

Enhancing Taxpayer Risk Prediction through LLM-Driven Profile Tuning

Shubhradeep Nandi
Sr. Data Scientist, Data Analytics
Unit, GoAP
Vijayawada, India
shubnandi.ai@gmail.com

Kalpita Roy
Data Scientist, KSA
Bengaluru, India
kalpita.roy99@gmail.com

Abstract—Taxpayer risk prediction is pivotal for governmental financial departments globally. While traditional methodologies have automated the detection of potential risky taxpayers, their efficacy is waning due to two primary impediments: the continued reliance on outdated algorithms amidst the rapid emergence of generative AI and Large Language Models (LLMs), and the absence of a standardized metric to delineate genuine taxpayers. Addressing these challenges, this research introduces a groundbreaking approach centred on LLMs for taxpayer risk assessment. By integrating taxpayer transactional data into predefined templates, we harness LLMs to generate comprehensive natural-language profiles. These profiles are then employed to fine-tune expansive pre-trained models, facilitating nuanced risk predictions. Preliminary evaluations against established baselines attest to the superior performance of our proposed methodology. This research not only underscores the transformative potential of LLMs in financial oversight but also paves the way for future explorations into AI-driven fiscal governance.

Keywords—Taxpayer Risk Prediction, Large Language Models (LLMs), Financial Oversight, AI-Driven Fiscal Governance, Natural-Language Profiles.

I. INTRODUCTION

Utilizing machine learning approaches[3] has significantly enhanced the domain of risk forecasting[29,33]. These strategies also empower governments to more precisely gauge the risks linked to their tax payers[28,31,34], facilitating wiser decision-making. The mechanization of this procedure not only diminishes the scope for human mistakes but also augments both the speed and precision in determining tax payer risks in Goods and Service Tax (GST)[28] regime.

The scrutiny of comprehensive tax data repositories has sufficiently revealed that tax payer risk endeavors can be distilled down to profile categorization activities[1]. These activities entail scrutinizing taxpayer portfolios to assess the likelihood[2] of them infringing upon tax regulations. A tax payers portfolio generally encompasses details such as trade demographic data, business concern and activities, commodities purchases, produced or traded, proprietors background, sector status, historical financial transactions and tax filing patterns. Leveraging the commonality in attributes across diverse datasets over years, we advocate for a harmonization of data to facilitate model training for large models.

Taking positive advantage of the burgeoning advancements and remarkable capabilities of large language models (LLMs)[4,6], we introduce an innovative method designed to execute Profile Tuning utilizing LLMs[4] to aid

in tax payer risk prognostication. A schematic representation of our methodology is depicted in Figure 1. Initially, we populate a predetermined instruction blueprint with data from the individual serial of the table. Subsequently, we commission substantial language models to craft natural language narratives encapsulating the entirety of the information present in each row of the table. In the final step, we employ the generated profile narratives to refine large foundational models, coupled with a compact classifier, to facilitate predictive analyses.

Despite the rapid progress in classification algorithms, there remains a notable deficiency in the availability of tax payer datasets. This absence of a consolidated tax payer benchmark has hindered the growth of algorithms aimed at predicting tax payer risks. To address this, we worked to introduce a doubly encrypted repository of high-caliber data encompassing approximately 147K labeled entries. This dataset facilitates a comprehensive analysis with distinct training, validation, and testing segments. Beyond offering numerical X-y pairs for standard machine learning algorithms, we extend additional statistical data pertaining to each table to aid specialized classification algorithms. Moreover, we furnish detailed instructions and profile texts for each entry to support our profile tuning endeavors.

To assess the efficacy of the newly devised technique, we implemented Profile Tuning on various publicly available foundational models including GPT[8] and LLaMA[26]. This method was then weighed with a range of distinguished classification algorithms encompassing robust machine learning benchmarks like GBM[10,11,20], RandomForest[12] and XGBoost[14,19], as well as specially crafted neural networks such as Tabnet[17], VIME[16] and DeepFM[15]. The F1-score[22] metric was utilized to gauge the performance of the models, given that all the datasets are centered around classification tasks. Furthermore, during the training phase, a larger loss penalty was applied to positive samples, which represent risky instances, to counterbalance the uneven distribution present in the dataset.

The trials conducted using the comprehensive dataset affirm the sustained improvement in performance achieved through our approach when measured against other predictive benchmarks. Our findings further illustrate that applying Profile Tuning yields superior results, outperforming conventional classification models restricted to single-table operations. Moreover, we delved into alternative tuning strategies for large language models (LLMs[4,6]), including in-context learning and instruction tuning. Despite not excelling as classifiers when compared to other benchmarks, the results indicate that prompted LLMs[4,6] are capable of offering valuable tax payer risk guidance.

Summarizing the key contributions of this paper:

- We introduce an innovative method that converts tabular tax payer information into profile narratives, leveraging large language models for predictive analysis through Profile Tuning.
- We establish a benchmark aimed at taxpayer risk assessment, assembled from a collection of datasets that focus on three prevalent tax payer risks: default, fraud[3,32], evasion.
- Through rigorous testing we validate the effectiveness of this novel method against a spectrum of notable baseline models, thereby enhancing the comprehension of the role and potential of LLMs[4,6] in tax payer risk forecasting.

TabNet[17]. In the present study, we leverage the innovative Profile Tuning technique to work with tax tabular data. This involves the creation of cohesive tax payer profile narratives from diverse tax payer datasets, ushering in a fresh approach to tabular data processing in the large language models (LLMs[4,6]) generation.

B. Baseline Models as Pretrained

In the past few years, the rapid evolution of Transformer-based[23] baseline models, including architectures like BERT-inspired, T5[24]-influenced, and GPT-derived, has showcased the effectiveness of the pretraining-fine-tuning framework across a range of applications. There have been initiatives to adapt BERT[23] for textual data to undertake tasks such as

bus_actvty	bus_con	age	division	circle	e_way	adj_itc_per	turnover_diff	hsn_sac	label
wholesale	proprietor	8	kurnool	nandyal	12	37	15	72044900	0
retailer	partnership	5	anantpur	hindupur	20	83	32	94037000	0
retailer	partnership	4	nellore	ongole	55	92	45	85362030	1

Step 1: Consider the selected row to construct a concise Taxpayer(TP) profile description using the below information
bus_actvty: retailer; bus_con: partnership; age: 5; division: anantpur; circle: hindupur; e_way_bill: 20; adj_itc_per: 83;
turnover_diff: 32; hsn_sac: 94037000

Step 2: Natural Language Taxpayer (TP) profile created using Large Language Model

This Tax Payer is a 5 month old partnership firm engaged as a retailer in plastic furnitures.
It is located in hindupur circle of anatpur in Andhra Pradesh.

It has a total turnover diff of 32 lacs adjusting 83% of ITC and generating e_way bills of 20 lacs

Step 3: LLM based Profile Tuning

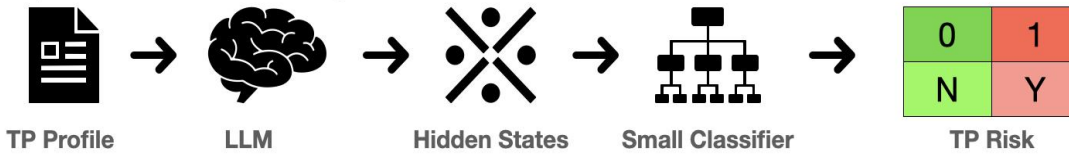


Figure 1: Overview of our novel method

II. PRIOR STUDIES

A. Tax Data Categorisation

Information related to taxation is traditionally structured in the form of tables, where numerous columns represent characteristics of tax payer or transactions. The machine learning sector has long focused on classification tasks involving tabular data[18], giving rise to an array of models designed for this purpose, including but not limited to RandomForest[12], XGBoost[14,19], and GBM[10,11,20]. The advent of deep learning has spurred the development of neural networks tailored to manage universal tabular data[18], with notable examples being DeepFM[15], VIME[16], and

sentiment interpretation, categorization of sentiments, and exploration of texts. Beyond the conventional fine-tuning approach that maintains consistency in the training procedure for subsequent tasks, newer large language models have demonstrated capabilities for in-context learning, which involves grasping tasks using minimal examples within a context. Consequently, experts have suggested fine-tuning LLMs[4,6] with meticulously crafted prompts to harness the vast potential inherent in LLMs[4,6].

In our novel method, we incorporate various publicly available large language models, such as LLaMA[26],

GPT[8] and BERT[23] while also weighing up our tuning approach with ICL[5] and instruction-based tuning.

III. OUR APPROACH

In this segment, we delineate the conceptual framework, mathematical representation, and execution strategy of the newly devised novel method. This innovative approach facilitates tax payer risk forecasting through Profile Tuning, utilizing the capabilities of large base models that are pre-trained.

A. Synopsis

The schematic representation of our novel approach can be observed in Figure 1, illustrating the three pivotal stages in our methodology, detailed below:

Stage 1: Initially, we establish an instructional blueprint, phrased as ‘‘Craft a succinct tax payer profile summary encompassing the ensuing details: PrfNGTP.’’ Here, PrfNGTP represents the amalgamation of ‘‘feature_name: feature_value,’’ pairings derived from individual rows of the proforma.

Stage 2: Subsequently, the directives formulated in the preceding step are fed into large language models, with the objective of generating articulate and logically structured tax payer profiles that encapsulate all the tabular data present in each row.

Stage 3: In the final phase, the linguistically crafted profiles from Stage 2 serve as the foundation for fine-tuning substantial baseline models such as BERT[23], GPT[8], and LLaMA[26]. Leveraging the hidden layers of these foundational models, a compact classifier — typically a feedforward neural network — is employed to discern the risk associated with each profile.

B. Data Security

In our endeavor to secure the Tax Payer data to the utmost degree, we have employed a robust encryption methodology synonymous with the one utilized in blockchain technology[30], namely the Secure Hash Algorithm (SHA) at the source of extraction. This cryptographic hash function ensures that data is encrypted in a manner that is virtually impervious to decryption, thereby forestalling any attempts at reverse engineering to retrieve the original data. The SHA function works by taking an input (or ‘message’) and returning a fixed-size string of bytes, typically a digest that is unique to each unique set of inputs. It is a one-way function, meaning that it is infeasible to reverse the process to reveal the original input. This method not only ensures the confidentiality and integrity of the data but also verifies that the data has not been altered or tampered with, thereby providing a secure and trustworthy foundation for our research. Leveraging the SHA encryption methodology, we have ensured that all sensitive information is shielded with a fortress of security, thereby upholding the highest echelons of ethical research and data privacy.

C. Methodology Outline

Let us consider a structured dataset S represented as $\{Z, b, r\}$. Here, $b \in \{0, 1\}^a$ signifies binary labels where $b_i = 0$ indicates the i -th entry is non-risky, and $b_i = 1$ indicates potential risky tax payer. $Z = \{z_1, \dots, z_a\} \in \mathbb{R}^{a \times c}$ represents a instances, with each $z_i \in \mathbb{R}^c$ having c attributes. The set r comprises c descriptors indicating the column headers.

Constructing Profiles: As outlined in the initial phase, we convert structured tax payer records into a predefined format Q_i by populating the PrfNGTP with $\{r_j : z_j^i\}_{j=1}^c$ for each i th entry. This results in a set of instructions Q with a elements. These instructions Q_i are then processed by prominent large language models like Open AI based GPT APIs. Post-generation, we acquire a collection of profiles, denoted as M , that encapsulates taxpayer related details in a coherent narrative format.

Tuning with Profiles: During the subsequent phase, we adapt pre-established foundational models $E: \mathbb{R}^s \rightarrow \mathbb{R}^{s \times f}$, such as BERT variants and GPT iterations, using the profiles. Here, s represents the token count in every input sequence, and f denotes the hidden state dimensions. Utilizing the official tokenizers associated with each model, the tokenized profile collection is represented as $M \in \mathbb{R}^{a \times s} = \{m_i\}_{i=1}^a$.

Given the extensive nature of some foundational models, adjusting all parameters can be computationally intensive.

As a solution, we attach a compact classifier $D: \mathbb{R}^{s \times f} \rightarrow \mathbb{R}^{f \times g}$ to the tail end of pre-trained models, where $g = 2$ represents the label categories. This classifier, a concise feed-forward neural network, performs binary classification based on the hidden states $h \in \mathbb{R}^{s \times f}$ generated by the foundational models E :

$$h = E(m) \quad (1)$$

For bidirectional foundational models like certain BERT[23] variants, we compute the average of the final hidden states h of all unmasked tokens:

$$h_{\text{avg}} = (1/s) \sum_{i=1}^s h_i \quad (2)$$

For auto-regressive foundational models similar to GPT, the hidden state of the last unmasked token is utilized for classification:

$$h_{\text{avg}} = h_s \quad (3)$$

Subsequently, h_{avg} is processed through classifier D to derive the taxpayer risk prediction $\hat{y}$:

$$\hat{y} = D(h_{\text{avg}}) \quad (4)$$

The cumulative loss L for all a instances is determined using Binary Cross Entropy (BCE) $O: \mathbb{R}^a \times \mathbb{R}^a \rightarrow \mathbb{R}$

Given the imbalanced nature of datasets, a higher weight $w_{\text{pos}} \in \mathbb{R}$ is applied to positive samples to facilitate cost-sensitive learning:

$$w_{\text{pos}} = \{|b| - \sum_{i=1}^a b_i\} / \sum_{i=1}^a b_i \quad (5)$$

The standard BCE loss vector $\bar{\ell}$ is:

$$\bar{\ell} = -\sum_{i=1}^a [b_i \log(\hat{y}_i) + (1 - b_i) \log(1 - \hat{y}_i)] \quad (6)$$

The weighted BCE loss is then computed as:

$$\ell = O(\hat{y}, b) = -(1/a) \sum_{i=1}^a [b_i \bar{\ell}_i w_{\text{pos}} + (1 - b_i) \bar{\ell}_i] \quad (7)$$

Multi-dataset Profile Tuning: For u structured datasets $S_{\text{all}} = \{S^i\}_{i=1}^u = \{Z^i, b^i, r^i\}_{i=1}^u$ and their corresponding profile sets $M_{\text{all}} = \{M^i\}_{i=1}^u$, we can utilize all profiles in M_{all} to fine-tune the foundational model E and assess performance on each dataset's test set S^i .

IV. PERFORMANCE METRIC

We introduce a standard design to assess the efficacy of machine learning models when handling both structured tabular data and narrative profile inputs. Initially, we aggregate a vast array of taxation data and subsequently narrow down our selection to a consolidated dataset tailored for tax payer risk assessment. Our selection criteria encompass dataset column relevance, and the efficacy of foundational models on the feature set of these datasets. This dataset categorizes taxation risks into three primary segments: default, fraud, and evasion.

This comprehensive dataset comprises approximately 147K labeled records. Each dataset is segmented into training, validation, and testing subsets. The testing subset constitutes 30% of the entire dataset, with the remaining data divided between training and validation in a 9:1 ratio. Every dataset is bifurcated into two label categories: '1' symbolizing a positive or precarious instance, and '0' indicating the negative or safe taxpayer. An examination of the positive (risky) sample ratio in Table 1 reveals an imbalance across datasets, underscoring the necessity for balance-centric techniques during the training phase. Our observations indicate that without implementing such techniques, F1-scores on the test set plummet to near zero.

Beyond the conventional X-y data pairs (X_{algo}, y) suited for standard machine learning models, we furnish supplementary statistical data for each table, catering to specialized classification algorithms. This includes data such as class count (class_total), feature count (feature_total), numerical column indices (num_indices), categorical column indices (cat_indices), categorical column dimensions (cat_dimensions), column titles (column_titles), and categorical column descriptors (cat_descriptions). Moreover, for enhanced adaptability, we supply both the directive ($X_{\text{guide_for_profile}}$) and the narrative profile ($X_{\text{narrative}}$) for each record, catering to Profile Tuning and other text-based input scenarios.

TABLE 1: STATISTICS OF TAXPAYER RISK DATASET

Tax Payer Risks	Description	Classes	Train[Positive %]	Validation[Positive %]	Test[Positive %]
Default	Dataset of tax defaulters or not	2	2.9	2.1	2.3
Fraud	Dataset of tax fraudsters or not	2	0.67	0.7	0.69
Evasion	Dataset of tax evaders or not	2	1.69	1.68	1.69

Taxpayer default: This refers to the failure of an individual or entity to meet their tax obligations, either by not paying the owed amount or by not filing the necessary tax returns.

This situation can arise among individuals, businesses, or larger entities, posing a challenge for tax authorities who must factor in the risk of taxpayer default when planning and executing their collection strategies.

Taxpayer fraud: This refers to intentional misrepresentation of financial details or transactions to evade tax obligations. This could encompass generation of false e-way bills, generation of large value of false invoices thereby acting as a facilitator of input tax credit(ITC) fraud, consumption of large false input invoices thereby manipulating tax due to be paid in cash to tax authorities. The primary objective is to determine the likelihood of a taxpayer to engage in fraudulent tax related activities.

Tax evasion: This refers to the act of deliberately avoiding the payment of tax owed to the government. This can be achieved through various means such as underreporting of sales, inflating expenses, not registering for GST[34] despite crossing the threshold, or engaging in shadow transactions that are not recorded. The main goal is to assess the probability of a taxpayer partaking in activities that lead to tax evasion.

V. EXPERIMENTAL FRAMEWORK

In this segment, we delve into the comprehensive experimental framework, encompassing the foundational models, core components of metrics, technical specifics, training procedures, and assessment criteria.

A. Foundational Models

To gauge the efficacy of dataset, we weigh up it against a spectrum of robust foundational models. This includes ensemble strategies rooted in decision trees and intricate neural networks tailored for structured data.

Algorithms driven by trees: We've opted for three models Random Forest[12], LightGBM[13], and XGBoost[19] as our tree-driven foundational models. Their selection is attributed to their stellar track record in diverse data-centric challenges. These models are either ensemble techniques or gradient-boosting algorithms that harness the power of decision trees.

Neural Network Architectures[18]: Our selection for neural networks tailored for tabular data comprises the following:

- DeepFM[15]: An industry-favored approach that amalgamates factorization mechanisms with deep learning, enabling feature extraction through a unique neural design.
- VIME [16]: A technique that integrates self-directed and semi-directed learning for structured data, leveraging value imputation and estimation masking.
- TabNet [17]: A transparent and efficient deep learning design that employs sequential attention for discerning relevant attributes in structured datasets.

B. Core Structure built on established Foundational Models

We integrate various pre-established foundational models to serve as the primary structural framework.

-BERT[23]: A renowned model rooted in bidirectional Transformer mechanisms, this encoder-centric language model is pre-trained using masked language modeling and subsequent sentence prediction tasks. It also has a Financial variant called FinBERT.

-GPT[7,8,9]: Rooted in the Transformer architecture, this decoder-centric language model is pre-trained using a progressive next-token prediction task, essentially focusing on language modeling.

-FLAN-T5[24,25]: A refined variant of the T-Model, it's tailored to handle a spectrum of zero-shot NLP and limited-shot contextual learning challenges.

-LLaMA[26]: A collection of publicly available, foundational language models that are large, trained on a vast token dataset. With models ranging from 7B to 65B parameters, the LLaMA-13B surpasses the performance of GPT across multiple benchmarks.

As highlighted in Section III, considering computational constraints, we employ a compact classifier a forward-propagating neural network to harness the hidden states generated by these foundational models. Furthermore, we selectively immobilize certain segments of the foundational models based on their complexity, as detailed in Table 2.

TABLE 2: EXPERIMENTAL SETTINGS FOR CORE

<i>model</i>	<i>params-all</i>	<i>params-trainable</i>	<i>Trainable modules</i>
BERT finBERT	110M	110M	All
GPT-2	117M	117M	All
FLAN-T5-Base FLAN-T5-XXL	220M 11B	220M 260M	Last Decoder Block
LLAMA-7B LLAMA-13B	7B 13B	202M 317M	Last Decoder Layer

C. Technical Specifics

We've constructed the foundational models utilizing their official source code or recognized libraries. For the Random Forest model, we've employed the scikit-learn framework. The XGBClassifier[19] is used for XGBoost, and LGBMClassifier[20] for LightGBM. Regarding neural network models, we've recreated DeepFM, VIME, and TabNet.

During the second phase of our primary process, each directive is processed through GPT via the OpenAI Python interface. The request is structured as follows:

```
import openai
response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",
    messages=[
        {"role": "system", "content": "a proficient taxation aide."},
        {"role": "user", "content": directive[i]},
    ], # The directive array is derived from Phase 1
    temperature=0,
)
```

For the foundational models in the third phase, we retrieve the model checkpoints from the HuggingFace platform, paired with their respective tokenization tools.

D. Training Specifics

Our experiments utilize two NVIDIA A40 GPUs, each having 48GB memory. Leveraging the Ampere architecture's BF16 mode, we employ mixed precision to optimize large foundational models. This approach hastens the training by executing operations in half-precision while retaining crucial network data by storing select details in single-precision.

For foundational model training, we set the batch size at 128 and cap epochs at 100. During dataset Profile Tuning, both the batch size and maximum sequence length are set to 128, with padding tokens mirroring the end-of-sentence tokens. If a dataset's sequence length exceeds 128 tokens, we adjust the batch size to 64 and the sequence length to 256. We employ the AdamW optimizer with a learning rate of 5e-5 and a weight decay of 0.01.

Training utilizes the provided sets from dataset, with models being updated using the training set and validated using the validation set. Post-training, the optimal checkpoint is loaded for testing. Experiments are replicated four times with varied random seeds, and average results are reported.

E. Assessment Criteria

Given the imbalanced nature of datasets, we adopt the F1-score as our primary evaluation metric for binary classification tasks. This metric offers a more nuanced assessment than accuracy, which could potentially yield a high number of false negatives in such scenarios.

VI. INSIGHTS AND OBSERVATIONS

This segment delves into the outcomes and interpretations of our experiments, as outlined in Section V.

A. Primary Findings

Table 3 showcases the F1-scores for tax payer risk predictions. Notably, our novel method consistently surpasses both tree-based techniques and prior neural frameworks, especially when applied to foundational models like GPT, T5, and Flan-T5. We further dissect the performance of various models below:

Tree-Based Model Insights: Among the tree-based techniques, Random Forest, XGBoost, or LightGBM XGBoost lead over others. The average F1-score between 43-46 suggests that our novel method poses a challenge, leaving scope for further enhancements.

Neural Network Model Observations: Within the neural networks tailored for tabular data, TabNet outshines its counterparts. Yet, these neural models don't match up to tree-based methods or our novel method, indicating that even specialized neural architectures struggle with risk prediction.

Full Tuning Analysis: Different foundational models exhibit varied predictive capabilities when integrated with our novel method. With Flan-T5 as the backbone, our novel solution consistently excels, highlighting its robustness in tax payer risk predictions. The edge of Flan-T5 underscores the value of instruction-based tuning. Additionally, GPT's performance surpassing BERT models indicates the potential of decoder-only Transformer models. The comparable outcomes between BERT and finBERT suggest that training foundational models on extensive financial data might not

always yield benefits, especially if the training and fine-tuning data diverge.

Selective Tuning Analysis: When tuning only the last Decoder layer/block, LLaMA consistently outperforms other foundational models. The superior performance of LLaMA-13B over LLaMA-7B aligns with LLaMA's findings. Impressively, even with limited trainability, LLaMA-13B exhibits commendable tax payer risk prediction scores, underscoring the effectiveness.

TABLE 3: TAXPAYER RISK PREDICTION F1 SCORE

Model Category	Model	Training	F1 Score
Tree-Based	Random Forest	Grid Search	43.11
	XGBoost	Grid Search	46.31
	LightGBM	Grid Search	45.91
Neural Network	DeepFM	Scratch	36.62
	VIME	Scratch	35.91
	Tabnet	Scratch	39.01
Tunning(All)	BERT	Tune All	44.63
	FinBERT	Tune All	44.08
	GPT	Tune All	46.83
	Flan-T5	Tune All	49.09
Tunning(Last)	Flan-T5	Tune Last	44.31
	LLaMA-7B	Tune Last	45.43
	LLaMA-13B	Tune Last	46.59

B. Alternative Approaches

As highlighted in Section II, leveraging in-context learning and instruction-based tuning has gained traction in the realm of expansive language models.

For in-context learning (ICL), we present ten samples to models like GPT, Flan-T5, and LLaMA, with the expectation that the model will generate an accurate response. Each sample is formatted as “Profile: S Response: Z” (Options: 1. Yes, 2. No. Response:”, where S is the target , and the model is expected to output the label Z in text format (“Yes” or “No”).

For instruction-based tuning (IT), we employ a directive to guide the model's predictions: “Assess if the subsequent taxpayer poses a risk. S Options: 1. Yes, 2. No. Response:”. Additionally, we've experimented with various detailed instruction prompts.

Given that foundational models occasionally produce ambiguous predictions, we consider the ICL and IT outputs accurate if the correct label text (“Yes” or “No”) is present in the response. Even with this leniency, both strategies yield subpar results (< 10 F1-score) when weighed up with other classification benchmarks. However, the outputs remain valuable as they offer supplementary insights, aiding human interpretation of the outcomes.

For instance: “Making a definitive assessment regarding a tax payers risk based on limited data is challenging. Yet, from the provided details, the taxpayer appears to be on the safer side. Taxpayer is a wholesaler, is older than 36 months in business, adjusts 40% of ITC, and has been associated for a considerable duration with a HSN[27]. Its increase in tax payment is moderately growing with respect to its sector, which isn't exceptional but isn't detrimental either. Hence, Option 2: No, the Taxpayer doesn't seem to be precarious, appears to be the more plausible conclusion.”

VII. CONCLUSION

For instance: In this paper, we introduce a novel method that transforms taxpayer tabular data into descriptive taxpayer profiles. These profiles are then optimized using Profile Tuning on foundational models(large) for predictive analysis. We also unveil a comprehensive dataset designed for assessing taxpayer risk predictions, encompassing datasets that focus on three prevalent tax payer risks: default, fraud, and evasion. The efficacy of our novel method is validated through comprehensive comparisons with several robust baseline models. Moreover, our in-depth analysis sheds light on the potential of utilizing foundational models in the realm of taxpayer risk forecasting.

REFERENCES

- [1] Murugan, M. & T., Sree Kala & Marappan, Raja. (2023). Large-scale data-driven financial risk management & analysis using machine learning strategies. Measurement: Sensors. 27. 100756. 10.1016/j.measen.2023.100756.
- [2] Lozano-Medina, Jessica & Hervet, Laura & Hernández-Gress, Neil. (2020). Risk Profiles of Financial Service Portfolio for Women Segment Using Machine Learning Algorithms. 10.1007/978-3-030-50436-6_42.
- [3] Chen, Yasheng & Wu, Zhuojun. (2022). Financial Fraud Detection of Listed Companies in China: A Machine Learning Approach. Sustainability. 15. 105. 10.3390/su15010105.
- [4] Wang, Yihan & Si, Si & Li, Daliang & Lukasik, Michal & Yu, Felix & Hsieh, Cho-Jui & Dhillon, Inderjit & Kumar, Sanjiv. (2022). Preserving In-Context Learning ability in Large Language Model Fine-tuning. 10.48550/arXiv.2211.00635.
- [5] Veres, Csaba. (2022). Large Language Models are Not Models of Natural Language: They are Corpus Models. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3182505.
- [6] Ruis, Laura & Khan, Akbir & Biderman, Stella & Hooker, Sara & Rocktäschel, Tim & Grefenstette, Edward. (2022). Large language models are not zero-shot communicators.
- [7] Veyseh, Amir & Lai, Viet & Démoncourt, Franck & Nguyen, Thien. (2021). Unleash GPT-2 Power for Event Detection. 6271-6282. 10.18653/v1/2021.acl-long.490.
- [8] Floridi, Luciano & Chiriatti, Massimo. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds and Machines. 30. 1-14. 10.1007/s11023-020-09548-1.
- [9] R, Roshini & A, Preethi & C, Arulmozhi & P, Sivaganga. (2023). Text Generator using GPT2 Model. INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. 07. 10.55041/IJSREM17751..
- [10] Natekin, Alexey & Knoll, Alois. (2013). Gradient Boosting Machines, A Tutorial. Frontiers in neurorobotics. 7. 21. 10.3389/fnbot.2013.00021..
- [11] Hosen, Md Saikat & Amin, Ruhul. (2021). Significant of Gradient Boosting Algorithm in Data Management System. Engineering International. 9. 85-100. 10.18034/ei.v9i2.559.
- [12] Ali, Jihad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.
- [13] Machado, Marcos & Karray, Salma & Sousa, Ivaldo. (2019). LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry. 1111-1116. 10.1109/ICCSE.2019.8845529.
- [14] Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.
- [15] Guo, Huifeng & Tang, Ruiming & Ye, Yunming & Li, Zhenguo & He, Xiuqiang. (2017). DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. 1725-1731. 10.24963/ijcai.2017/239.
- [16] Houthoofd, Rein & Chen, Xi & Duan, Yan & Schulman, John & De Turk, Filip & Abbeel, Pieter. (2016). VIME: Variational Information Maximizing Exploration.

- [17] Arik, Sercan & Pfister, Tomas. (2019). TabNet: Attentive Interpretable Tabular Learning.
- [18] Borisov, Vadim & Leemann, Tobias & Seßler, Kathrin & Haug, Johannes & Pawelczyk, Martin & Kasneci, Gjergji. (2021). Deep Neural Networks and Tabular Data: A Survey.
- [19] Santhanam, Ramraj & Uzir, Nishant & Raman, Sunil & Banerjee, Shatadeep. (2017). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets.
- [20] Alzamzami, Fatimah & Hoda, Mohamad & El Saddik, Abdulmotaleb. (2020). Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2997330.
- [21] Roumeliotis, Konstantinos & Tselikas, Nikolaos. (2023). ChatGPT and Open-AI Models: A Preliminary Review. Future Internet. 15. 192. 10.3390/fi15060192.
- [22] Goutte, Cyril & Gaussier, Eric. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Lecture Notes in Computer Science. 3408. 345-359. 10.1007/978-3-540-31865-1_25.
- [23] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- [24] Chung, Hyung & Hou, Le & Longpre, Shayne & Zoph, Barret & Tay, Yi & Fedus, William & Li, Eric & Wang, Xuezhi & deghani, Mostafa & Brahma, Siddhartha & Webson, Albert & Gu, Shixiang & Dai, Zhuyun & Suzgun, Mirac & Chen, Xinyun & Chowdhery, Aakanksha & Narang, Sharan & Mishra, Gaurav & Yu, Adams & Wei, Jason. (2022). Scaling Instruction-Finetuned Language Models. 10.48550/arXiv.2210.11416.
- [25] Suresh, Siddharth & Mukherjee, Kushin & Rogers, Timothy. (2023). Semantic Feature Verification in FLAN-T5.
- [26] Touvron, Hugo & Lavril, Thibaut & Izacard, Gautier & Martinet, Xavier & Lachaux, Marie-Anne & Lacroix, Timothée & Rozière, Baptiste & Goyal, Naman & Hambro, Eric & Azhar, Faisal & Rodriguez, Aurelien & Joulin, Armand & Grave, Edouard & Lample, Guillaume. (2023). LLaMA: Open and Efficient Foundation Language Models. 10.48550/arXiv.2302.13971.
- [27] Kar, Nilanjan & Sahore, Nidhi. (2018). Relevance of GST HSN Codes for Indirect Tax Implementation in India: A Review. 3. 39. 10.5958/2455-3298.2018.00006.5.
- [28] Kylasam Iyer, Deepa & Raghuram, G.. (2015). Goods and Services Tax: The Introduction Process. 10.13140/RG.2.1.4865.7520.
- [29] Coita, Ioana & Filip, Laura-Camelia & Kicska, Eliza-Angelika. (2021). TAX EVASION AND FINANCIAL FRAUD IN THE CURRENT DIGITAL CONTEXT. The Annals of the University of Oradea. Economic Sciences. 30. 187-194. 10.47535/1991AUOES30(1)020.
- [30] Faccia, Alessio & Mosteanu, Narcisa. (2019). Tax Evasion, Information Systems and Blockchain. Journal of Information Systems Management. 13.
- [31] López, César & Delgado Rodríguez, María Jesús & Santos, Sonia. (2019). Tax Fraud Detection through Neural Networks: An Application Using a Sample of Personal Income Taxpayers. Future Internet. 11. 86. 10.3390/fi11040086.
- [32] Abdallah, Aisha & Maarof, Mohd & Zainal, Anazida. (2016). Fraud Detection System: A survey. Journal of Network and Computer Applications. 68. 10.1016/j.jnca.2016.04.007.
- [33] Omar, Sinayobye & Kiwanuka, Fred & Swaib, Kaawaase. (2018). A state-of-the-art review of machine learning techniques for fraud detection research. 11-19. 10.1145/3195528.3195534.
- [34] Othman, Zaleha & Bidin, Zainol & Mansor, Muzainah & Abdullah, Yeop. (2019). GST fraud: Unveiling the truth. International Journal of Supply Chain Management. 8.