Bridging Data Silos in Finance via Federated Learning

Dongbin Chen, Shan Chang, *Member, IEEE*, Minghui Dai, Denghui Li, and Haoliang Zhao

***Abstract*—The financial industry faces the challenge of balancing data utilization and privacy protection. Federated learning (FL) offers a promising solution by enabling secure collaborative training. This paper focuses on an analysis of the key technologies for several financial applications that can benefit from FL. Specifically, we examine precision marketing based on multimodal FL (MMFL), anti-money laundering strategies leveraging federated graph learning (FGL), and credit card risk assessment utilizing vertical federated learning (VFL). Furthermore, we identify the key challenges in large-scale applications of FL in the financial industry. Additionally, we propose forward-thinking applications of FL in the finance sector, including the use of federated large language models (LLMs) for intelligent customer service (ICS) and decentralized FL integrated with blockchain for financial audit. Finally, we conduct a case study by using a consumer complaints dataset to verify the feasibility and effectiveness of federated LLMs in ICS.**

I. Introduction

Finance, as a typical data-intensive industry, widely employs artificial intelligence (AI) for modeling across diverse business practices. The significance of cross-domain data integration for analytical purposes is readily apparent. However, centralized model training poses a threat to data privacy. Despite the use of desensitization techniques, such as anonymization, critical data may still be compromised. The widespread concern regarding data privacy prompts several countries to enact regulations aimed at regulating the use of data to ensure compliance, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). In the absence of a joint modeling solution that ensures privacy protection, financial institutions are compelled to adopt more conservative modeling strategies, which ultimately hinder current business applications.

Federated learning (FL) enables collaborative model training while ensuring that raw data remains securely stored within its original domains, making it a promising solution to breaking down data silos. Beyond extensive research in academia, FL has also garnered attention across various industries, especially in the financial industry, which is characterized by high data value and strict privacy requirements. We observe that financial practitioners' limited understanding of FL principles, coupled with FL researchers' insufficient grasp of financial business scenarios, has hindered the large-scale integration of FL with financial practices. However, existing academic research often focuses on developing general FL frameworks rather than targeting specific financial applications. In contrast, white papers on FL from the industrial sector emphasize real-world business scenarios but often lack detailed analysis of the key technical processes underpinning relevant FL applications. Therefore, we hope that the study can help financial practitioners understand the current status of applications in the financial field and grasp the principles of FL technology through application cases, while assisting FL researchers in focusing on the practical issues and key technical challenges affecting its large-scale adoption in finance. Furthermore, we identify potential high-value business application directions for FL in the financial sector, providing path references for future research and application orientations.

In this paper, we emphasize several financial applications that can derive direct benefits from FL. These applications encompass the utilization of multimodal data for federated training to enhance precision marketing in the banking sector; the deployment of federated graph learning (FGL) for anti-money laundering purposes; and leveraging of vertical FL (VFL) to improve credit card risk assessment through cross-domain data integration. We briefly discuss potential solutions to the main challenges of these applications. Additionally, we identify the key challenges in the widespread deployment of FL in the financial industry. Subsequently, we highlight two innovative applications of FL in finance: the use of federated large language models (LLMs) for intelligent customer service (ICS) and the integration of decentralized FL with blockchain to meet financial auditing requirements. Finally, we present a case study based on a consumer complaints dataset, demonstrating the effectiveness of federated LLMs in ICS.

II. Practices of FL in Finance

Table I summarizes pilot applications of FL in finance, focusing on key areas of precision marketing, anti-money laundering, and credit risk control. In precision marketing, multimodal federated learning (MMFL) enables financial institutions to understand customer needs and enhance marketing effectiveness. In anti-money laundering, combining multi-institutional data via FGL can more effectively detect and prevent laundering activities, bolstering financial security. In credit risk control, VFL helps institutions improve the accuracy of credit decision-making and operational efficiency without sharing sensitive data.

A. MMFL Boosting Precision Marketing

Traditional structured data offers limited customer profiling, which is insufficient to capture risk appetites and investment

Dongbin Chen, Shan Chang, Minghui Dai, Denghui Li, and Haoliang Zhao are with the School of Computer Science and Technology, Donghua University, Shanghai, 201620, China. (Email: 1229821@mail.dhu.edu.cn; minghuidai@dhu.edu.cn; ldh@mail.dhu.edu.cn; zzzhl@mail.dhu.edu.cn) Corresponding author: Shan Chang (changshan@dhu.edu.cn)

TABLE I: Financial Application Cases on FL

Cases	Financial Industry	Financial Business	Participating Institutions
Credit Card Anti-Fraud	Banking	Anti-Fraud	Bank, Internet Company
Inclusive Financial Services for Small and Micro Merchants	Banking		Banking Alliance, Bank
Cross-Institution Joint Anti-Money Laundering	Banking	Anti-Money Laundering	Bank, Operator
Anti-Money Laundering Base on FGL	Banking		Bank, ISP, Social App
Insurance Existing Customer Marketing	Insurance	Precision Marketing	Insurance Company, Operator
Batch Customer Screening and Traffic Development	Banking		Bank, Government Department
Health Insurance Potential Customer Mining	Insurance		Bank, Insurance Company
Credit Card Cross-Marketing	Banking		Bank, Internet Consumer Platform
Precision Marketing and Customer Acquisition	Asset Management		Mutual Fund Company, Financial Portal, Short-Video App
Personal Consumer Credit Loan Risk Control	Banking	Credit Risk Control	Bank, Consumer Finance Company
Intelligent Risk Control for Banks	Banking		Bank, Financial Comprehensive Service Platform
Post-loan Enterprise Risk Control Monitoring	Banking		Bank, Group Company
Pre-loan Credit Approval for SMEs	Banking		Internet Bank, Banking Alliance
Enterprise Credit Risk Control	Banking		Bank, Operator
Bond Default Prediction	Asset Management		Mutual Fund Company, Bank

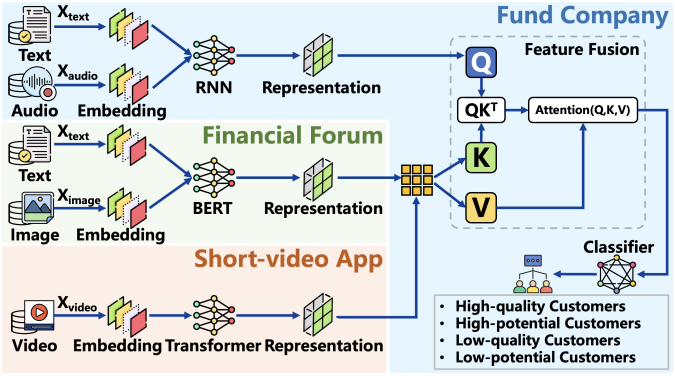

Fig. 1: The fund company, a financial forum, and a short-video app collaborate in MMFL. Following embedding processing, their multimodal data are separately processed by RNN (for the fund company's data), $BERT$ (for the forum's data), and Transformer (for the app's data) to extract features. The fund company then fuses these features using a cross-attention mechanism. The fused features are fed into a classifier to classify clients into four categories, enabling fund companies to accurately understand clients by integrating multi-source information.

needs. MMFL integrates multiple modalities of data for modeling, leading to the construction of comprehensive customer profiles and boosting precision marketing.

Without loss of generality, in a feature-level fusion federation consisting of K clients, the forward propagation of the k^{th} client (f^k) involves the sum of modality weights applied to its local model with parameters ω^k. The objective function is to minimize the loss function $\min_{\omega} \mathcal{L}\left(H^s\left(\sum_{k=1}^{K} \phi^k f^k, \omega^l\right), Y^l\right)$, where H^s denotes the server model, ϕ^k represents the weight of the k^{th} client, and l denotes the task party (label-holding client).

In the case illustrated in Fig. 1, a fund company collaborates with a financial forum and a short-video app to train a customer classification model, targeting existing clients for specific fund products. The intuition behind the design stems from the correlation between asset status, risk appetites, forum viewpoints, and short video preferences. Data include text and audio data from the fund company (text and voice conversations between fund clients and customer service), text and image data from the forum (articles of interest to forum users and their accompanying images), and video data from the app (videos favorited by users). During local training, clients embed private data and use matched models to train on it (e.g., RNN for text and audio, $BERT$ for text and image). Extracted lower-dimensional features are sent to the central server (i.e., the fund company), which then performs feature fusion to extract more representative and comprehensive features, utilizing a cross-attention mechanism in this case. The detailed design of the cross-attention mechanism is as follows: the representation vector of the fund company is linearly transformed via a learnable weight matrix to obtain the query (Q). The concatenated representation vectors of the financial forum and short-video app are linearly transformed via another learnable weight matrix to obtain the key (K) and value (V). By computing the product of Q and the transpose of K, scaling the result, and normalizing the scores using softmax, the weighted fused feature vector is obtained. This vector is subsequently connected to a classifier.

Based on the utilization of the existing attention mechanism, CADMR [1] involves a two-phase process: pretraining and fine-tuning. In the pretraining phase, an autoencoder and modality-specific feature extractors are trained separately to learn disentangled representations for textual and visual features. During the fine-tuning phase, a cross-attention mechanism integrates the user-item rating matrix with the unified multimodal representation, refining the matrix. The refined matrix is then processed through the trained autoencoder to produce the final reconstructed rating matrix.

MMFL integrates the characteristics of FL and multimodal learning (ML), while the introduction of multimodal data poses several challenges.

- **Modal Heterogeneity.** Existing research primarily addresses input-level heterogeneity, but comprehensive modality heterogeneity (encompassing diverse semantic structures) remains understudied. A promising direction is to enable deeper modality fusion by transmitting abstract representations or sharing encoder parameters, though the optimal implementation remains an open challenge.
- **Knowledge Transfer.** Cross-modal knowledge transfer represents a new challenge in MMFL. Although existing studies introduce unimodal knowledge into MMFL via distillation, there remains a need to explore additional effective approaches for transferring knowledge from one modality to another.
- **Benchmarks for MMFL.** MMFL still lacks practically validated benchmarks that have been verified in terms of accuracy, fairness, security, and generalization capability.

B. FGL Facilitating Anti-money Laundering

The estimated annual global money laundering volume ranges from \$800 billion to \$2 trillion, equivalent to 2%–5% of global GDP, with an annual growth of approximately \$100 billion[1].

Money laundering activities (MLAs) can be modeled as graph structures, where nodes represent bank accounts and edges represent transfer transactions in institutional subgraphs. Despite the effectiveness of graph convolutional networks (GCNs) in graph modeling, they face data silos when identifying suspect MLAs. Each bank holds only local transaction and feature data, while perpetrators use sophisticated strategies across regions and institutions to obfuscate illegal funds, limiting model generalization. FGL integrates multi-party graph data to detect global-level MLA patterns (e.g., circular and centralized laundering), with node-level tasks identifying suspicious accounts and edge-level tasks predicting illicit transactions for real-time prevention.

In horizontal FGL (HFGL), each client sub-graph is a part of the global graph, where nodes and edges share identical feature spaces, but node IDs have minimal overlap. Consider a HFGL involving three banks, as depicted in Fig. 2. The features of nodes include type, geographical consistency, age, offshore status, high-risk country origin, transaction volume spikes, multi-currency involvement, frequent large foreign currency withdrawals, inflow-outflow patterns, and balance reductions. Edge features encompass statistical features such as special amounts, transaction counts, time intervals, and counterparty types. Following the architectural of GCNs and $FedAvg$, the objective function of the global GCN model performing a node-level task can be denoted as $\min_{\omega} \frac{N^k}{N} \sum_{k=1}^{K} \mathcal{L}\left(H\left(X^k, \mathbb{A}^k, \omega\right), Y^k\right)$, where $\mathbb{A}^k$ represents adjacency matrix in the graph of the k^{th} client, H is a GCN model, $N^k = |D^k|$ and $N = \sum_{k=1}^{K} N^k$. The following is the typical training process of HFGL. An embedding layer is first used to convert node vectors into their embeddings, which are applied to a GCN to obtain interactions between nodes on the local subgraph. Then, a classifier that predicts whether a node is a money laundering account is directly connected to the final hidden layer of the GCN. Predictions are compared with local ground truth labels to compute the loss and derive local gradients, which are then uploaded to the server for aggregation, after which they are redistributed to clients for model updates.

[1]https://www.unodc.org/unodc/en/money-laundering/overview.html

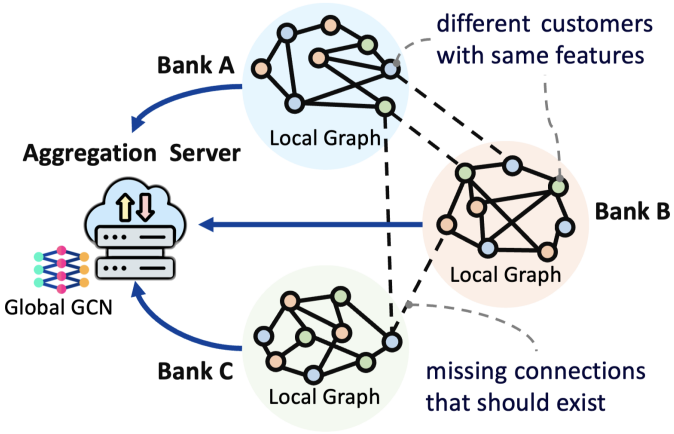

Fig. 2: Each bank has a local graph where nodes denote customers and edges represent fund transfers between them. MLAs exhibit specific patterns in node-node interactions and node-edge interactions. Leveraging HFGL, the global GCN captures more comprehensive MLA patterns. By incorporating cross-bank transactions, the model is able to capture more complex patterns, thereby enhancing the accuracy of MLA identification.

The subgraphs are vertically distributed in vertical FGL (VFGL), where clients share the same node ID space, but their feature spaces and labels differ. The aim of VFGL is to learn a GCN model by integrating $\left\{X_v^k \mid v \in V\right\}$ while utilizing $\left\{Y_v^k \mid v \in V\right\}$, where V denotes the set of shared nodes. The objective function is $\min_{\omega} \mathcal{L}\left(H^s\left(combine\left(\left\{h\left(X_V^k, \mathbb{A}_V^k, \omega^k\right)\right\}_{k=1}^{K}\right), \omega^l\right), Y_V^l\right)$, where $combine$ denotes the global node embedding and h is the local node embedding. For instance, a bank, an internet service provider (ISP) and a social app, follow the VFGL for collaboration, as illustrated in Fig. 3. The intuition behind this design is that criminals are likely to use ISP network services and communicate or interact via social applications before and after conducting money laundering transactions. The model training involves sequentially generating initial, local, and global node embeddings. Firstly, each client collaborates with others using the features of shared nodes to generate initial node embeddings protected by secret sharing or homomorphic encryption. Local embeddings are generated via multi-hop neighborhood aggregation and sent to a central server for integration into global embeddings. These embeddings are input into a classifier by the client with labels to predict money laundering accounts. The global model weights consist of four parts: the weights for initial node embeddings, the weights for neighborhood aggregation on subgraphs (held solely by clients), the weights for the hidden layers of the classifier (held by the central server),

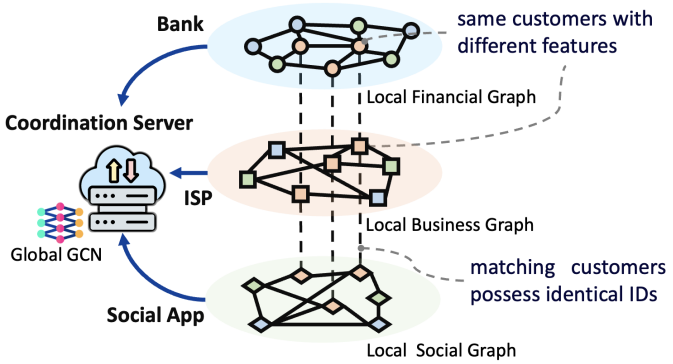

Fig. 3: Cross-subgraph nodes with the same color but different shapes represent shared customers (with distinct features) across participants. Edges in the local financial graph denote fund transfers between bank customers; those in the local business graph represent network service interactions between ISP customers; and those in the local social graph indicate social relationships among social app users. The three parties generate feature embeddings for shared customers using suspicious account labels and learn the global model's weights.

and the weights for the output layer of the classifier (held by the client with labels, which in this case is the bank).

When integrating GCN with FL, there are unique challenges beyond general FL:

- **Cross-client Missing Information.** In FL, each client holds a subgraph of the global graph, and some nodes may have neighbors owned by other clients. Due to privacy concerns, a node cannot access features from other clients, thereby leading to incomplete node representation. Yan *et al.* address this problem by proposing a two-stage perturbation mechanism [2] for user-item interaction publishing. In the first stage, the user's high-order patterns from shared HINs are perturbed using an exponential mechanism to balance utility and privacy. In the second stage, user-item interactions within selected shared HINs are perturbed in a degree-preserving manner, ensuring semantic consistency and enhancing interaction diversity while preserving privacy.
- **Secure Aggregation for Embeddings.** Current research mainly employs local differential privacy (LDP) and pseudo-instance sampling to mitigate privacy leakage during the embedding process [3], which leads to an impact on performance. Therefore, designing an aggregation method that balances efficiency and security remains an open problem.

C. VFL Enhancing Credit Risk Control

Banks often struggle to assess borrowers' credit status due to limited customer insights, leading to challenges in pre-loan approval and post-loan risk monitoring for small and micro enterprises (SMEs). VFL addresses this by expanding data dimensions through collaboration with entities such as banking alliances and government agencies. For instance, the bank integrate customer data (including economic income, overdue records, and credit ratings) with alliance data on transactions, payments, and bank card risks to build comprehensive user profiles, enabling more precise risk assessments.

In the standard VFL framework, sample ID alignment relies on the private set intersection (PSI) protocol. However, in practice, alignment scope is constrained by divergent customer groups/regions across institutions and their caution in data sharing, often resulting in a small intersection of training samples. To address this, one strategy is the semi-alignment of ID-mismatched samples using feature correlation coefficients. The rationale is that for samples with the same label, embeddings of unaligned features exhibit high similarity. This approach involves collaboration between a task party (label-holding client) and auxiliary parties (label-free clients). First, the task party trains an encoder and a transformer model simultaneously. The transformer learns to map task embeddings to auxiliary embeddings, improving model performance and feature correlation. To prevent overfitting, when processing unaligned task embeddings, the transformer outputs proxy embeddings instead of direct auxiliary embeddings. These proxies are then matched to unaligned auxiliary samples by similarity. The encoder is trained under a supervised contrastive learning framework to enhance discriminability between same/different label embeddings, even with scarce aligned samples. Semi-supervised learning predicts labels for all unaligned auxiliary embeddings, reducing mismatches from inconsistent labeling. Specifically, each proxy embedding is compared to same-label auxiliary embeddings, with the closest match selected as the best counterpart. Finally, corresponding task and auxiliary samples are combined into semi-aligned samples.

Some studies address the challenge of limited aligned samples. FedSim [4] introduces fuzzy identifiers into training, dynamically adjusting soft linking between parties based on record similarity. FedHSSL [5] utilizes all unlabeled samples (both aligned and unaligned) to pretrain local models. It employs cross-party-guided self-supervised learning (SSL), where each party uses a cross-party encoder to regularize its SSL training, improving local encoder discriminability and inter-party representation alignment. Pretrained models are fine-tuned on aligned labeled data in a conventional VFL setup to boost joint model performance.

III. Challenges of FL in Finance

We discuss the key challenges of particular concern to the financial sector that impede its large-scale commercialization.

A. Intersection Member Exposure Issue in VFL

In VFL, PSI is commonly used for secure sample alignment, enabling participants to compute the intersection of their datasets without disclosing raw data. However, in privacy-sensitive financial contexts, even exposing intersection sample IDs to other parties is unacceptable. For instance, in banks (where customer account information is highly confidential) revealing such IDs without consent may lead to legal consequences. Concerns over the exposure of intersection sample IDs thus remain a critical barrier to VFL collaboration among financial institutions. This raises the question: *Can sample*

alignment be achieved without disclosing intersection sample IDs, or can VFL operate without sample alignment?

Buddhavarapu *et al.* propose two PSI variants [6]: private-ID and private secret-shared set intersection (PS3I). The private-ID protocol allows clients to generate pseudorandom universal identifiers (UIDs) for their datasets, enabling downstream computations without disclosing which UIDs belong to the intersection. In contrast, PS3I privately computes additive shares of intersection records instead of exposing their plaintext. However, both protocols have significant limitations. Sun *et al.* introduce private set union (PSU) [7], generating synthetic data to compensate for unaligned samples in the union set. Yet the distribution mismatch between synthetic and real data often causes substantial performance degradation in federated models, and this approach is limited to imbalanced binary classification tasks.

B. Contribution Assessment

Well-designed contribution and incentive mechanisms are essential to balance interests among participants, facilitate cross-institutional cooperation, and meet compliance requirements. Consider a FL fraud detection project involving multiple banks, where data differ in type, volume, and quality. A bank that contributes more effective fraud-detection data should derive greater benefits, such as reduced model usage costs. However, practical challenges (including complex FL communication protocols and model architectures, as well as variations in data distribution and quality) hinder the precise tracking and quantification of participants' contributions across stages. For instance, in a joint credit card default prediction project, participants' data exhibit divergent distributions in key features (e.g., customer income levels, consumption frequency) and inconsistent quality, rendering traditional contribution assessment methods (e.g., data volume proportion) inadequate.

C. Interpretability

The financial industry is subject to strict regulation, with regulatory authorities requiring financial institutions to explain their decision-making processes to ensure fairness, impartiality, and transparency. The complexity and distributed nature of FL models make it difficult for financial institutions to meet such interpretability requirements, thereby impacting the widespread adoption of FL.

FL integrates data from multiple participants for training, which originates from diverse sources with heterogeneous features. Models must synthesize multi-source data to derive decisions. Explaining which specific data features and how their interactions drive final marketing decisions remains challenging. Moreover, when FL employs black-box models, their internal logic and decision pathways are often opaque. In critical financial scenarios such as risk assessment and investment strategy, institutions require clear insights into model decision-making. For instance, an FL-based loan default prediction model using deep neural network to process massive financial transactions and customer profiles poses challenges in explaining why a customer is classified as high-risk.

By quantifying the importance of multi-party distribution characteristics to the VFL model through federated counterfactual loss, EVFL [8] integrates prior knowledge and expert experience to generate interpretable counterfactual instances.

Shapley value (SV)-based contribution assessment is a key research focus [9]. DIG-FL [10] assesses contributions by measuring global model loss changes when adding a participant's data, a process that requires enumerating all participant combinations and incurs substantial retraining costs, thus necessitating methodological improvements.

D. Systematic Standards

The existing international standard IEEE 3652.1-2020[2] provides an architectural framework and application guidelines for FL. However, to ensure its effective implementation in the financial industry, the standards require enhancements in the following key aspects.

- **Model interpretability standards.** The financial industry has stringent requirements for model interpretability, as financial institutions must explain model decision-making processes to regulators and clients. For instance, in a credit risk assessment model using FL, institutions need to clearly explain how risks are evaluated using factors like customer income and credit history, as well as how the joint model training process operates.
- **Security and performance evaluation standards.** Quantitative evaluation metrics and hierarchical classification frameworks for FL platforms' security and performance should be established to assist financial institutions in selecting secure and efficient FL solutions.
- **Industry-specific application standards.** Detailed FL application guidelines and specifications tailored to the financial industry's requirements and characteristics must be developed, covering industry-specific data processing workflows, model architecture selection, and business scenario adaptation.

E. Platform Compatibility

Current FL platforms differ in communication protocols, data processing mechanisms, model architectures, algorithms, and system resource management frameworks. These disparities necessitate the development of additional adaptive layers for cross-platform collaboration, thereby reducing system efficiency and escalating security risks. Moreover, differences in security and privacy protection mechanisms across platforms further complicate interoperability among FL systems.

IV. Opportunities for FL in Finance

As FL applications in the financial industry evolve, some promising research directions remain underdeveloped. This section discusses two potential future pathways.

[2]https://standards.ieee.org/ieee/3652.1/7453/

A. Federated LLMs for Intelligent Customer Service

The financial industry requires customized ICS models, yet small and medium-sized institutions often lack sufficient data and computational resources for high-quality LLM training. Federated LLM fine-tuning and prompt engineering offer solutions to this challenge.

1) Federated LLMs Fine-Tuning: Traditional LLM fine-tuning by individual institutions may lead to overfitting, and small institutions often experience poor ICS performance due to limited local data, resulting in decreased answer accuracy and longer response times. Federated LLM fine-tuning, which leverages data from multiple institutions, enhances model generalization. However, parameter-efficient tuning (PETuning) for pre-trained language models (PLMs) poses the most critical challenge. FedPETuning [11] conducts a comprehensive empirical study of representative fine-tuning methods for PLMs in FL, addressing privacy attacks, performance comparisons, and resource-constrained scenarios.

2) Federated LLMs Prompt Engineering: Prompt engineering, which involves designing templates to guide model behavior without altering LLMs' parameters, is an effective strategy to minimize the number of updated parameters. The core of this approach lies in creating prompts that enable models to accurately understand and execute tasks, a process requiring meticulous input adjustments and iterative optimization. Federated LLM prompt engineering frameworks can enhance this process's efficacy. FedPrompt [12] explores prompt tuning using model splitting and aggregation in FL, demonstrating that this method reduces communication costs to just 0.01% of PLM parameters while maintaining nearly identical accuracy under both IID and non-IID data distributions. In practice, banks can collaborate with product issuers by refining ICS prompts to enhance marketing effectiveness and avoid missed opportunities caused by generic or simplistic responses.

B. Swarm Learning for Financial Audit

Financial auditing imposes strict regulatory requirements on the model training and aggregation processes of FL. Swarm Learning [13], a blockchain-based decentralized FL framework, meets these strict audit requirements through its traceability and immutability, enabled by its distributed ledger mechanism.

1) Blockchain-enabled Verifiable Aggregation: In FL's aggregation processes, client updates are recorded as transactions on the blockchain, each containing crucial information such as client identity, parameter data, and timestamps. Smart contracts embedded in the blockchain automatically validate submitted model parameters against preset verification rules. During this validation, the consensus mechanism of the blockchain [14] ensures all clients agree on the validation results. Model parameters verified by smart contracts and confirmed via consensus are stored on the blockchain and broadcast to the network. Due to the blockchain's immutability and traceability, auditors can access it at any time to review detailed submission records, validation processes, and aggregation results of model parameters. By tracing blockchain data, auditors can deeply analyze FL's aggregation process, verify its rationality and effectiveness, and ensure compliance with expected standards.

2) TEE-based Training Integrity Verification: Trusted execution environment (TEE) ensures the integrity of the training process and mitigates free-riding behavior among clients. TEE verifies training integrity through a "commit and prove" paradigm [15]: after each training round, the rich execution environment (REE) stores model parameters and sends a submission message to the TEE upon completing training. When the TEE requests to check a specific round, the REE returns input/output parameters and the corresponding training data. The TEE then verifies parameter consistency with the submission message, retrains using the data and input parameters, and compares its output with externally transmitted parameters. Successful verification of randomly selected rounds indicates clients have executed training tasks honestly. Finally, the TEE checks the last round's parameters, computes model updates, and generates proofs for server validation. The TEE framework ensures FL's training process is auditable, thus meeting the financial industry's stringent audit requirements.

V. Case Study

A. Methodology

We conducted an experiment to simulate a scenario where 10 banks participate in fine-tuning federated LLM using their local customer complaint text data, aiming to enable the ICS to achieve more accurate automatic classification of customer complaints. More precise classification of customer complaints allows banks to respond to customer needs more efficiently and optimize the allocation of human customer service resources, thereby enabling targeted enhancement of service quality and reducing manual processing costs. The experiment utilized the "Consumer Complaints Dataset for NLP"[3] from Kaggle, which contains 162,421 consumer submissions with narratives from the U.S. Consumer Financial Protection Bureau (CFPB). Each complaint submission is classified into one of five financial product categories: credit reporting, debt collection, mortgages and loans, credit cards, and retail banking. The dataset exhibits class imbalance, with 56% of samples belonging to the credit reporting class and the remainder roughly evenly distributed (8–14%) among the other classes. The dataset is split into training and test sets at a 0.8: 0.2 ratio. To simulate IID and non-IID scenarios, each client partitioned sample data using two strategies: random splitting (for IID) and Dirichlet distribution (α=1, for non-IID), with results compared against a single-client model.

We employ $BERT$[4] as the pre-trained model. During initialization, the model loads the pre-trained $BERT$ and adds a fully connected layer to map the 768-dimensional features output by $BERT$ to the specified number of classification categories. In the forward propagation process, input data is first encoded by $BERT$ to extract the pooled output which is then fed through a linear layer to generate the final classification results. The number of clients is set to 10, with all clients participating in each training iteration. The local model is trained for 1 epoch, with the learning rate set to

[3]https://www.kaggle.com/datasets/shashwatwork/consume-complaints-dataset-fo-nlp/data

[4]https://huggingface.co/google-bert/bert-base-uncased

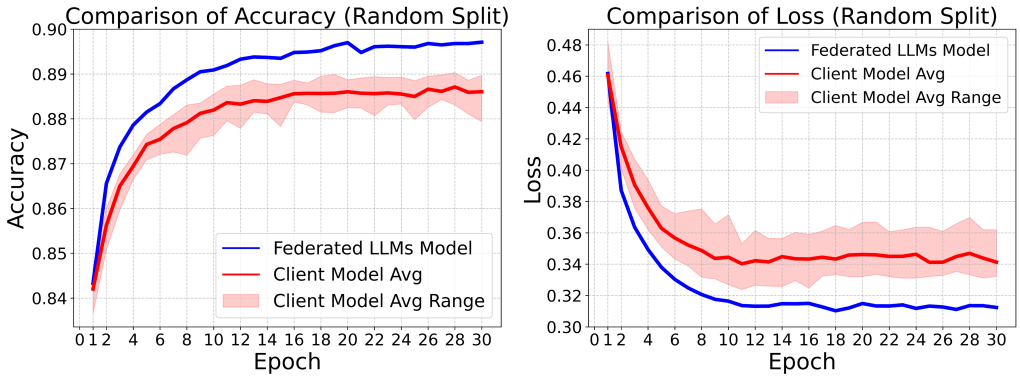

Fig. 4: Federated LLM under IID Setting.

0.00001 to balance training stability and convergence speed. The batch size is set to 16 for both training and testing phases. The cross-entropy loss is adopted as the loss function for the classification task, and the $Adam$ algorithm is selected as the optimizer.

The experiment runs on a server equipped with an Intel Xeon processor and NVIDIA GeForce RTX 3090 graphics cards, using the Ubuntu 20.04 LTS operating system and Python 3.10.13 for development. Model construction and training are mainly implemented using PyTorch. By analyzing the changes in accuracy and loss across epochs, we examine the performance differences between the federated model and the standalone client model.

B. Experimental Results

We present the results of fine-tuning federated LLM under both IID (random split) and non-IID (Dirichlet α=1) data partitions, evaluating performance via accuracy and loss metrics.

- **Performance Under IID Setting (Random Split).** As shown in Fig. 4, the global model achieves consistent improvement across epochs, with accuracy increasing from 0.843 to 0.897 and loss decreasing from 0.462 to 0.312. Notably, the global accuracy consistently surpasses the average accuracy of local client models (0.842–0.886), demonstrating effective knowledge aggregation across clients. The narrow range between the minimum and maximum local accuracy (e.g., 0.837–0.847 at epoch 1) and low loss variance (average loss of client models: 0.461–0.343) indicate stable client contributions under IID, where data homogeneity facilitates federated convergence. This stability is critical for banking scenarios requiring reliable cross-client generalization, as it ensures uniform complaint classification quality across institutions.
- **Robustness Under Non-IID Data (Dirichlet α=1).** Under non-IID conditions (as illustrated in Fig. 5), the global model converges rapidly in the initial stage but shows slow performance improvement in later stages, with accuracy increasing from 0.790 to 0.896 and loss decreasing from 0.599 to 0.335. The gap between the global model accuracy and the average accuracy of client models (0.778–0.866) highlights the federated framework's ability to mitigate data heterogeneity: despite skewed client distributions, the global model leverages aggregated parameters to overcome individual client biases. Notably, the minimum (e.g., 0.723 at epoch 1) and maximum (0.841 at epoch 1) accuracy of client models exhibit significant early-stage dispersion, reflecting non-IID-induced client disparities, yet convergence narrows this gap over time. Loss dynamics further confirm this trend: the average loss of client models (0.706–0.443) remains higher than the global loss, indicating that centralized aggregation reduces overfitting to client-specific data distributions. This is a key advantage for real-world banking datasets with inherent class imbalance (e.g., 56% credit reporting complaints).
- **Comparative Analysis and Practical Implications.** The federated model outperforms standalone client models in both scenarios, with global accuracy exceeding the maximum local accuracy in the non-IID scenario by epoch 3 (0.872 vs. 0.864) and maintaining lower loss across all epochs. This demonstrates that FL effectively pools diverse banking data while preserving privacy, enabling accurate cross-institutional complaint classification. The narrowing variance in local metrics across epochs signifies improved alignment among client models, a critical factor for collaborative service optimization. While non-IID conditions introduce initial training challenges, the framework's resilience underscores its suitability for real-world financial datasets, where data distributions naturally diverge across institutions.

In summary, the results validate that federated LLM enhance complaint classification accuracy and robustness, providing a scalable solution for inter-bank collaboration without compromising data privacy.

VI. Conclusion

FL offers a feasible technique for addressing the challenge of balancing data utilization and privacy protection in the financial industry. This study analyzes typical applications of FL in finance through in-depth case studies, helping financial practitioners understand the technical principles and the logic behind practical implementations. Furthermore, we address

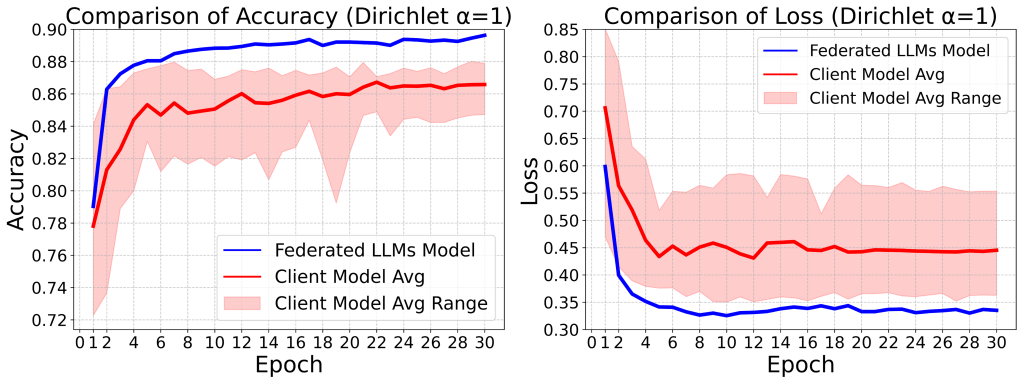

Fig. 5: Federated LLM under Non-IID Setting.

the key challenges hindering the large-scale adoption of FL in finance, offering targeted research agendas for scholars in this field. We identify high-value potential applications of FL in financial business, providing practical pathways for future research and implementation. A case study demonstrates the effectiveness of federated LLM in ICS, thereby providing empirical evidence for financial institutions to adopt FL in real-world scenarios.

VII. Acknowledgment

This research is supported in part by the National Natural Science Foundation of China (Grant No. 62472083), and the AI-Enhanced Research Program of Shanghai Municipal Education Commission (Grant No. SMEC-AI-DHUZ-01).

References

[1] Y. Khalafaoui, M. Lovisetto, B. Matei, and N. Grozavu, "CADMR: Cross-Attention and Disentangled Learning for Multimodal Recommender Systems," *arXiv preprint arXiv:2412.02295*, 2024.

[2] B. Yan, Y. Cao, H. Wang, W. Yang, J. Du, and C. Shi, "Federated Heterogeneous Graph Neural Network for Privacy-preserving Recommendation," in *Proceedings of the ACM Web Conference 2024*, WWW '24, (New York, NY, USA), p. 3919–3929, Association for Computing Machinery, 2024.

[3] Z. Liu, L. Yang, Z. Fan, H. Peng, and P. S. Yu, "Federated Social Recommendation with Graph Neural Network," *ACM Trans. Intell. Syst. Technol.*, vol. 13, Aug. 2022.

[4] Z. Wu, Q. Li, and B. He, "A Coupled Design of Exploiting Record Similarity for Practical Vertical Federated Learning," in *Advances in Neural Information Processing Systems* (S. Koyejo and S. Mohamed and A. Agarwal and D. Belgrave and K. Cho and A. Oh, ed.), vol. 35, pp. 21087–21100, Curran Associates, Inc., 2022.

[5] Y. He, Y. Kang, X. Zhao, J. Luo, L. Fan, Y. Han, and Q. Yang, "A Hybrid Self-Supervised Learning Framework for Vertical Federated Learning," *IEEE Transactions on Big Data*, pp. 1–13, 2024.

[6] P. Buddhavarapu, A. Knox, P. Mohassel, S. Sengupta, E. Taubeneck, and V. Vlaskin, "Private Matching for Compute." Cryptology ePrint Archive, Paper 2020/599, 2020.

[7] J. Sun, X. Yang, Y. Yao, A. Zhang, W. Gao, J. Xie, and C. Wang, "Vertical federated learning without revealing intersection membership," *arXiv preprint arXiv:2106.05508*, 2021. DOI:10.48550/arXiv.2106.05508.

[8] P. Chen, X. Du, Z. Lu, J. Wu, and P. C. Hung, "EVFL: An explainable vertical federated learning for data-oriented Artificial Intelligence systems," *Journal of Systems Architecture*, vol. 126, p. 102474, 2022.

[9] L. Xu, J. Chen, S. Chang, C. Wang, and B. Li, "Toward Quality-Aware Data Valuation in Learning Algorithms: Practices, Challenges, and Beyond," *IEEE Network*, vol. 38, pp. 213–219, Sep. 2024.

[10] J. Wang, L. Zhang, A. Li, X. You, and H. Cheng, "Efficient Participant Contribution Evaluation for Horizontal and Vertical Federated Learning," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 911–923, May 2022.

[11] Z. Zhang, Y. Yang, Y. Dai, Q. Wang, Y. Yu, L. Qu, and Z. Xu, "FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models," in *Findings of the Association for Computational Linguistics: ACL 2023* (Anna Rogers and Jordan Boyd-Graber and Naoaki Okazaki, ed.), p. 9963–9977, Association for Computational Linguistics (ACL), 2023. Annual Meeting of the Association of Computational Linguistics 2023, ACL 2023 ; Conference date: 09-07-2023 Through 14-07-2023.

[12] H. Zhao, W. Du, F. Li, P. Li, and G. Liu, "FedPrompt: Communication-Efficient and Privacy-Preserving Prompt Tuning in Federated Learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, June 2023.

[13] E. Shammar, X. Cui, and M. A. Al-qaness, "Swarm Learning: A Survey of Concepts, Applications, and Trends," *arXiv preprint arXiv:2405.00556*, 2024.

[14] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng, and Q. Yan, "A Blockchain-Based Decentralized Federated Learning Framework with Committee Consensus," *IEEE Network*, vol. 35, pp. 234–241, January 2021.

[15] G. Liang, S. Chang, M. Dai, and H. Zhu, "FLoomChecker: Repelling Free-riders in Federated Learning via Training Integrity Verification," in *2024 IEEE 30th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 194–201, Oct 2024.