# Pre-training a Large Tabular Model

**A. Scientific/Technical Goal**

**Extending Foundation Models to Tabular Data: Challenges and Limitations.** Foundation Models, particularly exemplified by Large Language Models (LLMs), have gained significant attention and demonstrated remarkable capabilities across a wide range of tasks and domains. However, while these models hold great potential for extending their power to tabular data, they face notable limitations. Existing LLMs are primarily trained on *unstructured* and web-scraped text data, which restricts their ability to effectively process *structured* tables of mixed-type data, and introduces the risk of perpetuating unfiltered biases from their text-based pre-training data sets. Previous research has explored leveraging language models via text interfaces to model tabular data, showing promising performance in both classification and regression tasks [**?**, **?**, **?**], as well as in generative tasks [**?**, **?**, **?**]. Despite these advances, the studies have primarily utilized models significantly smaller than the current state-of-the-art open-source models. These approaches either fine-tune models pre-trained on text data—risking the perpetuation of biases—or forego large-scale pre-training, leading to substantially downgraded performances and limited model applicability.

**Pre-training a Foundational Model for Tabular and Text Modalities.** To overcome limitations above, this proposal aims to pre-train a foundational model for tabular and text modalities, creating a versatile open-source resource for various domains heavily utilizing tabular data, including finance, healthcare, and marketing. This model leverages extensive pre-training, a robust neural network structure, and contextual information from related text and tables to perform all table-related machine learning tasks, enhancing performance across diverse data domains. It aims to advance table classification and regression tasks, such as user-behavior prediction, and assists in generating realistic synthetic tables and imputing missing data, especially in data-limited fields like healthcare. As our foundation model would be uniquely designed for tabular data and text in documents and reports, it could improve efficiency in applications such as IRS accounting, military logistics, internal government auditing/transparency and budget forecasting etc.

Literally, pre-trained large tabular models can serve as important roles akin to those of vision foundation models for image data. In general, synthetic tabular datasets have many applications [**?**, **?**, **?**] such as: enabling data sharing while maintaining privacy, reducing bias and improving representation of marginalized groups, simulating unseen domains, and augmenting the size of real data. Moreover, the large tabular model could be used for cleaning and pre-processing datasets [**?**]; finding relevant datasets (from different domains or general knowledge-bases, e.g. Wikipedia) [**?**]; augmenting existing datasets (e.g. performing SQL joins based on related data, such as adding a GDP column to a dataset with different countries); and conducting automated (meta)-analyses [**?**]. A notable example is a digital twin [**?**] based on tabular records of patients that can be used to personalize treatment plans by predicting disease progression and treatment outcomes [**?**].

Our research plan involves a phased approach: an initial research and data collection/cleaning phase (Months 1-3), followed by model development, i.e., pre-training, and iterative testing (Months 4-9), and culminating in an optimization and evaluation phase (Months 10-12). This timetable ensures systematic progress and allows for adjustments based on intermediate findings, ultimately leading to the delivery of a state-of-the-art tabular data foundational model by the end of the project year.

**B. Estimate of Compute, Storage and Other Resources**

The large tabular model will be pre-trained on a public dataset called Tablib [**?**, **?**], which consists of 627 million tables, totaling 69 TiB, and 867 billion context tokens. To support these experiments, substantial computational resources are required, particularly high-performance GPUs such as NVIDIA's A100 or H100, which will enable efficient parallel data processing and model training. Furthermore, the experiments necessitate a minimum of 100 TiB of storage space to accommodate the dataset and model outputs, as well as robust software platforms and tools suitable for handling such large-scale data, such as PyTorch.

**Computing need.** We will use a large neural architecture, such as Transformer or Mamba, to train our large tabular model. We split the entire training process into hundreds of independent computational tasks, each targeting a specific portion of the dataset. Each task is expected to require 50 GPU hours, which includes data loading, model iteration, and validation processes. To determine these computational costs, we analyzed historical data and training times for similar model training tasks, such as LLaMA model [**?**], taking into account the complexity of our model and the size of our dataset. To pre-train our large tabular model, we estimate that a total of approximately 50,000 A100 GPU hours would be required to complete the model training for this project. This estimate is based on the following analysis: Training the LLaMA-7B model using 1.5 trillion tokens would require 82,432 hours on A100 GPUs, and our training dataset contains about 0.867 trillion tokens, so we estimate that training of our base tabular data model will require about 50,000 A100 GPU hours. This estimate also takes into account the number of possible iterations and parameter tuning to ensure that sufficient resources are available to accomplish the research goals without being affected by resource constraints. In addition, regarding the distributed nature of computing, the computing tasks in this project can be executed in multiple resource partitions (including different physical or virtual machines) in a fully distributed manner.

**Storage need.** The storage requirements for this project are substantial, primarily due to the extensive volume of the datasets involved and the significant amount of model data generated during the training process. Specifically, the Tablib dataset employed in this project contains 69 TiB of data. When including intermediate data, model snapshots, and the final model files produced throughout the training sequence, the total estimated storage requirement for the project is approximately 100 TiB. Moreover, due to frequent data read and write operations during model training, a storage solution with not only ample capacity but also high-speed read and write capabilities is necessary. Such a solution is crucial to support efficient data processing and access speeds, thereby facilitating effective model training and data analysis.

**Usage need.** We plan to utilize this large tabular model to perform complex downstream tasks, especially in the understanding and generation of structured tabular data. This initiative will greatly increase the automation of processing large data sets, leading to increased efficiency and accuracy in a variety of application scenarios such as financial analysis, medical records management, and supply chain optimization. The automation provided by the model is designed to reduce human error and increase the speed of data insights, leading to a more informed decision-making process in these critical areas. In addition, the model will facilitate research and development of new algorithms and techniques. These innovations will enhance tabular data processing capabilities, introducing advanced predictive analytics and machine learning capabilities. This continuous development will ensure that our model remains at the forefront of technology, setting industry standards for data processing and analytics. As a result, the computational resources required

for large-scale tabular models should not be underestimated in terms of model usage. This often requires powerful GPU support to handle large amounts of data and complex calculations. Using a high-performance GPU such as the NVIDIA A100 or H100, a simple inference task such as a single inference on 10,000 records can take from a few minutes to more than ten minutes. For more complex or larger datasets, reasoning times can extend to hours.

**C. Support Needs**

Support from the staff at NAIRR Pilot Resource Providers will be essential for the successful completion of our project. Specifically, assistance with computational resources, software and tools, data management, and technical troubleshooting will be critical. For computational resources, we will need guidance on optimizing the use of high-performance GPUs, storage systems, and distributed computing environments. In terms of software and tools, support in setting up and configuring platforms like TensorFlow, PyTorch, and other data processing libraries will be necessary. Expertise in handling large datasets, including data preprocessing, storage management, and efficient data access during model training, will be crucial for effective data management. Additionally, help with debugging and resolving technical issues that may arise during the implementation and execution of our experiments will be indispensable.

The necessary support will include regular consultations, such as bi-weekly meetings with technical staff for project updates and troubleshooting, dedicated support hours for setting up and maintaining computational and storage resources, and access to documentation, best practices, and training materials related to the software and hardware being used.

Our project involves the use of a public dataset (Tablib [**?**]) and open-source software packages, so there are no export-controlled code or ITAR restrictions. We will ensure data privacy and security, even though the Tablib dataset does not contain personal health information (PHI) or HIPAA-restricted data. We will implement robust data security measures to protect the integrity and confidentiality of the data, including secure storage solutions, encryption, and access control mechanisms. If any proprietary data is used in future phases of the project, we will adhere to all relevant data usage agreements and confidentiality requirements. We do not anticipate any specific regional location restrictions for compute resources and we will ensure that our computational tasks are distributed accordingly to comply with any regional constraints that may arise.

**D. Team and Team Preparedness.**

This project is co-led by Prof. Guang Cheng and Dr. Chi-hua Wang at the Trustworthy AI Lab, UCLA. Prof. Cheng is the director of Trustworthy AI Lab and currently working on Generative AI with a particular focus on synthetic tabular data. Dr. Wang is a senior Postdoc at this lab with expertise on generative data science, and is in charge of administrative tasks, such as track the progress of project and communicate with stakeholders, in this project. Our team also consists of the following members in the lab: Mr. Minrui Gui is a first year computer science master student, in charge of the per-taining of the foundation tabular generation model. In the past, he successfully fine-tuned LLaMA7B; Mr. Xiaofeng Lin is a second year Ph.D student. He is in charge of the architectural design of the foundation model and had extensive experience working with large language models and deep learning algorithms in the table domain; Mr. Chenheng Xu is a first year master student, in charge of the data management and environment setup. He had experience in tabular synthesis; Mr. Peiyu Yu is a second year Ph.D. student, in charge of the downstream applications such as digital twin. He had experiences in multi-modal generative modeling.

## References


[1] Douglas Clements and Julie Sarama. Early childhood mathematics intervention. *Science (New York, N.Y.)*, 333:968–70, 08 2011.

[2] Linda Darling-Hammond. Teacher education and the american future. *Journal of Teacher Education*, 61(1-2):35–47, 2010.

[3] OECD. *Education at a Glance 2019: OECD Indicators*. OECD Publishing, Paris, 2019.

[4] Kun Zhou, Zhimeng Liu, Yizhou Wang, Wenbo Li, Jingyi Chen, Xingyu Xie, Yuxuan Wang, and Ziwei Liu. Emov2: Emotive avatar generation from text and speech. *arXiv preprint arXiv:2501.10687*, 2025.