

Evaluating SLMs for Predicting Tabular Data: An Essay on Higher Education Dropout

Fabrício G. Scaglioni¹, Marilton Aguiar¹, Júlio C. B. Mattos¹

¹Universidade Federal de Pelotas (UFPeI)
Pelotas – RS – Brazil

fabricao.scaglioni@inf.ufpel.edu.br,

marilton@inf.ufpel.edu.br, julius@inf.ufpel.edu.br

Abstract. *Student dropout in higher education represents a structural challenge for educational institutions, resulting in both academic and financial losses. Although classical machine learning models are successfully applied to this task, the consolidation of Large Language Models (LLMs) has enabled new predictive approaches. This study investigates the effectiveness of small-scale LLMs (7-14B parameters) adapted via Parameter-Efficient Fine-Tuning (PEFT) compared to specialized tabular-data models. Utilizing a UFPel dataset with 6,011 records, we evaluated three representation strategies: scaled (traditional), labeled (discretized), and storytelling (narrative), inferred by the Phi-4, Qwen2.5-7B, and Mitra models. The results indicate that the specialized tabular model (Mitra) outperformed the LLMs, achieving 80.30% accuracy and an F1-Score of 0.8203, compared to 72.07% and 0.7176 for the best-performing LLM (Phi-4 scaled). The labeled and storytelling linguistic formats yielded performance levels comparable to simple numerical formatting, invalidating the narrative advantage hypothesis in static prediction. Evidence shows that 4-bit-quantized small-scale LLMs perform worse on isolated tabular tasks. We conclude that tree-based and foundationally tabular classifiers remain the most effective architecture for dropout prediction under resource-constrained conditions.*

1. Introduction

Dropout in higher education impacts institutional sustainability and human capital formation. The early identification of students at risk of dropping out provides a foundation for public retention policies. Although tabular educational data is abundant, predictive modeling in this domain faces challenges, including inherent class imbalance. Recently, language models (LLMs) have demonstrated proficiency in unstructured tasks, but their isolated applicability to structured prediction from tabular datasets remains unquantified, especially under severe computational constraints. This study seeks to answer the following research questions:

- RQ1:** Are SLMs (7-14B parameters) fine-tuned with PEFT/LoRA effective for predicting student dropout based on tabular data?
- RQ2:** Which tabular data representation strategy (scaled, labeled, or storytelling) is most suitable for maximizing the predictive performance of LLMs?
- RQ3:** How do SLMs compare to specialized tabular data models (such as Mitra) and traditional machine learning models in scenarios with limited computational resources?

The main contributions of this work include several critical advancements in applying language models to educational datasets. First, it provides a **systematic empirical evaluation** of SLMs, specifically Phi-4 and Qwen2.5-7B, fine-tuned using PEFT/LoRA for dropout prediction in tabular data and compared against the specialized Mitra model. Additionally, this study presents the **first systematic comparison** of three distinct tabular data representation strategies for LLMs. These strategies are: (i) scaled, involving traditional ML preprocessing; (ii) labeled, utilizing discretization into textual categories; and (iii) storytelling, converting data into natural language narratives. Furthermore, the work identifies the **practical limitations** inherent to LLMs in tabular tasks, including persistent bias toward the majority class, inconsistent results across executions, and significant infrastructure challenges due to API and hardware constraints. These insights lead to the formulation of **evidence-based guidelines** regarding the optimal allocation of SLMs versus GBDTs in educational binary predictions under computational constraints. Finally, a **new institutional tabular benchmark** is established through the Mitra model, reaching an F1-Score of 0.82 on native data and thereby challenging the performance boundaries previously proposed by [Ma et al. 2024].

The remainder of this article is organized as follows: Section 2 reviews related work on LLMs, PEFT/LoRA, dropout prediction, and models for tabular data. Section 3 describes the academic dataset and the preprocessing strategies. Section 4 details the experimental methodology, including LoRA configurations, hyperparameters, and evaluation metrics. Section 5 presents and discusses the obtained results. Finally, Section 6 concludes the article and points out directions for future research.

2. Related Work

The field of Natural Language Processing (NLP) presents a technical fragmentation: the widespread interest in Large Language Models (LLMs) has coexisted with the development of Small Language Models (SLMs). This realignment is driven by computational limitations, data privacy regulations (**on-premise**), and **edge computing** deployments. The subsequent review isolates three areas: scalar adaptive techniques, injecting SLMs into tabular matrices, and domain funneling.

The practical application of language models expanded with the emergence of Parameter-Efficient Fine-Tuning (PEFT). [Hu et al. 2022] introduced Low-Rank Adaptation (LoRA), based on the premise that adapting pre-trained models occurs in a low intrinsic dimension. LoRA injects low-rank matrices into the attention layers, which reduces the number of trainable parameters by up to 10,000 times without compromising performance compared to full fine-tuning.

Shortly thereafter, [Dettmers et al. 2023] developed QLoRA, which combines 4-bit quantization with LoRA adapters. Consequently, it became possible to train state-of-the-art models with significantly reduced hardware requirements; the authors fine-tuned a 65-billion-parameter model on a single 48GB GPU, achieving accuracy comparable to that of 16-bit models.

Parallel to strict language injection, the rise of native Tabular Foundation Models proposes vector processing free of semantics. Dense architectures such as TabPFN-2.5, Real-TabPFN, Transformers with free cross-attention (TabICLv2), and continuous Em-

beddings (TaRL, TabEmbedBench) model distributions without transforming logarithmic values into rigid grammatical morphologies.

[Zhang and Robinson 2025] fundamentally structured this abstraction around the Mitra concept, pre-trained on synthetic relational datasets and operating via numerical distances over inactive matrices. This foundational model outperformed exclusive Gradient Boosting architectures in classifying static short limits, prompting proficiency comparisons with pure sequential Transformers (SLMs) on specific deterministic tasks.

While approaches such as GTL [Borisov et al. 2022] and graph-based adapters like BOLERO seek to mitigate the Modality Gap [Liang et al. 2022], their technical viability depends on high-performance infrastructure. In contrast, this study adopts the philosophy of efficient adaptation [Hu et al. 2022], prioritizing hardware constraints (T4 GPU) via Parameter-Efficient Fine-Tuning (PEFT).

Other researchers pursue a middle ground, such as [Do et al. 2023] and [Kesarkar et al. 2025], who utilize LLMs to create new features and enrich data semantics, subsequently feeding them into models such as XGBoost. [Han et al. 2024] follows a similar logic with FeatLLM, automating feature engineering. Furthermore, [Rabbani et al. 2024] demonstrated that smaller models, such as DistilGPT2, can compete with specialized methods when properly fine-tuned for simple tables.

Several cases exist where specialized SLMs outperform massive models or traditional methods. In **healthcare**, [Hosseini et al. 2025] utilized BERT and GPT-4 to predict opioid-induced respiratory depression, achieving superior results compared to linear regressions. [Hu et al. 2025] created SLM-MEEF for ECG reports, employing a Qwen-0.5B model with Mixture of Experts. Furthermore, [Li et al. 2025] developed a system to detect mental disorders by “distilling” the logic of larger models into a LLaMA3-8B.

In **security and code**, [Prasad et al. 2025] applied LoRA to a CodeT5+ model to classify vulnerabilities, outperforming boosting algorithms. In the field of reverse engineering, [Armengol-Estapé et al. 2024] introduced SLDe, a decompiler with only 200M parameters that outperformed Ghidra on optimized binaries.

SLMs have also reached **IoT and IT operations**. [Fukuda et al. 2025] demonstrated an agent for ICT operations that surpassed GPT-4 in specific tasks using nested thought reasoning techniques. [Paul et al. 2024] focused on smart home control, and [Ridwansyah and Saputro 2025] deployed music lyric generation directly on mobile devices, utilizing 8-bit quantization.

Finally, **multimodality** in compact models was explored by [Koska and Horváth 2024] with the EAGLE model, while [Thakur and Tyagi 2024] optimized spatial data retrieval with a Llama-2 model. These examples reinforce that fine-tuning for a specific domain is often more efficient than simply increasing the number of model parameters.

The quantified transition reported in the literature consolidates the shift from billion-parameter scalability to restricted situational efficiency. SLMs injected into a fine-tuning funnel retain a practical monopoly aimed at circumventing the inherent budgetary constraints of local laboratories and educational institutions.

The Unsloth library [Daniel Han and team 2023] optimizes fine-tuning with LoRA by applying techniques such as gradient checkpointing, optimized CUDA kernels, and efficient memory management. Benchmarks indicate up to 2x acceleration in training and an approximate 40% reduction in VRAM usage compared to standard implementations, enabling experiments on mid-range GPUs (such as the T4 available in Google Colab).

3. Dataset and Data Preparation

This section describes the utilized dataset, the preprocessing strategies, and the three data representation approaches evaluated in this study.

3.1. Academic Dataset

The dataset consists of academic records of students from the *Omitted for Review*. The predictive objective is the binary classification of **Dropout**, where instances of students with a “Graduated” status represent the negative class (0) and all other statuses (Dropout, Cancellation, Transfer, etc.) represent the positive dropout class (1).

The original file contains 7,191 raw records, of which 6,011 remain in the global view after unification and cleaning. A cross-sectional structure was adopted, in which each row aggregates the accumulated totals for a student in their first academic semester. Focusing on the initial period allows for dropout prediction as early as possible and ensures that each student is represented by a single row in the dataset. The variables (features) are divided into two main groups:

Demographic and Enrollment Data. Entry age, marital status, gender, race, SiSU code, public school indicator, higher education quotas, shift, time elapsed between high school and undergraduate enrollment, and the degree program’s fundamental area.

Performance and Attendance Variables for the Reference Semester. The baseline academic semester is linked to its consolidated quantitative assessments. It enumerates the number of courses, total credits, and grade point average, along with approved, exempted, failed, and withdrawn credits. It also records non-attendance records and the number of exams administered.

It is fundamental to clarify the structural matrix adopted to nullify any methodological deviation, such as temporal leakage or breach of independence. The modeling framework, by design, establishes that **each row in the dataset represents a unique student**. The performance markers **do not constitute elongated longitudinal series** and **did not undergo horizontal pivoting from subsequent quarters**. On the contrary, each occurrence exhibits absolute cross-sectional markers contained under the restricted static encapsulation of the individual’s transition. Therefore, when randomly partitioning the database into global training and validation splits, deterministic segregation is achieved between mutually exclusive individuals evaluated within the same static baseline window. This renders structurally unfounded the conjecture that future chronologies could act as reverse leakers or oracles of the final target class.

Unlike previous work on the subject, this dataset revealed that dropout accounts for approximately 58% of the evaluated clean baseline, making the dropout class the **majority**, requiring reevaluation of how bias and performance metrics are analyzed.

3.2. Cleaning and Numerical Preprocessing

The preprocessing architecture included removing null instances and duplicate rows. Integrity filters were applied, excluding anomalous records for entry age (values ≤ 0 or ≥ 100) and `time_between_highschool_university` (negative values). Identifying columns (`id`, `cod_course`, and `original_status`) were removed from the modeling stage.

Continuous variables were treated using logarithmic compression (`log1p`) to mitigate asymmetry, while categorical data received Label Encoding followed by Standard Scaling for standardized modeling.

3.3. Three Data Representation Strategies

In order to investigate which representation format optimizes LLM performance, three distinct formats were developed using the data:

1. **Scaled Format (Traditional ML):** Applies the pipeline detailed in the previous subsection, with log-transformed numerical variables and standardized general data. This format feeds the models as conventional tabular key-value pairs.
2. **Labeled Format (Discretization into Textual Categories):** Instead of continuous scalar normalization, the numerical delimitations of grades and distances (binning) were linearly partitioned into fixed static quintiles, without exhaustive analytical exploration for optimal splits. These categorical intervals were transposed into semantic keys in the “between_X_and_Y” format. This approach replaces extensive continuous vectors with pre-delimited categorical blocks, theoretically easing the LLM’s abstract inference.
3. **Storytelling Format (Natural Language Narrative):** Each student record is converted into a textual paragraph. To eradicate multilingual noise inherent in the attention processes of smaller models, the injected mask used only the native language (English) and mapped state variables via f-strings. The rigid template was structured as follows (with no-ablations): “This student is {age} years old, identifies as {gender} and {color}. The student started the course in {year}. Over this reference semester, they enrolled in {nr_tot} courses with an overall average grade of media. Are they at risk of dropping out? Answer strictly yes or no.” This standardization enforces adherence to the conversational linearity of the generators.

3.4. Data Division and Isolation (Data Leakage Prevention)

To separate the training and test sets from the main database, an 80/20 stratified split was used. The partitioning constituted the unified mandatory preliminary stage of the research, executed in isolation, a single time, prior to any modeling experiment. The subsequent subsets (Scaled, Labeled, and Storytelling) were strictly instantiated from the generated matrices, which were then maintained in an immutable state. This architectural structure prevents dynamic reprocessing at each model iteration, shielding the research from data leakage and ensuring that the same isolated test sample serves as the final, definitive benchmark for all tested LLMs and traditional models.

Oversampling mechanisms (such as the SMOTE algorithm) were discarded from the primary experimentation pipeline, as inserting synthetic neighbors into linguistic

neural networks degrades inference based on the true sample prevalence of educational dropout.

Methodological note regarding the test sample size: The 80/20 stratified split over the 6,011 valid instances generated a test subset (holdout set) containing $N = 1,203$ instances. All models, including the traditional baselines, the specialized Mitra model, and the SLMs (Phi-4 and Qwen2.5), were strictly evaluated and reported on this exact matrix of $N = 1,203$ records under the consolidated configurations, ensuring sample parity and invalidating class distortions.

4. Experimental Methodology

This section details the evaluated models, fine-tuning configurations with PEFT/LORA, hyperparameters, computational infrastructure, and evaluation metrics. We evaluate two small language models (SLMs) and a specialized tabular model: Phi-4 (Microsoft), an optimized Transformer with $\sim 14B$ parameters for edge infrastructure, and Qwen2.5-7B (Alibaba Cloud), a $\sim 7B$ parameter instructional model selected to evaluate inferential boundaries under long reasoning induction. Both SLMs were quantized to 4 bits using the `bitsandbytes` library to accommodate training constraints on an NVIDIA Tesla T4 GPU (15GB VRAM). As a specialized tabular baseline, we integrate Mitra [Zhang and Robinson 2025], a tabular foundation model that abstracts features from unknown domains in zero-shot inference via continuous encoding, bypassing textual prompting by optimizing networks directly over numerical arrangements. To isolate its baseline performance, automated orchestration (AutoML) and GBDT branches were disabled.

4.1. Tabular Modeling and Inference via Reasoning (Chain-of-Thought)

The tabular adaptation for **Storytelling** replaces feature engineering with linguistic conversion. Argumentative modulation injections (Chain-of-Thought) guide the SLM to draft the analytical chaining (e.g., “Analyzing the entry age combined with the numerical deficiencies...”) before issuing the definitive dichotomous decision $\langle yes \rangle$ or $\langle no \rangle$. Calculating the ROC AUC metric isolates the value judgment via Regular Expression (Regex) for direct mathematical measurement.

To neutralize continuous distribution misalignments and latency regression in $O(N)$ inference, the modular prompts (SYSTEM_MESSAGE) were statically parameterized over asynchronous batch processing (Batch Inference).

4.2. LoRA Configurations

The Low-Rank Adaptation (LoRA) quantization tensors were applied, isolating the predictive architecture to capture nonlinear correlations without altering the Foundation Model mesh. Structural details (Table 1(a)) and training parameterization (Table 1(b)) are provided in the following definitions. The Unsloth module and the **Sequence Packing** paradigm optimized token grouping, increasing decoding speed.

4.3. Training Hyperparameters and Constraint

The architecture excluded Early Stopping routines, as supporting independent evaluation datasets would excessively fragment the residual capacity on the entry-level T4 GPU

Table 1. LoRA Configuration and Training Hyperparameters

Parameters	Values
Rank (r)	32
Alpha (lora_alpha)	16
Dropout	0
Target modules	Auto-Maped
Bias	none
Gradient checkpointing	unsloth

(a) LoRA Configuration Parameters

Hyperparameters	Values
Max sequence length	1024 tokens
Learning rate	2×10^{-4}
Scheduler	Cosine Decay
Otimizador	AdamW 8-bit
Weight decay	0.01
Max steps	100
Warmup steps	5
Batch size efetivo	8
Max new tokens	500
Seed	3407

(b) Training Hyperparameters

(15GB of VRAM). Predatory overfitting was neutralized through a cosine Learning Rate Scheduler and a parameterizable numerical ceiling on active epochs, limited to 100 fine-tuning episodes.

4.4. Experimental Setup and Evaluation Metrics

The training environment was configured on the Google Colaboratory (Colab) platform utilizing an NVIDIA Tesla T4 GPU with 15GB of VRAM. The software stack comprised the `unsloth`, `transformers`, `trl`, `peft`, and `bitsandbytes` libraries, with an execution time of approximately 1 hour per experiment. Following this setup, a total of 7 configurations were evaluated: Mitra in raw format as a specialized baseline, along with the Scaled, Labeled, and Storytelling formats implemented for both Qwen2.5-7B and Phi-4.

Performance evaluation relies on standard metrics for binary classification with unbalanced classes, prioritizing the Weighted F1-Score as the primary metric. Since the instances operate on unbalanced sampling, using the macro formulation (raw average of harmonics) distorts rigorous comparisons. The reported metric isolates and weights each class's F1 score, then consolidates the scores based on the actual volumetric load (support) tested, mathematically justifying the separation observed in the combined balance of Precision and Recall. Regarding the ROC-AUC calculation in LLMs, liminal textual inferences (Phi-4, Qwen) were computed by fixing the generated binary prediction ('yes'/'no') converted to a closed absolute limit (point binary threshold of 0.5). Since extracting collinear probability matrices from exhausted logits (calibrated scores) did not remain consistent in the adopted LoRA/4-bit loading, the vector quantification of this curve is invariably reflected as flattened (AUC equalizing the exact Recall/Balanced Accuracy of a single class), substantially undersizing the extended continuous scalar expressiveness.

5. Results and Discussion

The **Mitra** model, performing inference on raw tabular data, demonstrated performance comparable to gradient-based models (XGBoost) and ensembles (Random Forest), recording an **F1-Score** of 0.8203 and an AUC of 0.8943. Considering its tabular foundation model architecture geared toward zero-shot inferences, Mitra establishes itself as a

state-of-the-art baseline compared to the traditional NLP implementations mapped in this study. Table 2 consolidates the performance of all evaluated models.

Table 2. Model Performance Comparison

Model	Format	Accuracy	Precision	Recall	F1-Score	ROC AUC
Mitra	Raw	0.8030	0.8050	0.8200	0.8203	0.8943
XGBoost	Scaled	0.8071	0.8000	0.8100	0.8000	0.8915
SVM	Scaled	0.8055	0.8000	0.8200	0.8000	0.8836
Random Forest	Scaled	0.8047	0.8000	0.8100	0.8000	0.8973
Phi-4	Scaled	0.7207	0.7201	0.7295	0.7176	0.7295
Phi-4	Labeled	0.7190	0.7124	0.7200	0.7134	0.7200
Phi-4	Storytelling	0.7099	0.7183	0.7264	0.7087	0.7264
Qwen2.5-7B	Scaled	0.6293	0.6412	0.8700	0.7383	0.5683
Qwen2.5-7B	Labeled	0.7016	0.9062	0.5615	0.6934	0.7370
Qwen2.5-7B	Storytelling	0.7232	0.8707	0.6335	0.7334	0.7459

The Mitra confusion matrix (Table 3(a)) shows that 541 students were identified as at risk of dropout. Fifty-five False Positives were observed in relation to graduate students. The ROC AUC of 0.8943 attests to the stratified tabular model's predictive separability.

The Phi-4 architecture configured in the scaled format achieved 72.07% accuracy and identified 496 true dropout instances (recall of 68.60%). Under the scaled paradigm, SLMs achieve satisfactory discriminatory precision, though their performance remains inferior to the Mitra classifier's ceiling. The confusion matrix can be observed in Table 3(b).

Table 3. Confusion Matrices

	Graduate	Dropout
Graduate	425	55
Dropout	182	541

(a) Mitra: Specialized Model

	Graduate	Dropout
Graduate	371	109
Dropout	227	496

(b) Phi-4 with Scaled Format

The conversion into linguistic attributes maintained the prediction bias toward the majority class. The Labeled format recorded an F1-Score of 0.7134, while the Storytelling format stabilized at 0.7087, demonstrating parity with the raw numerical Scaled format (0.7176). Strictly contextual descriptive additions did not result in linear gains in predictive abstraction. The confusion matrices are shown in Tables 4(a) and 4(b).

Table 4. Confusion Matrices

	Graduate	Dropout
Graduate	348	132
Dropout	206	517

(a) Phi-4 with Labeled Format

	Graduate	Dropout
Graduate	388	92
Dropout	257	466

(b) Phi-4 with Storytelling Format

Trials under the **Qwen2.5-7B** infrastructure empirically delimited the instability of SLMs subjected to under-training for majority classes (dropout $\approx 60\%$). In initial direct inference experiments (Scaled, limiting generation to `max_new_tokens=50` to extract

only the final label), the architecture exhibited severe generalization failure: the model began to predict the positive class unanimously (classifying 100% of cases as dropout). In the absence of a reasoning chain before the explicit verdict, the default learning rate ($2e-4$) induced immediate overfitting to the dominant class. Stabilizing the Loss function and restoring class distinction required reducing the learning magnitude to $2e-5$.

The inductive chain-of-thought prompting (Pseudo-CoT) reversed the static collapse previously imposed by sampling bias. Once the restriction in the generator library was corrected, the parameter quantization of the Qwen base model achieved 72.32% global accuracy and an ROC AUC of 0.7459 (F1-Score reached 0.7334).

This finding dismisses premature assumptions based on blocking narrative induction. Allocating a prominent space for referential chaining before the final output bypasses the original architectural constraints of the lightweight system (7B). This generates parallel scalar competitiveness against the limits attained by the denser structural family tested (14B). Table 5 compares the results with the baseline from [Ma et al. 2024].

Table 5. Comparison with Baseline [Ma et al. 2024]

Model	Study	Precision	Recall	F1-Score
GPT-3.5 Fine-tuned	Ma et al.	0.91	0.94	0.92
SVM	Ma et al.	0.81	0.83	0.82
Random Forest	Ma et al.	0.82	0.80	0.80
XGBoost	Ma et al.	0.80	0.80	0.80
Mitra	This study	0.81	0.82	0.82
Phi-4 Scaled	This study	0.72	0.73	0.72
Phi-4 Labeled	This study	0.71	0.72	0.71
Phi-4 Storytelling	This study	0.72	0.73	0.71
Qwen2.5-7B Scaled	This study	0.64	0.87	0.74
Qwen2.5-7B Labeled	This study	0.91	0.56	0.69
Qwen2.5-7B Storytelling	This study	0.87	0.63	0.73
SVM	This study	0.80	0.82	0.80
Random Forest	This study	0.80	0.81	0.80
XGBoost	This study	0.80	0.81	0.80

The fine-tuned GPT-3.5 model evaluated by [Ma et al. 2024] achieved an **F1-Score of 0.92**. This asymmetry stems from structural discrepancies and massive computational power beyond the scope of this metric. Key factors include the absence of restrictive local 4-bit quantization and a significant parameter disparity, totaling approximately 175 billion versus the $\leq 14B$ ceiling in this study. In turn, the decision tree-based architecture of the Mitra classifier obtained an F1-Score of 0.8203. These findings highlight the predictive limits of local SLMs when compared to rigorous purely tabular processing.

The observed performance proximity suggests that **Phi-4** has a robust capacity to interpret structured numerical prompts. Paradoxically, providing extensive and explicit narrativity to numerical data (Storytelling) did not yield qualitatively superior performance compared to the format that only relates keys and clean data (Scaled). Table 6 summarizes the performance of Phi-4 across the three formats.

Table 6. Comparison of Representation Formats (Phi-4)

Format	Accuracy	F1-Score	ROC AUC
Scaled	72.07%	0.7176	0.7295
Labeled	71.90%	0.7134	0.7200
Storytelling	70.99%	0.7087	0.7264

Unlike the Microsoft model’s invariable semantic robustness, the Alibaba-derived model (**Qwen2.5-7B**) showed severe sensitivity to basal structuring. The purely numerical format (Scaled), stripped of grammatical ties, led to atypical spikes in predicting the dropout class, resulting in a flattened discriminant threshold (AUC 0.5683). In this environment, providing precise contextual grounding via induction (Storytelling) fundamentally restored the model’s discriminative capability, equalizing precision and recall while consolidating the arrangement with the highest Accuracy (72.32%) and ROC AUC (0.7459). Table 7 summarizes the comparative performance of the Qwen2.5-7B architecture across the different morphological arrangements.

Table 7. Comparison of Representation Formats (Qwen2.5-7B)

Format	Accuracy	F1-Score	ROC AUC
Scaled	62.93%	0.7383	0.5683
Labeled	70.16%	0.6934	0.7370
Storytelling	72.32%	0.7334	0.7459

The limited inclusion of prior linguistic reasoning, specifically Pseudo Chain-of-Thought (Pseudo-CoT), reduced inferential stability and increased quantization sensitivity relative to the base architectures. Regarding stability, **Phi-4** successfully accommodated the verbosity overhead of Pseudo-CoT, maintaining its predictive capacity without degrading the extraction of the final class. Conversely, an architectural conflict was observed in **DeepSeek-R1-Distill**, which internally requires the free generation of reflection tokens; artificially limiting the generational window through the `max_new_tokens` parameter corrupted the continuous logical structure and degraded the reliability of the final binary inferences. Furthermore, **Qwen2.5-7B** exhibited significant hyperparameter sensitivity. The addition of induced and abstract narrative text in the prompt necessitated reducing the learning rate to 2×10^{-5} to prevent immediate overfitting to the majority class. This highlights the fragility of generalization to new syntactic formats. Ultimately, fixing universal generation parameters penalized models designed for variant structures. Imposing context truncation ceilings neglects models fundamentally based on free reasoning. Moreover, textual inductions without hyperparameter fine-tuning lead direct SLMs to classify instances based purely on inherent system imbalances.

The results reveal fundamental challenges in applying SLMs to tabular data. LLMs were pre-trained on massive textual corpora, developing internal representations optimized for linguistic semantics. Numerical tabular data represent a **fundamentally different modality**, where numerical values lack intrinsic semantics, and relationships are quantitative rather than qualitative. This limitation is corroborated by [Fang et al. 2024], who highlight the persistent hegemony of gradient-boosted decision trees (GBDTs) in high-dimensional settings and in purely numerical data, where LLMs frequently fail to

capture fine-grained statistical correlations without external assistance. Fine-tuning with LoRA may be insufficient to overcome this discrepancy, especially when only approximately 1% of the parameters are modified.

The success of **Mitra** (F1-Score 0.8698) validates the approach of models specialized in tabular data. Advantages include specific synthetic priors, the absence of extensive pre-processing, computational efficiency, and robustness in small datasets. This study presents the following methodological and practical limitations:

Transversal Structure and Airtight Isolation. The elementary criticism regarding dynamic leakage biases (individuals sequentially divided across chronological assessment blocks) is inapplicable to the established modeling, given that the matrix-vector required each row to represent a unique and unrestricted individual. The descriptive arrangement confines attributes solely within the closed perspective of the window linked to the `reference_semester`. By distancing predictive models from the premise of capturing multiple sequential sparse bands across the extended curriculum grid of a student, the mere random partitioning of samples yielded profiles that were inherently orthogonal to the grids. Thus, it is established that vectors and leaks originating from chronological contamination (Data Leakage) are prevented.

Hardware Limitations and Undertraining. Large Language Models were tested under limited computational resources, limiting their full parameterization. Training was constrained by the limited availability of a single T4 GPU (15GB), forcing a reliance on dense 4-bit quantization. The models were trained for at most 100 optimization steps, with a batch size confined to 8 and a single seed. Both Early Stopping and cross-validation partitioning were mandatorily suppressed, resulting in the structured absence of an isolated static validation matrix in the primary inferential temporal loop. It remains unclear how to decouple the syntactic restriction of the SLM architecture itself from the biases of failures extensively caused by undertraining.

Baseline Limitations and Analytical Robustness. In addition to the extensive parametric absence, comprehensive 10-fold cross-validation routines were not implemented in the extended scope. The 80/20 point partitioning, fixed at a single reference limit, excludes evidence-based assessments of standard deviations and Confidence Intervals (CIs). Furthermore, the addressed static baselines were restricted solely to XGBoost and Random Forest, omitting modern comparison with CatBoost, LightGBM, or recent equivalent dense tabular foundation models (TabPFN-2.5, Real-TabPFN, TabICLv2).

6. Conclusion and Future Work

This study investigated the feasibility of applying small-scale Large Language Models (7-14B parameters) fine-tuned with PEFT/LoRA for school dropout prediction using tabular data. In the evaluated context (7-14B models, 4-bit quantization, T4 GPU), **LLMs were not effective under the evaluated conditions (RQ1)**, as they were outperformed by specialized models (Mitra and XGBoost achieving global accuracy performance in the range of 80%). The evaluation confirmed comparable performance across the three syntactic formulations (**RQ2**). The numerical Scaled format demonstrated tangible mathematical

and computational stability, proving suitable for sequential processing. Injecting analytical morphological markers into running text (Storytelling, Labeled) matched the standardized baseline reference but failed to surpass it, which empirically refuted the premise of substantial gains linked to dense contextual semantics. Specialized models (Mitra and tree-based methods from AutoGluon) significantly outperformed LLMs in this experimental setup, outperforming LLMs by approximately 10 percentage points in F1-Score (**RQ3**). Mitra (AUC ≈ 0.89) confirms the dominance of tabular solutions in these educational datasets.

The practical implications of this study suggest that educational institutions should prioritize specialized models, such as Mitra, XGBoost, and TabNet, for tasks involving academic tabular data. Within these environments, LLMs should be reserved for complementary qualitative analyses, such as processing free-text responses and student feedback. Consequently, institutions must perform thorough cost-benefit assessments, as the resource-intensive process of fine-tuning LLMs may not be justified when compared to the performance and efficiency of classical models. From a research perspective, further investigation is required to identify the specific scale thresholds between models that fail to compete at 14B parameters and those that demonstrate superior performance at 175B. Moreover, the development of hybrid architectures that integrate LLMs with specialized numerical encoders represents a critical path for future inquiry.

Based on the limitations and insights identified in this study, several directions for future research are proposed. These initiatives include evaluating intermediate-scale LLMs, specifically those with 30B to 70B parameters, without using aggressive quantization. Furthermore, hybrid methods integrating LLMs for textual processing with specialized models for numerical data should be explored. Applying advanced prompt engineering techniques, such as Chain-of-Thought and Self-Consistency, and testing diverse datasets across multiple educational contexts and other tabular domains are also essential. Additionally, future work should prioritize developing explainability techniques to interpret LLM decisions in tabular contexts, alongside implementing calibration methods to stabilize predictions and reduce variability. Finally, investigating tabular-pretrained LLMs, which are models specialized for tabular data from the pre-training stage, represents a significant opportunity for further advancement.

Data demonstrate that specialized tabular models efficiently outperform local quantized SLMs (7B and 14B) in binary prediction tasks with class bias. The hypothesis that detailed textual narratives (Storytelling) would enhance the structural proficiency of language models was not mathematically confirmed, as performance remained analogous to the standardized use of isolated numerical variables (Scaled). For static academic dropout prediction, under severe computational hardware budget constraints with limited base memory (≈ 16 GB GPU RAM), tree-based foundation classifiers objectively maintain superior comparative effectiveness compared to textual processing networks. As SLMs iteratively become more capable through continuous distillation and as inferential optimizations (e.g., PEFT) overcome associated quantization bottlenecks, the cost-benefit ratio of these integrations must be reevaluated. Currently, for institutional scenarios with limited infrastructure, specialized models remain the technically justified choice.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Generative AI tools, such as Gemini and Grammarly, were used to refine the language during preparation. The authors did not fabricate data, alter experimental results, create references, or misrepresent their scientific interpretation. They reviewed and verified all content and take full responsibility for the final version of the manuscript.

References

- Armengol-Estapé, J., Woodruff, J., Cummins, C., and O’Boyle, M. F. P. (2024). SLaDe: A portable small language model decompiler for optimized assembly. In *2024 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 67–80.
- Borisov, V., Seiffert, T., Mayrhauser, H., Molnar, C., and Kasneci, G. (2022). Deep learning for tabular data: An extensive survey. *IEEE Transactions on Neural Networks and Learning Systems*.
- Daniel Han, M. H. and team, U. (2023). Unsloth.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized llms.
- Do, T., Shrestha Gurung, B. D., Aryal, S., Khanal, A., Chataut, S., Gadhamshetty, V., Lushbough, C., and Gnimpieba, E. Z. (2023). Utilizing XGBoost for the Prediction of Material Corrosion Rates from Embedded Tabular Data using Large Language Model. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4497–4499.
- Fang, X., Xu, W., Tan, F. A., Zhang, J., Hu, Z., Qi, Y., Nickleach, S., Socolinsky, D., Sengamedu, S., and Faloutsos, C. (2024). Large language models (llms) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*.
- Fukuda, N., Nozue, H., and Oishi, H. (2025). Small language model agent for the operations of continuously updating ICT systems. *IEEE access : practical innovations, open solutions*, pages 37522–37533.
- Han, S., Yoon, J., Arik, S. O., and Pfister, T. (2024). Large language models can automatically engineer features for few-shot tabular learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 1–26.
- Hosseini, E., Srinivas, A., Nazari, N., Hale, C., Rafatirad, S., and Homayoun, H. (2025). Large language models for opioid-induced respiratory depression prediction in hospitalized patients: A retrospective study. *ACM Transactions on Computing for Healthcare*, 6(2).
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

- Hu, F., Fang, Y., Si, W., Dou, J., Yang, Y., Qiang, Y., and Dong, X. (2025). SLM-MEEF: ECG report generation based on small language model and multi-expert ensemble framework. In *Proceedings of the 44th Chinese Control Conference (CCC)*, pages 8743–8748.
- Kesarkar, A. G., Deepakbhai Patel, T., and Vinayak Deshmukh, P. (2025). Hybrid Intelligence in Business Analytics: Integrating Large Language Models with Traditional Predictive Analytics. In *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pages 114–122.
- Koska, B. and Horváth, M. (2024). Towards multi-modal mastery: A 4.5B parameter truly multi-modal small language model. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 587–592.
- Li, M., Xu, X., Li, S., and Wu, B. (2025). Making Small Language Model Excellent Symptom Inference Expert for Mental Disorders Detection. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Nantes, France. IEEE.
- Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 17612–17625.
- Ma, X., Liu, W., Zhao, C., and Tikhvatulina, L. R. (2024). Can Large Language Model Predict Employee Attrition? In *Proceeding of the 2024 5th International Conference on Computer Science and Management Technology*, pages 1164–1172, Xiamen Guangdong China. ACM.
- Paul, S., Zhang, L., Shen, Y., and Jin, H. (2024). Enabling device control planning capabilities of small language model. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12066–12070. IEEE.
- Prasad, R., Singh, P. P., B, L., Maurya, A., D, S. N., and Singh, A. (2025). Sink vulnerability type prediction using small language model (SLM). In *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, pages 496–501. IEEE.
- Rabbani, S. B., Kowsar, I., and Samad, M. D. (2024). Transfer Learning of Tabular Data by Finetuning Large Language Models. In *2024 13th International Conference on Electrical and Computer Engineering (ICECE)*, pages 287–292.
- Ridwansyah, A. and Saputro, I. P. (2025). Implementation of small language model for on-device creative lyrics generation using LoRA and 8-bit quantization methods. In *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pages 552–555.
- Thakur, S. K. and Tyagi, N. (2024). Spatial data discovery using small language model. In *2024 International Conference on Computing, Power, and Communication Technologies (IC2PCT)*, pages 899–905.
- Zhang, X. and Robinson, D. M. (2025). Mitra: Mixed synthetic priors for enhancing tabular foundation models.