompt engineering techniques for handling imbalanced classification tasks across diverse LLMs.

---

4. Not All Features Deserve Attention: Graph-Guided Dependency Learning for Tabular Data Generation with Language Models

Authors: Zheyu Zhang, Shuo Yang, Bardh Prenkaj, Gjergji Kasneci | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Language Models (LLMs) have shown strong potential for tabular data generation by modeling textualized feature-value pairs. However, tabular data inherently exhibits sparse feature-level dependencies, where many feature interactions are structurally insignificant. This creates a fundamental mismatch as LLMs'self-attention mechanism inevitably distributes focus across all pairs, diluting attention on critical relationships, particularly in datasets with complex dependencies or semantically ...

Relationship Analysis

Both papers belong to the Prompt Engineering and In-Context Example Selection category, focusing on improving LLM-based tabular data generation through enhanced context design. The original paper introduces maximal information auxiliary variables (MIAVs) constructed via rank-matching to improve TabPFN's in-context learning for weakly associated variables, while the candidate paper (GraDe) addresses the structural mismatch between LLMs' dense attention and tabular data's sparse dependencies by incorporating graph-guided attention mechanisms with functional dependency constraints. The key difference is that the original paper focuses on auxiliary variable construction for better context examples, whereas GraDe modifies the attention mechanism itself to align with tabular data structure.

---

5. TabGen-ICL: Residual-Aware In-Context Example Selection for Tabular Data Generation

Authors: Fang Lian-cheng, Liu Aiwei, Liancheng Fang, Zhang Hengrui, Aiwei Liu, et al. (12 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Language models (LLMs) have achieved encouraging results in tabular data generation. However, existing approaches require fine-tuning, which is computationally expensive. This paper explores an alternative: prompting a fixed LLM with in-context examples. We observe that using randomly selected in-context examples hampers the LLM's performance, resulting in sub-optimal generation quality. To address this, we propose a novel in-context learning framework: TabGen-ICL, to enhance the in-contex...

Relationship Analysis

Both papers belong to the 'Prompt Engineering and In-Context Example Selection' category, focusing on improving LLM-based tabular data generation through strategic selection of in-context examples. While the original paper (MIAV) addresses weak variable associations by constructing maximal information auxiliary variables through rank-matching to enhance TabPFN's generation quality, the candidate paper (TabGen-ICL) tackles the problem of LLM prior distribution dominance by iteratively selecting residual-aware in-context examples that represent underrepresented regions of the data distribution. The key distinction is that MIAV uses auxiliary variables as a preprocessing step for TabPFN, whereas TabGen-ICL employs dynamic, iterative retrieval of real samples based on distribution residuals to guide general-purpose LLMs like GPT-4o.

Contributions Analysis

Overall novelty summary. The paper introduces the Maximal Information Auxiliary Variable (MIAV) strategy to improve synthetic tabular data generation when variables exhibit weak associations. It positions itself within the 'Prompt Engineering and In-Context Example Selection' leaf of the taxonomy, which contains six papers total. This leaf sits under 'Large Language Model-Based Generation', a moderately populated branch addressing prompt-based and retrieval-augmented approaches. The focus on enhancing in-context learning for TabPFN-like models places the work in an active but not overcrowded research direction, where recent efforts explore example selection, retrieval strategies, and dependency encoding.

The taxonomy reveals neighboring leaves addressing LLM fine-tuning, zero-shot generation, and diffusion-based methods, indicating that the field explores multiple paradigms beyond prompt engineering. Within the same parent branch, sibling papers like HARMONIC and EPIC tackle retrieval-augmented generation and example diversity, while the paper's MIAV strategy focuses on maximizing information content through auxiliary variables constructed via rank-matching. This positions the work as complementary to retrieval-focused methods, addressing a distinct challenge—weak variable associations—rather than competing directly on example selection or diversity metrics.

Among seventeen candidates examined, no contribution was clearly refuted. The MIAV strategy itself was compared against two candidates with no refutable overlap. Theoretical properties of the approach were assessed against five candidates, again with no refutations. The TabICL-based implementation was evaluated against ten candidates, yielding no clear prior work providing the same auxiliary variable construction mechanism. These statistics suggest that within the limited search scope, the specific combination of rank-matching auxiliary variables and theoretical justification for weak associations appears distinct from existing prompt engineering and in-context learning techniques.

Based on the top-seventeen semantic matches examined, the work appears to occupy a recognizable niche within LLM-based tabular generation. The analysis does not cover the full landscape of auxiliary variable methods or information-theoretic approaches outside the examined candidates. The absence of refutations within this scope suggests novelty in the specific MIAV construction, though broader exhaustive searches might reveal related techniques in adjacent fields such as data augmentation or feature engineering for tabular models.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Maximal Information Auxiliary Variable (MIAV) strategy for synthetic data generation

Description: The authors propose a novel strategy that constructs auxiliary variables by rank-matching random noise to real data variables. This approach addresses the limitation of TabPFN-based synthetic data generation when variables are weakly associated, by providing informative context for in-context learning.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Novel uncertainty quantification through perturbation-assisted sample synthesis

URL: View paper

Brief Assessment

Perturbation Sample Synthesis[59] focuses on uncertainty quantification through rank-preserving data perturbation for unstructured data with deep learning models, while the original paper addresses synthetic tabular data generation using TabPFN foundation models with rank-matched auxiliary variables for in-context learning. The technical approaches and application domains differ substantially.

---

2. TabSDS: a Lightweight, Fully Non-Parametric, and Model Free Approach for Generating Synthetic Tabular Data

URL: View paper

Brief Assessment

TabSDS[61] uses rank-matching transformations for synthetic data generation but does not employ auxiliary variables constructed specifically for in-context learning with foundation models like TabPFN. The approaches address different technical problems in synthetic data generation.

Contribution 2: Theoretical properties of MIAV approach

Description: The authors prove that the MIAV approach has two key theoretical properties: (i) conditional on its MIAV, a variable is independent of all other variables, and (ii) the MIAV retains maximal information about the variable in a non-parametric, information-theoretic sense (Theorem 1).

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. High-dimensional Kalman filtering: a review

URL: View paper

Brief Assessment

High Dimensional Kalman[57] focuses on Kalman filtering with auxiliary variables for state estimation in dynamical systems, not synthetic data generation or information-theoretic properties of auxiliary variables in tabular data contexts.

---

2. Wise-ale: Wide sample estimator for aggregate latent embedding

URL: View paper

Brief Assessment

Wise ALE[56] focuses on aggregate latent embeddings for generative modeling, not on auxiliary variable methods for synthetic data generation. The theoretical framework addresses variational inference for auto-encoders rather than conditional independence properties in tabular data synthesis contexts.

---

3. An Introduction to PottsUtils

URL: View paper

Brief Assessment

PottsUtils Introduction[58] focuses on Potts models for spatial statistics and graph-based simulations, not on auxiliary variable methods for tabular data generation or information-theoretic properties of MIAV approaches.

---

4. Towards causal representation learning with observable sources as auxiliaries

URL: View paper

Brief Assessment

Causal Representation Auxiliaries[54] focuses on causal representation learning with observable sources as auxiliary variables in a causal graph framework, not on synthetic data generation or information-theoretic properties of auxiliary variables for tabular foundation models.

---

5. A mixed approach for data fusion of HBS and SILC

URL: View paper

Brief Assessment

Mixed Approach Fusion[55] focuses on data fusion techniques for combining household budget surveys (HBS) and survey on income and living conditions (SILC), not on synthetic data generation or tabular foundation models. The technical contexts are fundamentally different.

Contribution 3: TabICL-based implementation demonstrating generality of MIAV

Description: The authors demonstrate that the MIAV strategy is not limited to TabPFN by implementing it with TabICL, a more scalable tabular foundation model. This implementation shows the approach can be directly applied to other PFN-based foundation models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Language Models are Realistic Tabular Data Generators

URL: View paper

Brief Assessment

Language Models Realistic[18] focuses on using language models (specifically GPT-2) for generating synthetic tabular data through textual encoding, not on foundation models for in-context learning like TabICL or TabPFN. The paper does not address auxiliary variables or the MIAV strategy.

---

2. Ctab-gan+: Enhancing tabular data synthesis

URL: View paper

Brief Assessment

Ctab GAN Plus[51] focuses on conditional GANs for tabular data synthesis using different encoding strategies and differential privacy. It does not address foundation models or in-context learning approaches like TabICL or TabPFN, which are central to the original paper's contribution.

---

3. FinDiff: Diffusion Models for Financial Tabular Data Generation

URL: View paper

Brief Assessment

FinDiff[34] focuses on diffusion models for financial tabular data generation, not on foundation models or in-context learning approaches like TabICL. The technical approaches are fundamentally different.

---

4. Mixed-type tabular data synthesis with score-based diffusion in latent space

URL: View paper

Brief Assessment

Mixed Type Diffusion[53] focuses on score-based diffusion models for mixed-type tabular data synthesis in latent space, not on foundation models using in-context learning or the MIAV strategy for synthetic data generation.

---

5. Modeling Tabular data using Conditional GAN

URL: View paper

Brief Assessment

Conditional GAN Tabular[22] focuses on generating synthetic tabular data using conditional GANs for mixed data types, not on implementing foundation model strategies like MIAV with TabICL. The paper addresses different technical challenges (mode-specific normalization, conditional generation) rather than demonstrating generality of MIAV across foundation models.

---

6. Diffusion Models for Tabular Data Imputation and Synthetic Data Generation

URL: View paper

Brief Assessment

Diffusion Tabular Imputation[3] focuses on diffusion models for tabular data imputation and generation, not on foundation models using in-context learning or the MIAV strategy for synthetic data generation.

---

7. Ctsyn: A foundational model for cross tabular data generation

URL: View paper

Brief Assessment

Ctsyn[21] focuses on cross-table synthetic data generation using diffusion models with autoencoder architectures, not on implementing MIAV strategies with TabICL or other PFN-based foundation models for in-context learning.

---

8. Tabula: Harnessing language models for tabular data synthesis

URL: View paper

Brief Assessment

Tabula[25] focuses on using LLMs for tabular data synthesis through tokenization and compression strategies, not on foundation models for in-context learning with auxiliary variables like MIAV. The technical approaches are fundamentally different.

---

9. Risk In Context: Benchmarking Privacy Leakage of Foundation Models in Synthetic Tabular Data Generation

URL: View paper

Brief Assessment

Risk In Context[17] focuses on privacy leakage evaluation of foundation models for synthetic tabular data generation, not on implementing MIAV strategies with TabICL or demonstrating the generality of MIAV approaches across different foundation models.

---

10. Tab-VAE: A Novel VAE for Generating Synthetic Tabular Data.

URL: View paper

Brief Assessment

Tab-VAE[52] focuses on VAE-based synthetic tabular data generation using Gumbel-softmax sampling for categorical variables, not on foundation models or in-context learning approaches like TabICL or TabPFN.

Appendix: Text Similarity Detection

Textual similarity detection checked 23 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Using maximal information auxiliary variables to improve synthetic data generation based on TabPFN foundation models: preliminary results

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Using maximal information auxiliary variables to improve synthetic data generation based on TabPFN foundation models View paper
- [1] Accurate predictions on small data with a tabular foundation model View paper
- [2] Is API Access to LLMs Useful for Generating Private Synthetic Tabular Data? View paper
- [3] Diffusion Models for Tabular Data Imputation and Synthetic Data Generation View paper
- [4] TabICL: A Tabular Foundation Model for In-Context Learning on Large Data View paper
- [5] TTVAE: Transformer-based generative modeling for tabular data generation View paper
- [6] Foundation Models for Tabular Data within Systemic Contexts Need Grounding View paper
- [7] A text-to-tabular approach to generate synthetic patient data using LLMs View paper
- [8] Real-tabpfn: Improving tabular foundation models via continued pre-training with real-world data View paper
- [9] A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data View paper
- [10] TABGEN-RAG: iterative retrieval for tabular data generation with large language models View paper
- [11] Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding--A Survey View paper
- [12] Generating Synthetic Relational Tabular Data via Structural Causal Models View paper
- [13] HARMONIC: Harnessing LLMs for tabular data synthesis and privacy protection View paper
- [14] Mitra: Mixed Synthetic Priors for Enhancing Tabular Foundation Models View paper
- [15] Tabby: Tabular Data Synthesis with Language Models View paper
- [16] Why tabular foundation models should be a research priority View paper
- [17] Risk In Context: Benchmarking Privacy Leakage of Foundation Models in Synthetic Tabular Data Generation View paper
- [18] Language Models are Realistic Tabular Data Generators View paper
- [19] Synthetic Data View paper
- [20] Mitigating and Assessing Bias and Fairness in Large Language Model-Generated Synthetic Tabular Data View paper
- [21] Ctsyn: A foundational model for cross tabular data generation View paper
- [22] Modeling Tabular data using Conditional GAN View paper
- [23] In-context learning of evolving data streams with tabular foundational models View paper
- [24] A comprehensive survey of synthetic tabular data generation View paper
- [25] Tabula: Harnessing language models for tabular data synthesis View paper
- [26] Advancing web science through foundation model for tabular data View paper
- [27] Synthcity: a benchmark framework for diverse use cases of tabular synthetic data View paper
- [28] Generating Realistic Tabular Data with Large Language Models View paper
- [29] Generative AI for Tabular Data Synthesis View paper
- [30] SynLLM: A Comparative Analysis of Large Language Models for Medical Tabular Synthetic Data Generation via Prompt Engineering View paper
- [31] TabPFGen - Tabular Data Generation with TabPFN View paper
- [32] Data-centric AI: tabular data synthesis with deep generative models View paper
- [33] TabDDPM: Modelling Tabular Data with Diffusion Models View paper
- [34] FinDiff: Diffusion Models for Financial Tabular Data Generation View paper
- [35] Creating artificial students that never existed: Leveraging large language models and CTGANs for synthetic data generation View paper
- [36] CTSyn: A Foundation Model for Cross Tabular Data Generation View paper
- [37] Generating Synthetic Tabular Data for DDoS Detection Using Generative Models View paper
- [38] Differentially Private Normalizing Flows for Synthetic Tabular Data Generation View paper
- [39] OmniTab: Pretraining with natural and synthetic data for few-shot table-based question answering View paper
- [40] Deep Learning within Tabular Data: Foundations, Challenges, Advances and Future Directions View paper

- [41] EPIC: Effective Prompting for Imbalanced-Class Data Synthesis in Tabular Data Classification via Large Language Models View paper
- [42] Tabdpt: Scaling tabular foundation models on real data View paper
- [43] Curated LLM: Synergy of LLMs and data curation for tabular augmentation in low-data regimes View paper
- [44] Not All Features Deserve Attention: Graph-Guided Dependency Learning for Tabular Data Generation with Language Models View paper
- [45] Generative adversarial networks vs large language models: a comparative study on synthetic tabular data generation View paper
- [46] DP-Tabula: Differentially Private Synthetic Tabular Data Generation with Large Language Models View paper
- [47] TabGen-ICL: Residual-Aware In-Context Example Selection for Tabular Data Generation View paper
- [48] Differentially Private Tabular Data Synthesis using Large Language Models View paper
- [49] CoddLLM: Empowering Large Language Models for Data Analytics View paper
- [50] Causal Data Augmentation for Robust Fine-Tuning of Tabular Foundation Models View paper
- [51] Ctab-gan+: Enhancing tabular data synthesis View paper
- [52] Tab-VAE: A Novel VAE for Generating Synthetic Tabular Data. View paper
- [53] Mixed-type tabular data synthesis with score-based diffusion in latent space View paper
- [54] Towards causal representation learning with observable sources as auxiliaries View paper
- [55] A mixed approach for data fusion of HBS and SILC View paper
- [56] Wise-ale: Wide sample estimator for aggregate latent embedding View paper
- [57] High-dimensional Kalman filtering: a review View paper
- [58] An Introduction to PottsUtils View paper
- [59] Novel uncertainty quantification through perturbation-assisted sample synthesis View paper
- [60] Using maximal information auxiliary variables to improve synthetic data generation based on TabPFN foundation models: preliminary results View paper
- [61] TabSDS: a Lightweight, Fully Non-Parametric, and Model Free Approach for Generating Synthetic Tabular Data View paper