Expertise-Centric Prompting Framework for Financial Tabular Data Generation using Pre-trained Large Language Models

Subin Kim*
Financial Tech Lab, Kakaobank
South korea
luna.ns@lab.kakaobank.com

Jungmin Son*
Financial Tech Lab, Kakaobank
South korea
elena.son@lab.kakaobank.com

Minyoung Jung*
Korea Electronics Technology Institute
South korea
minyoung.jung@keti.re.kr

Youngjun Kwak†
Financial Tech Lab, Kakaobank
South korea
vivaan.yjkwak@lab.kakaobank.com

Abstract

Access to financial tabular data is often restricted owing to strict regulations surrounding personal information. Despite the advanced generative capabilities of large language models (LLMs), methodologies for the effective creation or expansion of financial tabular datasets remains undeveloped. The complexity of attribute relationships and the diverse data ranges in financial services present significant challenges in processing and understanding these datasets. To address these issues, we propose an expertise-centric prompting framework for synthesizing realistic and accessible pseudo-financial data. This framework involves a collaboration between financial experts and LLMs, focusing on schema calibration and attribute constraints. Moreover, we introduce new metrics to evaluate the realism of these pseudo datasets. We validated the effectiveness of the proposed framework and metrics on both English and Korean datasets, encompassing card transactions, loan statements, and deposits and savings, utilizing pre-trained LLMs such as KoGPT, ClovaX, LLAMA 2-Chat, GPT-3.0, and ChatGPT-3.5/4.0.

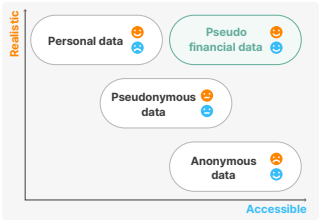
1 Introduction

Artificial intelligence has rapidly advanced financial services in sectors such as online banking, payment systems, investment, and fraud detection [1, 2, 3, 4]. However, the existing stringent policies and regulations pertaining to privacy and security restrict access to the financial tabular datasets of customers [5, 6, 7]. Consequently, pseudonymized or anonymized datasets are often employed to mitigate these concerns [8, 9], as illustrated in Fig. 1a. However, the limited availability of public datasets and the alteration of their original distributions through anonymization or pseudonymization pose significant challenges in the accurate representation of real-world scenarios. This shortage of authentic and accessible financial datasets impedes the advancement of artificial intelligence models for various financial tasks, including anti-money laundering and business confidence indices [10, 11].

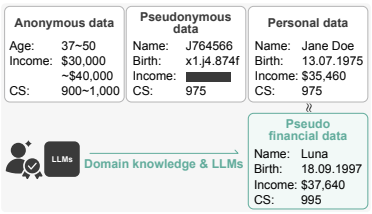
Large language models (LLMs), such as the GPT series [19, 20], exhibit remarkable proficiency in comprehending and generating texts for a wide range of tasks [21, 22, 23], including data construction

*Equal contribution
†Corresponding author

Table Representation Learning Workshop at NeurIPS 2024



(a) Comparison of financial data



(b) Example of pseudo-financial data

Figure 1: Comparative Analysis of Financial Data: (a) The y-axis represents the degree to which the data resembles authentic data from financial institutions, whereas the x-axis denotes accessibility. (b) Example of pseudo-financial data alongside public financial data, illustrating how pseudo-financial data, created through collaboration between financial experts and pre-trained LLMs, closely resembles personal data, with "CS" indicating credit score.

Table 1: Comparison between existing public financial datasets and pseudo-financial datasets.

Datasets	# Rows / Subj.	# Subj.	# Att.	Multilingual Data	Financial Constraint Check	Online Status Availability	Data Source	Data Year	Descriptions
Credit Card Payment [12]	15.0	100	13				PSPD Bank	-	Card Application Details / Payment History
Loan Approval Data [13]	1.0	614	12				Statista	-	Loan Application Information
Luxury Loan [14]	1.0	1,678	25				Data.world	1999	Loan Execution Details
Default of Credit Card Clients [15]	1.0	30,000	8				UCI	2005	Credit Card Default Data
Bank Loan [16]	1.0	5,000	14			✔	Thera Bank	-	Sole Internet Banking
German Credit Data [17]	1.0	1,000	21				UCI	1975	Anonymized Credit Card Details
Completed Order [18]	1.7	3,758	6				Czech Bank	1999	Bank Transfer Details
Ours (Card Transactions)	4.0	1,000	10	✔	✔	✔	LLM-generated	-	On / Offline Card Application Details
Ours (Loan Statements)	4.0	1,000	14	✔	✔	✔	LLM-generated	-	On / Offline Loan Application Details
Ours (Deposits and Savings Statements)	4.0	1,000	12	✔	✔	✔	LLM-generated	-	On / Offline Deposit Application Details

and augmentation [24, 25, 26, 27, 28]. Despite the impressive generative capabilities of LLMs, significant challenges remain in the generation and enhancement of financial tabular datasets, stemming from the intricate attribute relationships and diverse data ranges inherent in financial services [8, 29].

In this study, we address the deficiency of public financial tabular datasets and propose a novel approach termed the **expertise-centric prompting (ECP) framework**. Our ECP framework involves a collaboration between financial experts and pre-trained LLMs. The framework integrates two essential components into in-context learning: 1) schema calibration, which assesses the alignment of LLMs with primitive financial attribute prompts, and 2) attribute constraints, which produce balanced instances that closely resemble real-world financial data, as depicted in Fig. 1b. With this framework, we can generate pseudo-financial datasets, as illustrated in Fig. 1a, which are realistic and accessible. Furthermore, combining experts and LLMs, our framework enables the creation of unique multilingual datasets with new attributes not found in public datasets, as shown in Table 1.

Along with the ECP framework, we also propose two evaluation metrics for assessing the realism of the generated datasets. When evaluating the outputs from generative language models, metrics such as ROUGE [30], BERTscore [31], BLEURT [32], and COMET-22 [33] have been employed. However, evaluation metrics that are specifically tailored to synthetic tabular datasets in the financial domain have not yet been developed. To assess the realism of financial datasets, we introduce two evaluation techniques: one for assessing dataset diversity, encompassing inter-instance and intra-attribute diversities, and another for evaluating numeric constraint satisfaction. Our findings reveal that the collaboration between superior pre-trained LLMs and experts, facilitated by the proposed ECP framework, consistently produces superior results according to the proposed metrics.

Our contributions are summarized as follows:

- In this paper, we propose an expertise-centric prompting (ECP) framework to generate pseudo-financial tabular datasets, addressing the limited accessibility of realistic financial data.
- We introduce novel evaluation metrics to assess the diversity of the generated financial tabular datasets and the confidence level in constraint satisfaction.
- This paper demonstrates the efficacy of the proposed ECP framework and evaluation metrics on both English and Korean financial contexts across three distinct data types.

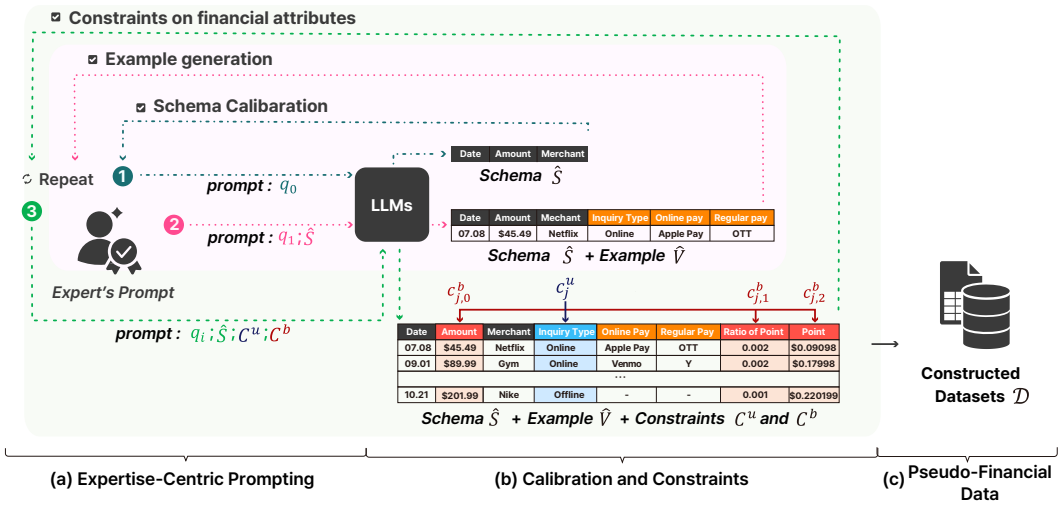


Figure 2: Overview of our expertise-centric prompting framework

2 Related Work

Public Financial Tabular Datasets. As listed in Table 1, several financial tabular datasets have been publicly released while adhering to regulations on personal data protection [7, 6, 5]. However, these datasets often exhibit limitations in terms of the availability of limited financial attributes and loss of personal information due to anonymization or pseudonymization [12, 13, 14]. Furthermore, these datasets are limited to English and lack the sophistication required for real-world applications owing to the absence of constraint satisfaction checks. To address these limitations, this paper presents the expertise-centric prompting framework for constructing more practical financial tabular datasets.

Financial Tabular Dataset Generation via Large Language Models. Despite numerous approaches leveraging LLMs for text-to-table construction [34, 35, 36, 37], these methods have limitations in generating financial tabular datasets. Specifically, models like BloombergGPT [38] and FinGPT [39] struggle with foundational financial data due to their training on non-tabular sources (e.g., news and social media), lacking a comprehensive understanding of financial tables. Inspired by the work of KiPT [40], which enhances prompts with domain-specific knowledge, we introduce a novel task to generate more informative financial data resembling publicly available datasets.

Metrics for Validating Diversity on Large Language Models. Recent research has focused on evaluating outcomes from pre-trained LLMs. CRITIC [41] uses external tools like search engines and code interpreters to refine outputs, enhancing quality iteratively. In contrast, SelfCheckGPT [42] identifies hallucinations by cross-referencing multiple outputs for consistency.In the finance domain, FinLMEval [43] proposes an evaluation metric tailored to appraise the performance of LLMs, with a particular emphasis on financial classification tasks. Unlike prior methods, our approach assesses the quality of LLM-generated financial datasets by distinguishing between numeric and categorical data, evaluating dataset diversity for the first time. We also introduce metrics to measure constraint satisfaction confidence, crucial for producing realistic and accurate data.

3 Methodology

In this section, we propose a new approach called an ***expertise-centric prompting framework*** for financial tabular data generation. Figure 2 illustrates this framework, which integrates schema calibration and attribute constraints through in-context learning with pre-trained LLMs [44, 45, 20].

Our framework constructs financial datasets that closely resemble real-world data by incorporating qualitative insights from experts, including schema calibration and constraints. Along with this framework, we introduce novel evaluation metrics designed to assess the realism and professionalism

of the financial dataset. By leveraging these metrics, we can ensure the quality of datasets generated through our proposed framework.

The following subsections elaborate on the framework and the formulation of metrics.

3.1 Prompt Design

We design prompts to interact with pre-trained LLMs for the construction of financial tabular datasets. Each prompt incorporates templates that format schema calibration and attribute constraints. In practice, the prompt template is designed as follows:

$$t(q_i) = \begin{cases} q_i, & \text{if } i = 0 \\ q_i; C, & \text{otherwise} \end{cases} \tag{1}$$

where a query q_i denotes the i^{th} prompt of a financial tabular dataset $\mathcal{D}$ for in-context learning [19]. C represents a set of K conditions $\{c_{j=0,\ldots,K-1}\}$ for schema calibration and constraints related to previous queries and $\mathcal{D}$, and the semicolon indicates a concatenation operator.

3.2 Expertise-Centric Prompting (ECP)

3.2.1 Calibration of Tabular Schema

As shown in Fig. 2, we employ a query-answer pair (q_0, S), where S is a tabular schema crafted by domain experts to refine the schema $\hat{S}$ from the pre-trained LLMs for financial tabular dataset $\mathcal{D}$. The schema $\hat{S}$ is formulated as follows:

$$\hat{S} = \mathcal{LLM}(t(q_0)), \tag{2}$$

where q_0 is the initial formatted prompt designed to assess the comprehensive understanding of the $\mathcal{LLM}$. $\hat{S} = \{\hat{a}_{i=0,\ldots,|\hat{S}|-1}\}$ contains the estimated attributes $\hat{a}_i$ of $\mathcal{D}$ obtained by the $\mathcal{LLM}$. To conform to the schema $S = \{a_{i=0,\ldots,|S|-1}\}$, which is provided by human experts, the generated schema $\hat{S}$ is retained or substituted as follows:

$$h(\hat{S}) = \begin{cases} \hat{S}, & \text{if } S - \hat{S} = \emptyset \\ S, & \text{otherwise} \end{cases} \tag{3}$$

where $h(\hat{S})$ produces the calibrated attributes intended for human experts' descriptions. This allows for the creation of new attributes, such as online-related variables or interest rates, which are not present in publicly available datasets, through the collaboration between pre-trained LLMs and financial experts.

3.2.2 Constraints on Financial Attributes

Based on the $\hat{S}$ of each dataset $\mathcal{D}$, we generate examples $\hat{\mathcal{V}} = \{\hat{v}_{h,w}, 0 \leq h < H, 0 \leq w < W\}$ for each corresponding to row h and column w using $\mathcal{LLM}$:

$$\hat{\mathcal{V}} = \mathcal{LLM}(t(q_1) = q_1; \hat{S}), \tag{4}$$

where $\hat{S}$ represents our calibrated tabular schema of C for the requisition of initial examples.

Finally, attribute constraints in in-context learning apply unary and binary constraints when generating instances for the given attributes, preserving the characteristics of financial datasets. These constraints can help the model focus on the precise distribution of each attribute, ensuring the generation of realistic and accurate values. The prompt template with schema calibration and the constraints in Eq. 1 is re-formulated as follows:

$$t(q_i) = q_i; C = q_i; \hat{S}; C^u; C^b, \tag{5}$$

where $C^u = \{c^u_{j=0,\ldots,K^u-1}\}$ denotes a set of K^u unary constraints for single-attributes, whereas $C^b = \{c^b_{j=0,\ldots,K^b-1}\}$ indicates a set of K^b binary constraints for relationships between multi-attributes. We specify that the unary constraints C^u in Eq. 5 impose bounds on row values in c^u_j-conditioned columns as follows:

$$\Psi : c_j^u \rightarrow \hat{v}_{\bullet,c_j^u}, \tag{6}$$

where Ψ represents unary constraints on the c_j^u-conditioned columns to generate pseudo-examples while ensuring realistic instantiation. We define a binary operation $\Phi : \hat{\mathcal{V}} \times \hat{\mathcal{V}} \rightarrow \hat{\mathcal{V}}$ that underlies our binary constraints, based on either a single binary operation or a combination of multiple binary operations. This is applied as follows:

$$\Phi(c_j^b) = \hat{v}_{\bullet,c_{j,l}^b} \odot \hat{v}_{\bullet,c_{j,m}^b} = \hat{v}_{\bullet,c_{j,n}^b}, \tag{7}$$

where c_j^b is a binary constraint, and $\odot$ denotes a binary attribute constraint involving columns l, m, and n to construct pseudo-examples while ensuring a predefined logical relationship. As depicted in Fig. 2, an example of a binary operation $\Phi(c_j^b)$ can be described as follows: $\hat{v}_{\bullet,c_{j,0}^b} \odot \hat{v}_{\bullet,c_{j,1}^b} = \hat{v}_{\bullet,c_{j,2}^b}$.

3.3 Evaluation Metrics for Pseudo-Financial Data

Diversity, a pivotal metric, seeks to bridge the gap between the generated and real-world financial data, which often experience a loss of diversity due to anonymization. This diminution in diversity adversely affects the performance of machine learning models in finance [46, 47, 48, 49]. Our aim is to generate data that preserves the diversity inherent in real-world financial data. Therefore, we introduce a novel evaluation metric tailored to gauge the diversity of pseudo-financial datasets generated by our ECP framework.

Nevertheless, an excessive divergence from expected ranges or a lack of consistency in the relationships between financial attributes raises concerns about unrealistic diversity. To mitigate this issue, we propose another new metric for constraint satisfaction to quantify the confidence of numeric operations. With the proposed metrics, we evaluate whether the generated values adhere to feasible constraints, ensuring that the exhibited diversity is grounded in financial reality.

3.3.1 Evaluation of Diversity

To assess the variety of the generated financial tabular datasets, we introduce the inter-instance and intra-attribute diversities. Inter-instance diversity measures the extent to which LLMs generate a set of values in a row without repetition. In contrast, intra-attribute diversity assesses whether each column contains varied values, ensuring that attributes are not redundant across instances.

Inter-Instance Diversity. We assume $\mathcal{F} : X \rightarrow R^d$ indicates a pre-trained ChatGPT-3.5 tasked with extracting latent features e for each row, where $X = \underset{w}{concat}[\hat{v}_{h,w}]$ is constructed by concatenating all columns. To evaluate the diversity among instances, we utilize the uniformity metric $Unif$ [50] defined as follows:

$$Unif \simeq f_{unif}(\underset{h}{concat}[e_h]), \tag{8}$$

where f_{unif} is implemented in Fig. 3 of Appendix B. The *uniformity score* is subsequently normalized to the $[0, 1]$ range using a sigmoid function.

In addition, we apply principal component (PC) analysis [51] to reduce the dimensionality of the embedding e, transforming it into PC while retaining the embedding's semantic meaning. This reduction aids in assessing the diversity coverage of the generated examples, enabling a more tractable analysis of the data's variance and distribution. The ratio of PCs indicates the number of PCs required to describe the data. A higher ratio implies that the data is diverse and not monotonous.

Intra-Attribute Diversity. With the pseudo-financial instances $\hat{\mathcal{V}}$, we employ entropy to evaluate the diversity within individual attributes as follows:

$$\mathcal{H} = -\sum_{i \in I} p_i * \log p_i, \tag{9}$$

where I indicates a set of unique examples on a certain attribute, and p_i represents the probability of each distinct value i. In a multi-attribute set J, we apply average and maximum operations such as $Avg(\mathcal{H}) = \frac{1}{J}\sum_j \mathcal{H}_j$ and $Max(\mathcal{H}) = \max_j[\mathcal{H}_j]$.

3.3.2 Evaluation of Constraint Satisfaction

To evaluate the alignment between the generated dataset and specific constraints, we introduce metrics for each unary and binary constraint. We define a metric for unary constraint satisfaction as follows:

$$\rho = \frac{1}{K^u} \sum_{c_j^u \in C^u} \mathbb{1}(\Psi(c_j^u)), \tag{10}$$

where ρ denotes the confidence of LLMs in data generation adhering to the corresponding constraint set C^u. Furthermore, we devise a metric for binary constraint satisfaction as follows:

$$\tau = \frac{1}{K^b} \sum_{c_j^b \in C^b} \mathbb{1}(\Phi(c_j^b)), \tag{11}$$

where τ represents the confidence of the outputs generated by LLMs in adhering to the corresponding constraint set C^b. Consequently, ρ and τ ensure that the pseudo-financial examples exhibit realistic diversity while satisfying specific attribute bounds and maintaining accurate logical relationships.

Table 2: Evaluation results related to diversity on the card transactions dataset: Inter-instance and intra-attribute diversities of the datasets generated by our approach are evaluated against those of the public datasets.

Language	Card Transactions Dataset	ECP	Diversity of Instance		Diversity of Attribute↑					
			Embedding-Level		Categorical-Level		Binary-Level		Numerical-Level	
			Uniformity↓	PC↑	Avg($\mathcal{H}$)	Max($\mathcal{H}$)	Avg($\mathcal{H}$)	Max($\mathcal{H}$)	Avg($\mathcal{H}$)	Max($\mathcal{H}$)
English	*Credit Card Payment* [12]		0.351	0.056	3.714	3.714	N/A	N/A	8.287	8.287
	KoGPT [52]	✔	0.365	0.040	2.652	5.174	1.072	1.284	4.823	4.823
	ClovaX [44]	✔	0.369	0.003	1.953	2.574	1.093	1.119	2.574	2.574
	LLAMA 2-Chat [45]	✔	0.339	0.036	3.737	4.934	0.971	0.971	5.057	5.057
	GPT-3.0 [19]	✔	0.340	0.060	3.190	5.461	0.963	1.000	3.874	3.874
	ChatGPT-3.5 [20]	✔	0.320	0.080	4.627	6.115	0.998	1.000	4.689	4.689
	ChatGPT-4.0 [53]	✔	0.327	0.081	3.987	6.483	0.954	0.990	5.751	5.751
Korean	KoGPT [52]	✔	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	ClovaX [44]	✔	0.396	0.006	1.807	2.722	0.961	1.000	3.322	3.322
	LLAMA 2-Chat [45]	✔	0.379	0.028	3.718	4.663	1.526	1.640	3.715	3.715
	GPT-3.0 [19]	✔	0.396	0.053	2.743	4.941	0.993	1.133	4.060	4.060
	ChatGPT-3.5 [20]	✔	0.388	0.107	3.412	6.723	1.024	1.101	6.096	6.096
	ChatGPT-4.0 [53]	✔	0.387	0.082	3.879	7.085	0.321	0.584	6.176	6.176

Table 3: Evaluation results related to diversity on the loan statements dataset: Inter–instance and intra-attribute diversities of the datasets generated by our approach are evaluated against those of the public datasets.

Language	Loan Statements Dataset	ECP	Diversity of Instance		Diversity of Attribute↑					
			Embedding-Level		Categorical-Level		Binary-Level		Numerical-Level	
			Uniformity↓	PC↑	Avg($\mathcal{H}$)	Max($\mathcal{H}$)	Avg($\mathcal{H}$)	Max($\mathcal{H}$)	Avg($\mathcal{H}$)	Max($\mathcal{H}$)
English	*Loan Approval Data* [13]		0.437	0.054	1.577	1.577	0.882	0.985	6.225	7.102
	Luxury Loan [14]		0.367	0.062	6.790	10.713	N/A	N/A	5.141	7.922
	Bank Loan [16]		0.440	0.058	7.484	7.484	0.652	0.967	4.305	6.414
	GPT-3.0 [19]	✔	0.435	0.046	7.018	7.295	0.956	0.999	3.025	4.634
	ChatGPT-3.5 [20]	✔	0.430	0.062	6.332	6.969	0.937	1.000	3.395	5.631
	ChatGPT-4.0 [53]	✔	0.431	0.056	6.995	7.146	0.999	1.000	3.205	4.268
Korean	GPT-3.0 [19]	✔	0.454	0.027	3.409	3.661	0.930	0.990	2.425	3.908
	ChatGPT-3.5 [20]	✔	0.446	0.078	7.084	7.417	0.702	0.995	3.267	5.769
	ChatGPT-4.0 [53]	✔	0.449	0.064	7.047	7.162	0.963	1.000	3.606	5.551

4 Experiments

In this section, we validated the efficacy of our expertise-centric prompting framework by generating three distinct types of pseudo-financial datasets. A comparative analysis using the proposed metrics was conducted to evaluate their quality against publicly available reference financial datasets.

4.1 Experiment Settings

We employed multilingual pre-trained LLMs, including KoGPT [52], ClovaX [44], LLAMA 2-Chat [45], GPT-3.0 [19], and ChatGPT-3.5/4.0 [20, 53] to empirically evaluate the efficacy of constructing financial tabular datasets using the proposed ECP framework. For embedding-level diversity, we utilized pre-trained ChatGPT-3.5 to extract embeddings for the rows of the data.

4.2 Datasets

To validate our framework, we generated three pseudo-financial datasets simulating typical banking activities: card transactions, loan statements, and deposits and savings. *i) Card transactions (D_{ct})*

Table 4: Evaluation results related to diversity on the deposits and savings dataset: Inter-instance and intra-attribute diversities of the datasets generated by our approach are evaluated.

Language	Deposits and Savings Dataset	ECP	Diversity of Instance		Diversity of Attribute↑					
			Embedding-Level		Categorical-Level		Binary-Level		Numerical-Level	
			Uniformity↓	PC↑	Avg($\mathcal{H}$)	Max($\mathcal{H}$)	Avg($\mathcal{H}$)	Max($\mathcal{H}$)	Avg($\mathcal{H}$)	Max($\mathcal{H}$)
English	GPT-3.0 [19]	✔	0.412	0.073	3.523	7.043	0.948	0.997	3.900	5.754
	ChatGPT-3.5 [20]	✔	0.412	0.079	3.472	7.080	0.955	0.981	3.468	4.455
	ChatGPT-4.0 [53]	✔	0.405	0.091	3.617	7.452	0.997	1.000	4.786	6.911
Korean	GPT-3.0 [19]	✔	0.434	0.026	2.422	3.635	0.912	0.942	2.606	3.754
	ChatGPT-3.5 [20]	✔	0.424	0.078	3.485	7.183	0.952	0.988	3.088	4.275
	ChatGPT-4.0 [53]	✔	0.419	0.070	3.519	7.064	0.985	1.000	4.433	6.418

encompass a comprehensive collection of financial records about credit and debit card transactions. *ii) Loan statements (D_{ls})* compile detailed information regarding loans and their associated financial transactions. Lastly, *iii) deposits and savings (D_{ds})* include data concerning deposits made into various financial accounts, encompassing savings and checking accounts. To encourage broader applicability, we have incorporated pseudonymization of user names into these datasets, facilitating their utilization in various financial scenarios [54, 55, 56] such as financial planning, credit risk assessment, and loan default prediction. The details of dataset properties and constraints designed with our framework are presented in Appendix C, while the dataset examples are provided in Appendix D.

4.3 Experimental Results

4.3.1 Analysis of Diversity

Diversity Analysis of the Card Transactions Dataset. Table 2 presents the diversity comparison between a public dataset (Credit Card Payment [12]) and the pseudo-financial datasets generated by our proposed framework. The results indicate that our framework constructs pseudo-financial datasets that closely resemble the reference dataset. Notably, models from the ChatGPT series outperformed other LLMs on average in generating realistic datasets. With ChatGPT-based LLMs, the average entropy $\mathcal{H}$ for attribute diversity is 4.10 for English and 4.15 for Korean. In contrast, other LLMs exhibit averages of 2.99 for English and 2.78 for Korean. Since the performance of LLMs correlates positively with their capability to generate diverse outputs based on a deeper understanding of the finance domain, our framework and evaluation metrics align well with these expectations.

Diversity Analysis of the Loan Statements Dataset. Table 3 demonstrates that the loan statements dataset generated through our approach is comparable in diversity to existing public datasets and exhibits strong generation performance with ChatGPT-based LLMs. On the contrary, KoGPT, ClovaX, and LLAMA 2-Chat failed to generate loan statement data.

Diversity Analysis of the Deposits and Savings Dataset. Table 4 illustrates the diversity of the generated dataset pertaining to deposits and savings. The diversity evaluation for the baseline was excluded due to the absence of publicly available datasets for this particular data type. Nevertheless, we present this table as a reference for future research on financial tabular data generation. Consistent with the previous two datasets, we confirmed that the ChatGPT-based models exhibit a higher capability in generating diverse datasets.

Consistency in Multilingual Dataset Generation. Tables 2, 3, and 4 present consistent results across both English and Korean datasets in terms of diversity inspection. This indicates the robust operation of our framework in generating diverse datasets for both languages.

Table 5: Evaluation results of constraints satisfaction: Unary and binary constraint satisfaction, denoted as ρ and τ respectively, are evaluated on the multilingual datasets generated by our approach.

Datasets	Lang.	Pre-trained LLMs	ρ	τ
Card Transactions	EN	KoGPT [52]	0.92	0.86
		ClovaX [44]	0.97	0.97
		LLAMA 2-Chat [45]	1.00	1.00
		GPT-3.0 [19]	1.00	1.00
		ChatGPT-3.5 [20]	1.00	0.99
		ChatGPT-4.0 [53]	1.00	1.00
	KR	KoGPT [52]	N/A	N/A
		ClovaX [44]	0.95	0.95
		LLAMA 2-Chat [45]	0.82	0.67
		GPT-3.0 [44]	0.97	0.96
		ChatGPT-3.5 [20]	0.97	0.94
		ChatGPT-4.0 [53]	1.00	1.00

Datasets	Lang.	Pre-trained LLMs	ρ	τ
Loan Statements	EN	GPT-3.0 [19]	1.00	0.68
		ChatGPT-3.5 [20]	1.00	0.77
		ChatGPT-4.0 [53]	1.00	0.83
	KR	GPT-3.0 [19]	1.00	0.68
		ChatGPT-3.5 [20]	1.00	0.70
		ChatGPT-4.0 [53]	1.00	0.77
Deposits and Savings	EN	GPT-3.0 [19]	1.00	0.92
		ChatGPT-3.5 [20]	1.00	0.90
		ChatGPT-4.0 [53]	1.00	0.97
	KR	GPT-3.0 [19]	1.00	0.75
		ChatGPT-3.5 [20]	1.00	0.69
		ChatGPT-4.0 [53]	1.00	0.73

Table 6: Ablation study on the influence of expertise-centric prompting (ECP) in our proposed framework is conducted on loan statements, and deposits and savings datasets, which encompass attributes necessary for combination of multiple binary constraints.

		Diversity of instance		Diversity of attribute			Constraint Satisfaction	
		Embedding-Level		Categorical-Level	Binary-Level	Numerical-Level		
Dataset	ECP	Uniformity↓	PC↑	Avg($\mathcal{H}$)	Avg($\mathcal{H}$)	Avg($\mathcal{H}$)	ρ	τ
Loan Statements Dataset		0.42	0.06	5.33	0.90	3.82	0.6	0.5
	✔	0.43	0.06	6.33	0.93	3.39	**1.0**	**0.7**
Deposits and Savings Dataset		0.39	0.06	2.99	1.48	4.48	0.1	0.4
	✔	0.41	0.08	3.47	0.95	3.46	**1.0**	**0.9**

Table 7: Ablation study on the key components of our ECP framework for loan statements (D_{ls}), and deposits and savings (D_{ds}) datasets, which encompass attributes necessary for combination of multiple binary constraints.

Method	$h(\hat{S})$	C^u	C^b	D_{ls}		D_{ds}		Avg(Std)
				ρ	τ	ρ	τ	
ChatGPT-3.5 [20]	✗	✔	✔	1.00	0.73	0.93	0.92	$0.90^{\ddagger}(\pm 0.12)$
	✔	✗	✔	1.00	0.76	1.00	0.88	$\underline{0.91}(\pm 0.15)$
	✔	✔	✗	1.00	0.74	1.00	0.73	$0.87(\pm 0.11)$
	✔	✗	✗	1.00	0.75	0.66	0.39	$0.70^{\dagger}(\pm 0.25)$
	✔	✔	✔	1.00	0.77	1.00	0.90	**0.92**(± 0.11)

4.3.2 Analysis of Constraint Satisfaction

Table 5 depicts the level of constraint satisfaction across three distinct datasets: card transactions, loan statements, and deposits and savings. ρ measures the degree of satisfaction for our unary constraints (Eq. 10), while τ quantifies the degree of satisfaction for our binary constraints between multiple columns (Eq. 11). A value close to 1.0 indicates strong adherence to these constraints.

Alignment with Unary Constraints. The datasets generated by models, particularly those from the GPT series, demonstrate strong alignment with unary constraints, indicating high levels of unary constraint satisfaction. Furthermore, the consistent ρ values for both English and Korean across most datasets represent that our proposed framework operates robustly across different languages.

Alignment with Binary Constraints. While the results of the binary constraint satisfaction were robust across languages, it is noteworthy that τ values were relatively lower than ρ values, suggesting a need for further development for understanding the complex operations and calculations of the current LLMs. In cases where complex constraints are imposed in specialized domains, such as the deposits and savings dataset, the level of compliance for the Korean data appears to be marginally lower as compared to that for English data. For such domains, additional language-specific model training is recommended.

4.3.3 Ablation Studies

Effectiveness of Our Expertise-Centric Prompting Framework in Generating Realistic Datatsets. Table 6 presents the comprehensive evaluation results on diversity and constraint satisfaction with ECP. We observed that while our approach resulted in increased in instance and attribute diversities, certain binary and numerical-level diversities showed slight decreases. According to Appendices C.2 and C.3, constraints were imposed on binary-level attributes and those pertaining to mathematical relationships. Consequently, the diversities related to the aforementioned attributes were reduced to satisfy the conditions, leading to a notable enhancement in constraint satisfaction through ECP. This indicates the effectiveness of our ECP framework in realistically adjusting datasets while preserving diversity.

Effectiveness of Our Key Components in the Expertise-Centric Prompting Framework. Table 7 shows the results of an ablation study to explore the effects of key components - schema calibration $h(\hat{S})$, unary constraints C^u, and binary constraints C^b - within our ECP framework. The results indicate that the highest average values of ρ and τ are achieved when all components are utilized across all combinations, highlighting the significant role of both calibrated schema and constraints in producing accurately calculated data aligned with specified operations. Notably, the average constraint satisfaction decreased by 23.91% (from **0.92** to $0.70^{\dagger}$) when both constraints were removed, whereas the average decreased by only 2.17% (from **0.92** to $0.90^{\ddagger}$) when schema calibration was removed. Particularly, in D_{ds}, the absence of our proposed binary constraints leads to a significant reduction in

understanding and satisfaction of complex operations. This suggests that leveraging binary constraints can significantly benefit the generation of more complex datasets, enabling the model to produce realistic datasets with a more sophisticated understanding.

5 Conclusion

This paper proposes a novel approach, the expertise-centric prompting framework, designed to generate realistic financial tabular datasets. We integrate pre-trained LLMs with financial experts and devise a structured prompting framework that improves the quality of the generated tabular dataset. Moreover, we suggest new evaluation metrics to verify the diversity and confidence of constraint satisfaction in datasets. We introduce three multilingual financial tabular datasets generated by our framework and demonstrate their diverse and accurate resemblance to real-world datasets. Furthermore, ablation studies demonstrate that both schema calibration and attribute constraints play significant roles in generating financial-specific tabular datasets.

6 Acknowledgments

This work was supported by KakaoBank Corp., and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multi-modal interactions). In particular, we would like to thank the designer Hyeji Jo (*johyeji.official@gmail.com*) for well-presented Figures 1, 2, and 4. We also like to thank the service marketers Jaeheung Yoo (*jayden.yoo@kakaobank.com*), Yeonghun Jang (*james.jang@kakaobank.com*) for planning and launching an innovative service called *Today's Mini Diary* using our methodology.

References

[1] David Byrd. Learning not to spoof. In Proceedings of the Third ACM International Conference on AI in Finance, ICAIF '22, page 139–147, New York, NY, USA, 2022. Association for Computing Machinery.

[2] Jingwei Ji, Renyuan Xu, and Ruihao Zhu. Risk-aware linear bandits with application in smart order routing. In Proceedings of the Third ACM International Conference on AI in Finance, ICAIF '22, page 334–342, New York, NY, USA, 2022. Association for Computing Machinery.

[3] Anubha Pandey, Alekhya Bhatraju, Shiv Markam, and Deepak Bhatt. Adversarial fraud generation for improved detection. In Proceedings of the Third ACM International Conference on AI in Finance, ICAIF '22, page 123–129, New York, NY, USA, 2022. Association for Computing Machinery.

[4] Shubhi Asthana and Ruchi Mahindru. Mapping of financial services datasets using human-in-the-loop. In Proceedings of the Third ACM International Conference on AI in Finance, ICAIF '22, page 183–191, New York, NY, USA, 2022. Association for Computing Machinery.

[5] Tina Piper. The personal information protection and electronic documents act-a lost opportunity to democratize canada's technological society. Dalhousie LJ, 23:253, 2000.

[6] Lissa L Broome and Jerry W Markham. Banking and insurance: before and after the gramm-leach-bliley act. J. Corp. L., 25:723, 1999.

[7] General Data Protection Regulation. General data protection regulation (gdpr). Intersoft Consulting, Accessed in October, 24(1), 2018.

[8] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In Proceedings of the First ACM International Conference on AI in Finance, ICAIF '20, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Edgar Alonso Lopez-Rojas and Stefan Axelsson. A review of computer simulation for fraud detection research in financial datasets. In 2016 Future Technologies Conference (FTC), pages 932–935, 2016.

[10] Hao-Yuan Chen, Shang-Xuan Zou, and Cheng-Lung Sung. Pluto: A deep learning based watchdog for anti money laundering. In Proceedings of the First Workshop on Financial Technology and Natural Language Processing, pages 93–95, Macao, China, August 2019.

[11] Hiroki Sakaji, Ryota Kuramoto, Hiroyasu Matsushima, Kiyoshi Izumi, Takashi Shimada, and Keita Sunakawa. Financial text data analytics framework for business confidence indices and inter-industry relations. In Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors, Proceedings of the First Workshop on Financial Technology and Natural Language Processing, pages 40–46, Macao, China, August 2019.

[12] Darpan Bajaj. Credit Card Exploratory Data Analysis. https://www.kaggle.com/datasets/darpan25bajaj/credit-card-exploratory-data-analysis?select=Customer+Acqusition.csv/, 2018.

[13] Mazaharul Hasnine Mirza. Loan data set, 2023.

[14] Anandaram Ganapathi. BWorld Robot Control Software. https://data.world/lpetrocelli/retail-banking-demo-data/workspace/file?filename=LuxuryLoanPortfolio.csv/, 2020.

[15] I-Cheng Yeh. default of credit card clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

[16] Sunil Jacob. Bank Loan modelling. https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling/, 2018.

[17] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[18] Lpetrocelli. Retail Banking Demo Data. https://data.world/lpetrocelli/retail-banking-demo-data/workspace/file?filename=completedorder.csv/, 2020.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, 2020.

[20] OpenAI. Chatgpt: A conversational ai language model, 2022. https://www.openai.com/research/chatgpt.

[21] Ruixue Liu, Shaozu Yuan, Aijun Dai, Lei Shen, Tiangang Zhu, Meng Chen, and Xiaodong He. Few-shot table understanding: A benchmark dataset and pre-training baseline. In Proceedings of the 29th International Conference on Computational Linguistics, pages 3741–3752, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[22] Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Context-tuning: Learning contextualized prompts for natural language generation. In Proceedings of the 29th International Conference on Computational Linguistics, pages 6340–6354, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[23] Ashish Upadhyay and Stewart Massie. Content type profiling of data-to-text generation datasets. In Proceedings of the 29th International Conference on Computational Linguistics, pages 5770–5782, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[24] Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), pages 208–224, Toronto, Canada, July 2023. Association for Computational Linguistics.

[25] Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoungpil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park, and Jung-Woo Ha. SQuARe: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6692–6712, Toronto, Canada, July 2023. Association for Computational Linguistics.

[26] Junyu Luo, Junxian Lin, Chi Lin, Cao Xiao, Xinning Gui, and Fenglong Ma. Benchmarking automated clinical language simplification: Dataset, algorithm, and evaluation. In Proceedings of the 29th International Conference on Computational Linguistics, pages 3550–3562, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[27] Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. Augmenting legal judgment prediction with contrastive case relations. In Proceedings of the 29th International Conference on Computational Linguistics, pages 2658–2667, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[28] Shweta Yadav and Cornelia Caragea. Towards summarizing healthcare questions in low-resource setting. In Proceedings of the 29th International Conference on Computational Linguistics, pages 2892–2905, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[29] Yuanfei Luo, Hao Zhou, Wei-Wei Tu, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Network on network for tabular data classification in real-world applications. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2317–2326, 2020.

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81, 2004.

[31] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.

[32] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. arXiv preprint arXiv:2004.04696, 2020.

[33] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, Proceedings of the Seventh Conference on Machine Translation (WMT), pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[34] Michał Pietruszka, Michał Turski, Łukasz Borchmann, Tomasz Dwojak, Gabriela Pałka, Karolina Szyndler, Dawid Jurkiewicz, and Łukasz Garncarek. Stable: Table generation framework for encoder-decoder models, 2022.

[35] Wenhu Chen. Large language models are few(1)-shot table reasoners. In Findings of the Association for Computational Linguistics: EACL 2023, pages 1120–1130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[36] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Evaluating and enhancing structural understanding capabilities of large language models on tables via input designs, 2023.

[37] Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, Tao Zhang, Chen Zhou, Kaizhe Shou, Miao Wang, Wufang Zhu, Guoshan Lu, Chao Ye, Yali Ye, Wentao Ye, Yiming Zhang, Xinglong Deng, Jie Xu, Haobo Wang, Gang Chen, and Junbo Zhao. Tablegpt: Towards unifying tables, nature language and commands into one gpt, 2023.

[38] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.

[39] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. FinLLM Symposium at IJCAI 2023, 2023.

[40] Haochen Li, Tong Mo, Hongcheng Fan, Jingkun Wang, Jiaxi Wang, Fuhao Zhang, and Weiping Li. Kipt: knowledge-injected prompt tuning for event detection. In Proceedings of the 29th International Conference on Computational Linguistics, pages 1943–1952, 2022.

[41] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing, 2023.

[42] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023.

[43] Yue Guo, Zian Xu, and Yi Yang. Is ChatGPT a financial expert? evaluating language models on financial natural language processing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 815–821, Singapore, December 2023. Association for Computational Linguistics.

[44] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3405–3424, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[46] Saqib Aziz, Michael Dowling, Helmi Hammami, and Anke Piepenbrink. Machine learning in finance: A topic modeling approach. European Financial Management, 28(3):744–770, 2022.

[47] Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. Ieee Access, 7:64323–64350, 2019.

[48] SUG Hyontai. Performance of machine learning algorithms and diversity in data. In MATEC Web of Conferences, volume 210, page 04019. EDP Sciences, 2018.

[49] Jaeyoung Choe, Keonwoong Noh, Nayeon Kim, Seyun Ahn, and Woohwan Jung. Exploring the impact of corpus diversity on financial pretrained language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2101–2112, Singapore, December 2023. Association for Computational Linguistics.

[50] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In International Conference on Machine Learning, pages 9929–9939. PMLR, 2020.

[51] I.T. Jolliffe. Principal Component Analysis. Springer Verlag, 1986.

[52] Ildoo Kim, Gunsoo Han, Jiyeon Ham, and Woonhyuk Baek. Kogpt: Kakaobrain korean(hangul) generative pre-trained transformer. `https://github.com/kakaobrain/kogpt`, 2021.

[53] OpenAI. Gpt-4 technical report, 2023.

[54] John Lee, Jow-Ran Chang, Lie-Jane Kao, and Cheng-Few Lee. Financial Analysis, Planning, and Forecasting, pages 433–455. Springer International Publishing, Cham, 2023.

[55] Jonathan N Crook, David B Edelman, and Lyn C Thomas. Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3):1447–1465, 2007.

[56] Lili Lai. Loan default prediction with machine learning techniques. In 2020 International Conference on Computer Communication and Network Security (CCNS), pages 5–9. IEEE, 2020.

A Limitations

Despite the strengths and novel contributions of the proposed expertise-centric prompting framework for generating financial tabular data, several limitations must be acknowledged. First of all, there still remains a deviation between our pseudo-financial data and the gold standard of actual personal data. With the digitalization of financial institutions, all customer financial activities conducted through both online and offline channels are meticulously logged and stored with various attributes, making it challenging to perfectly replicate actual personal data. Moreover, inexperienced users, especially those without financial expertise, should follow detailed instructions and undergo elaborate training before using our framework to ensure reliable results. This procedure potentially decreases efficiency and increases the time consumption of using the framework. Furthermore, the quality and characteristics of the generated pseudo financial data are influenced by the initial examples. If constraints are not satisfied or values in the examples are not sufficiently diverse, the final output quality may not be optimal. Lastly, similar to other fine-grained evaluation methods, our approach primarily focuses on certain facets of financial data like diversity and constraint satisfaction while potentially neglecting others, necessitating further research into integrating diverse evaluation metrics to achieve a comprehensive evaluation. By acknowledging these limitations, we aim to provide a balanced view of our framework's capabilities and areas for further improvement, highlighting the need for further research and development efforts to enhance the robustness and applicability of our methodology in diverse financial contexts.

B Implementation Details

Fig.3 shows the implementation details for assessing uniformity as presented in Eq.8.

```
def f_unif(e, t=2):
  pdst = torch.pdist(e, p=2).pow(2)
  return pdst.mul(-t).exp().mean().log()
```

Figure 3: Implementation of f_{unif} using PyTorch.

C Financial Activity Dataset Generation

We applied our expertise-centric prompting framework to the pre-trained LLM, ChatGPT-3.5 [20], to create a realistic financial tabular dataset $\mathcal{D}$ encompassing card, loan, and deposits and savings datasets. As a result, the presented pseudo-financial data exhibit significant value in addressing issues related to construction costs, anonymity, and personal data protection.

C.1 Card Transactions Dataset

The card transactions dataset, which was constructed using the proposed approach, incorporates nine columns with various types of values and up-to-date attributes that are not readily available in public datasets [12, 15]. To ensure data authenticity, human experts have designed the attributes within schema $\mathcal{S}_{\mathcal{D}}$ for our card transactions dataset D_{ct}, in which attributes encompass commonly encountered information such as transaction time, amount, and merchant-related information, as well as emerging attributes such as online payment.

Properties. The transaction time column includes both dates and timestamps, which facilitates detailed analyses of transaction records on an hourly basis. Our dataset also provides valuable information from the online financial ecosystem. In response to the growing significance of digital payments, we introduced columns related to online payment status and institution. Moreover, columns were added for regular payment status and type, indicating lifestyle patterns, such as streaming services and management fees.

Constraints. The proposed unary and binary constraints were applied to the columns associated with online and regular payments. Both online and regular payment statuses must be assigned values of either "Yes" or "No," representing unary constraints. Furthermore, the values of online payment institutions and regular payment types should only be filled when both online payment and regular payment statuses are set to "Yes," while they should be left vacant in cases where these statuses are "No" in accordance with binary constraints.

C.2 Loan Statements Dataset

Our loan data, which was meticulously generated through our proposed expertise-centric prompting and rigorous adherence to financial constraints, includes 14 essential loan-related attributes. To ensure realism, human experts have carefully designed the attributes within schema $\mathcal{S}_{\mathcal{D}}$ for the loan statements tabular dataset D_{ls}, as shown in Fig. 4. These attributes cover fundamental information such as amount, interest rate, and maturity, as well as detailed information trends, including loan product, new/renewal type, and delinquency.

Properties. We incorporated properties such as loan product type, new/renewal status, and bank visit/online inquiry type to provide a more comprehensive portrayal of the loan application context. In addition, delinquency count and duration can effectively quantify loan delinquency severity, making them valuable for loan default prediction and related analyses.

Constraints. In this dataset, we have applied our proposed unary constraints to binary-level attributes, such as application/execution type, loan product, and approval status. The total interest

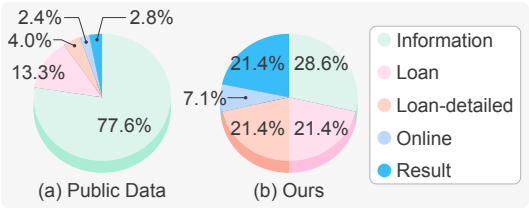


Figure 4: Comparison of attribute statistics on loan statements dataset: (a) for unbalanced attributes in public datasets and (b) for well-calibrated attributes in our generated dataset.

amount is determined based on binary constraints, and calculated as the product of loan amount, interest rate, and duration. Moreover, when the delinquency count is "> 0" or "$= 0$," the delinquency duration must also be "> 0" or "$= 0$," respectively, thus adhering to the binary constraints.

C.3 Deposits and Savings Dataset

The deposits and savings data, constructed using our novel prompting method and tailored by our proposed financial constraints is distinct from prior public datasets [13, 14, 16] in terms of real-world relevance. To achieve realistic datasets, human experts have designed the schema $\mathcal{S}_{\mathcal{D}}$ for the deposits and savings tabular datasets D_{ds}, incorporating fundamental properties such as amount, interest rate, and maturity, as well as updated financial indicators such as frequency and automatic transfer status.

Properties. The most significant attribute in the deposits and savings domain is the amount-related information, and our generated dataset D_{ds} provides this information down to the smallest unit. In contrast, Bank Loan [16] only denotes the presence or absence of a savings account, and German Credit Data [17] categorizes the amount information into "little," "moderate," "quite rich," and "rich," limiting the depth of analysis. Our dataset also encompasses practical details that are commonly used in real-world scenarios, such as expected maturity, interest, and total number of payments. Moreover, our dataset features digitalized financial trends, including the categorization of bank visits/online inquiry type, frequency, and automatic transfer status.

Constraints. We enforced our proposed binary constraints on both the maturity date and the expected maturity amount properties. Savings accounts have maturity durations ranging from 1 to 36 months, where a binary constraint ensures that the difference between the maturity date and the account opening dates precisely matches the specified duration. Furthermore, a mathematical relationship dictates that the expected maturity amount equals the product of the initial amount and the total number of payments.

D Pseudo-Financial Dataset Examples

We generated pseudo-financial data for the card transactions, loan statements, and deposits and savings domains in English and Korean languages. The tabular dataset examples are shown in Tables 8, 9, 10, 11, 12, and 13 for each domain.

Unary Constraint Examples The Card Transactions dataset includes the attributes Online Payment Status and Regular Payment Status as unary constraints, while the Loan Statements dataset includes Application/Execution Type, Loan Product, Approval Status, Inquiry Type, and New/Renewal Type as unary constraints. Additionally, the Deposits and Savings dataset includes Product Type and Inquiry Type as unary constraints. These attributes are binary, with options such as "Y/N," "Application/Execution," "Credit Loan/Secured Loan," "Approval/Rejection," "Bank Visit/Online Inquiry," "Renewal/New," and "Deposit/Saving."

Binary Constraint Examples Binary constants involve three types of operations: conditional operations, the four basic arithmetic operations, and self-combination. The Card Transactions dataset includes conditional operations, such as pairing Online Payment Status with Online Payment Institution and Regular Payment Status with Regular Payment Type. The Loan Statements dataset includes self-combination operations, such as Loan Amount × Interest Rate × Duration = Total Interest Amount and Loan Amount × Interest Rate / 12 = Monthly Interest Amount. Additionally, the Deposits and Savings datasets also include self-combination operations, such as Amount × Total Number of Payments = Maturity Expected Amount.

Table 8: Example of the card transactions dataset in English.

Customer ID	Payment Time	Payment Amount	Card Merchant Type	Card Merchant Location	Card Merchant Name	Online Payment Status	Online Payment Institution	Regular Payment Status	Regular Payment Type
A0001	2023-01-26 8:30	75	Laundry Service	Houston, TX	Washio	N	-	N	-
A0001	2023-05-12 14:30	56.78	Online Shopping	California, Los Angeles	Amazon	Y	PayPal	Y	OTT
A0001	2023-08-23 9:45	103.45	Gas Station	Texas, Houston	ExxonMobil	N	-	N	-
A0001	2023-11-05 11:55	250.75	Restaurant	Florida, Miami	Red Lobster	N	-	N	-
A0002	2023-03-17 16:00	79.99	Online Storage Service	Online Payment	Google Drive	Y	PayPal	Y	Cloud Storage Subscription
A0002	2023-06-23 16:40	150.25	Hotel	New Orleans, LA	Marriott	N	-	N	-
A0002	2023-07-19 18:20	250.75	Home Improvement Store	Atlanta, GA	Lowe's	Y	Google Pay	Y	Home Maintenance
A0002	2023-09-01 11:55	67.8	Grocery Store	Miami, FL	Publix	N	-	N	-

Table 9: Example of the card transactions dataset in Korean.

고객번호	결제시간	결제금액	가맹점종류	가맹점위치	가맹점이름	온라인결제여부	온라인결제기관	정기결제여부	정기결제속성
A0001	2023-02-14 21:00	35600	한식	서울시 강남구	한우랑 강남점	N	-	N	-
A0001	2023-04-01 16:40	9500	편의점	서울시 서대문구	미니스톱 연희점	N	-	N	-
A0001	2023-04-22 16:55	35000	음식배달업	서울시 서초구	배달의민족 서초점	Y	카카오페이	N	-
A0001	2023-12-24 22:30	80000	호텔/숙박	부산광역시 해운대구	신라스테이 해운대점	N	-	N	-
A0002	2023-02-17 20:45	35200	한식	경기도 수원시	한국관	N	-	N	-
A0002	2023-04-25 9:10	7800	커피/음료전문점	서울시 마포구	스타벅스 홍대입구점	N	-	N	-
A0002	2023-06-12 14:30	18500	양식	서울시 강남구	더그리핀	N	-	N	-
A0002	2023-12-24 14:20	52100	온라인쇼핑몰	온라인결제	쿠팡	Y	카카오페이	Y	식품배송

Table 10: Example of the loan statements dataset in English.

Customer ID	Date	App/Exec Type	Loan Product	Approval Status	Loan Amount	Interest Rate(%)	Duration(year)	Total Interest Amount	Monthly Interest Amount	Maturity Date	Inquiry Type	New/Renewal Type	Delinquency Count	Delinquency Duration
A0001	2023-10-17	Application	Credit Loan	Approval	30000	8	2	4800	200	2025-10-17	Bank Visit	Renewal	0	0
A0001	2023-10-18	Execution	Credit Loan	Approval	18000	7.25	2	2610	109	2025-10-18	Online Inquiry	Renewal	0	0
A0001	2023-10-19	Application	Credit Loan	Approval	22000	8.25	3	5445	151	2026-10-19	Bank Visit	New	0	0
A0001	2023-10-20	Application	Secured Loan	Approval	125000	7.75	7	68750	982	2030-10-20	Online Inquiry	New	0	0
A0002	2023-10-18	Application	Secured Loan	Rejection	0	0	5	0	0	2028-10-18	Online Inquiry	New	1	15
A0002	2023-10-25	Execution	Secured Loan	Approval	150000	7	5	52500	875	2028-10-25	Online Inquiry	New	0	0
A0002	2023-11-15	Execution	Secured Loan	Approval	75000	8.25	4	24750	516	2027-11-15	Online Inquiry	New	0	0
A0002	2023-12-10	Application	Credit Loan	Approval	5000	7	1	350	29	2024-12-10	Online Inquiry	New	0	0

Table 11: Example of the loan statements dataset in Korean.

고객번호	일자	신청/실행구분	대출 상품	승인 여부	대출 금액	이자(%)	기간(년)	전체이자	월납부이자	만기일자	조회구분	신규/연장 구분	연체건수	연체기간
A0001	2023-09-01	신청	신용대출	승인	50000000	5.49	1	2745000	228750	2024-09-01	온라인조회	신규	0	0
A0001	2023-10-05	실행	신용대출	승인	35000000	4.5	3	4725000	131250	2026-10-05	은행방문	신규	0	0
A0001	2023-11-10	신청	담보대출	거절	0	0	0	0	0	2023-11-10	은행방문	연장	0	0
A0001	2023-12-15	신청	신용대출	승인	20000000	3.5	2	1400000	58333	2025-12-15	은행방문	신규	0	0
A0002	2023-06-15	신청	신용대출	승인	5000000	7.5	2	750000	31250	2025-06-15	은행방문	신규	0	0
A0002	2023-07-20	실행	담보대출	승인	20000000	4	5	4000000	333333	2028-07-20	온라인조회	신규	0	0
A0002	2023-08-01	신청	신용대출	거절	0	0	0	0	0	2023-08-01	은행방문	신규	0	0
A0002	2023-09-05	신청	담보대출	승인	100000000	6.5	10	65000000	833333	2033-09-05	은행방문	신규	0	0

Table 12: Example of the deposits and savings dataset in English.

Customer ID	Date	Product Type	Amount	Interest Rate(%)	Duration(month)	Frequency	Total Number of Payments	Maturity Date	Expected Maturity Interest	Maturity Expected Amount	Inquiry Type	Automatic Transfer
A0001	2023-10-01	Deposit	20000	5.5	24	-	1	2025-10-01	11000	20000	Online Inquiry	-
A0001	2023-10-20	Deposit	10000	5.5	12	-	1	2024-10-20	5500	10000	Online Inquiry	-
A0001	2023-11-15	Savings	2000	4	12	Monthly	12	2024-11-15	160	24000	Online Inquiry	Y
A0001	2023-12-20	Savings	50	5	3	Daily	90	2024-03-20	2.29	150	Bank Visit	Y
A0002	2023-10-05	Deposit	5000	4.2	24	-	1	2025-10-05	5040	5000	Bank Visit	-
A0002	2023-11-12	Savings	100	3	3	Monthly	3	2024-02-12	0.75	300	Online Inquiry	Y
A0002	2023-12-01	Savings	500	3.5	6	Weekly	26	2024-05-01	17.5	13000	Bank Visit	Y
A0002	2023-12-15	Deposit	2000	5	12	-	1	2024-12-15	1000	2000	Online Inquiry	-

Table 13: Example of the deposits and savings dataset in Korean.

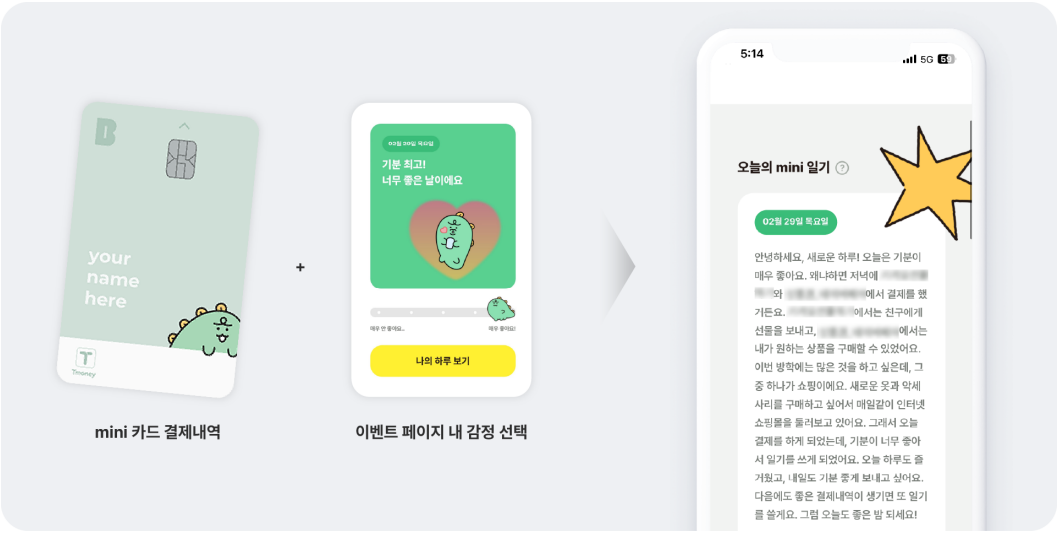
고객번호	일자	수신 상품	납입금액	이율(%)	기간(월)	납입주기	납입횟수	만기일자	만기예상이자	만기원금	조회구분	자동이체여부
A0001	2023-06-20	적금	10000	2.5	3	일	90	2023-09-20	1125	900000	온라인조회	Y
A0001	2023-10-05	예금	20000000	4.5	36	-	1	2026-10-05	9000000	20000000	은행방문	-
A0001	2023-12-01	적금	50000	3	12	주	52	2024-12-01	1800	600000	은행방문	Y
A0001	2023-12-31	예금	10000000	3.5	24	-	1	2025-12-31	875000	10000000	온라인조회	N
A0002	2023-05-15	예금	5000000	2	6	-	1	2023-11-15	50000	5000000	은행방문	-
A0002	2023-07-01	적금	10000	3	12	월	12	2024-07-01	3600	120000	온라인조회	Y
A0002	2023-09-10	적금	50000	3.5	24	주	104	2025-09-10	45500	5200000	은행방문	Y
A0002	2023-11-20	예금	20000000	4.5	24	-	1	2025-11-20	1800000	20000000	은행방문	-

E Case Study: Application of Pseudo-Financial Datasets in Generative AI Service

Figures 5 and 6 present a case study conducted by KakaoBank, an internet-only bank in South Korea, utilizing the methodology proposed in this paper. Using the ECP framework, we generated financial table datasets that were then used to develop and validate models for a generative AI service. The service, named *Today's Mini Diary*, automatically generates daily diary entries by taking a single day's debit card transaction table, as shown in Tables 8 and 9, along with the user's emotions for that day, as inputs to an LLM (Large Language Model).

'오늘의 mini 일기' 이벤트 안내

Figure 5: Event banner for the *Today's Mini Diary* service.



'오늘의 mini 일기' 일기 생성 예시

Figure 6: Example of the *Today's Mini Diary* service in action.