Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Knowledge distillation for financial large language models: a systematic review of strategies, applications, and evaluation*

Jiaqi SHI[§1,2], Xulong ZHANG[§1], Xiaoyang QU[1], Junfei XIE[1,2], Jianzong WANG[‡1]

[1]*Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518000, China*

[2]*University of Science and Technology of China, Hefei 230027, China*

E-mail: civilizwa@mail.ustc.edu.cn; zhangxulong@ieee.org; quxiaoy@gmail.com;
xiejunfei@mail.ustc.edu.cn; jzwang@188.com

Received Apr. 30, 2025; Revision accepted Sept. 5, 2025; Crosschecked

Unedited

Abstract: Financial large language models (FinLLMs) offer immense potential for financial applications. While excessive deployment expenditures and considerable inference latency constitute major obstacles. As a prominent compression methodology, knowledge distillation (KD) offers an effective solution to these difficulties. A comprehensive survey is conducted in this work on how KD interacts with FinLLMs, covering three core aspects such an strategy, application, and evaluation. At the strategy level, this review introduces a structured taxonomy to comparatively analyze existing distillation pathways. At the application level, this review puts forward a logical "upstream-midstream-downstream" framework to systematically explain the practical value of distilled models in the financial field. At the evaluation level, to tackle the absence of standards in the financial field, this review constructs a comprehensive evaluation framework that proceeds from multiple dimensions such as financial accuracy, reasoning fidelity, and robustness. In summary, this research aims to provide a clear roadmap for this interdisciplinary field, to accelerate development of distilled FinLLMs.

Key words: Financial large language models; Knowledge distillation; Model compression; Quantitative trading

https://doi.org/10.1631/FITEE.2500282 **CLC number:**

1 Introduction

Over the past few years, substantial progress in natural language processing (NLP) has been achieved through the development of large language models (LLMs), with their powerful capabilities in contextual understanding, text generation (TG), and reasoning indicating broad application prospects across various industries (Zhao et al., 2024). The financial industry, being highly dependent on information processing, analysis, and decision-making (DM), naturally becomes an important application scenario for LLM technology (Li et al., 2023). Financial large language models (FinLLMs) are trained or fine-tuned on general LLMs using specialized data from the finan-

‡ Corresponding author
§ These two authors contributed equally to this work
* Project supported by the National Key Research and Development Program of China (Youth Scientist Project) (No. 2024YFB4504300)
ORCID: Jianzong WANG, https://orcid.org/0000-0002-9237-4231;
© Zhejiang University Press 2025

cial domain, aiming to better understand financial terminology, capture market dynamics, and execute finance-specific tasks (Lee et al., 2025). The emergence of models marks the rise of FinLLMs research (Liu et al., 2023; Wu et al., 2023; Xie et al., 2023; Zhang and Yang, 2023; Bhatia et al., 2024). These models have demonstrated application potential in various aspects such as financial sentiment analysis, market prediction, quantitative trading, risk management (RM), report generation and summarization, and intelligent customer service (Raza et al., 2025).

However, the deployment of FinLLMs faces several major challenges. First, FinLLMs require high costs (Nie et al., 2024), and training a model such as BloombergGPT is estimated to cost millions of dollars. FinLLMs' large size also makes them difficult to deploy on standard hardware such as mobile devices or regular servers. Additionally, FinLLMs' latency is a critical issue for tasks that need instant decisions, such as algorithmic trading executed at millisecond speeds. These challenges are the main roadblocks preventing FinLLMs from turning their potential into practical value.

KD, recognized for its effectiveness in model compression and the transmission of knowledge, constitutes a fundamental strategy for resolving the difficulties FinLLMs encounter in real-world applications (Acharya et al., 2024). Through a "teacher-student" paradigm (Li et al., 2023), it enables smaller, simpler student models with fewer parameters to learn and inherit the key capabilities of large teacher models, thereby significantly reducing computational resource requirements, shortening inference times, and supporting deployment in resource-constrained environments. Consequently, KD technology can effectively bridge the gap between the powerful potential of FinLLMs and the practical implementation needs of the financial industry.

FinLLMs and KD are key research areas in artificial intelligence (AI). Nie et al. (2024) and Li et al. (2023) have published comprehensive surveys on FinLLMs, detailing their applications and challenges in financial tasks. Similarly, some researchers have provided thorough reviews on KD methods and applications (Xu et al., 2024; Yang et al., 2024). However, there is currently no systematic survey focusing on the integration of KD with FinLLMs. This paper addresses this gap by systematically investigating the synergy between KD and FinLLMs, offering a comprehensive survey to guide future research.

This paper delivers an exhaustive, systematic review of the domain by integrating strategies, application scenarios, and evaluation methods into a cohesive analytical framework. First, this paper divides the distillation strategy into black-box and white-box categories, then analyzes examples according to three different financial scenarios, and finally discusses the challenges of KD for FinLLMS. Next, it examines applications by proposing a logical "upstream-midstream-downstream" framework to more clearly elucidate the practical value of distilled models in the financial filed. Finally, to address the critical lack of evaluation metrics, this paper proposes a set of corresponding, quantifiable specific measurement indicators for the three different financial task scenarios of trading strategies.

The rest of the paper is structured in the following manner: section 2 offers a distinct introduction to the foundations of FinLLMs and KD. Section 3 divides the distillation strategy into black-box and white-box categories and then analyzes examples according to three different financial scenarios. Section 4 proposes an "upstream-

Unedited

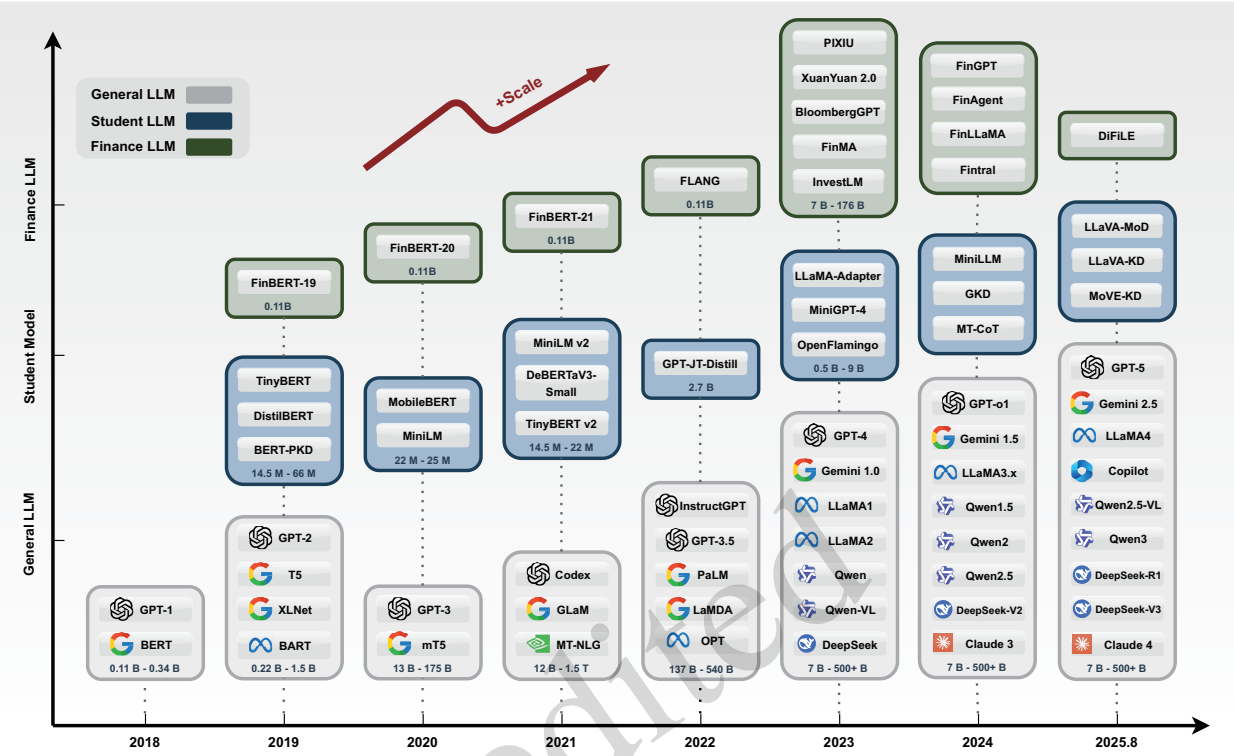


Fig. 1 A chronological diagram showcasing the transition of LLMs from broad-domain architectures to finance-oriented versions, accompanied by advances in KD

Unedited

midstream-downstream" framework to explain the practical applications of distilled models in finance. Section 5 provides evaluation indicators for three special financial tasks. Section 6 provides the concluding remarks.

2 Background

2.1 FinLLMs

FinLLMs are LLMs specifically trained or adapted for tasks within the financial domain (Lee et al., 2025). They are typically built upon general LLM architectures such as the Transformer (Raza et al., 2025). FinLLMs often contain billions of parameters (frequently exceeding 7 billion) and acquire core capabilities for processing financial information through pre-training or fine-tuning on corpora comprising extensive financial documents—spanning news pieces, annual filings, academic studies, and social-media chatter—as well as, in some cases, structured data including market quotes and financial indicators. In addition to inheriting the powerful natural language understanding and generation abilities of general-purpose LLMs, FinLLMs develop specialized expertise in financial knowledge and reasoning (Li et al., 2023). Fig. 1 depicts the progression from general-purpose LMs to finance-specific models alongside the development of KD.

As depicted in Fig. 2, the essential functionalities encompassed by FinLLMs include: precise comprehension of financial terminology, industry conventions, and complex reporting structures; efficient extraction of key information or generation of summaries from large volumes of financial text (Raza et al., 2025); analysis of market sentiment and entity-specific sentiment (Nie et al., 2024); engagement in financial question answering (QA),

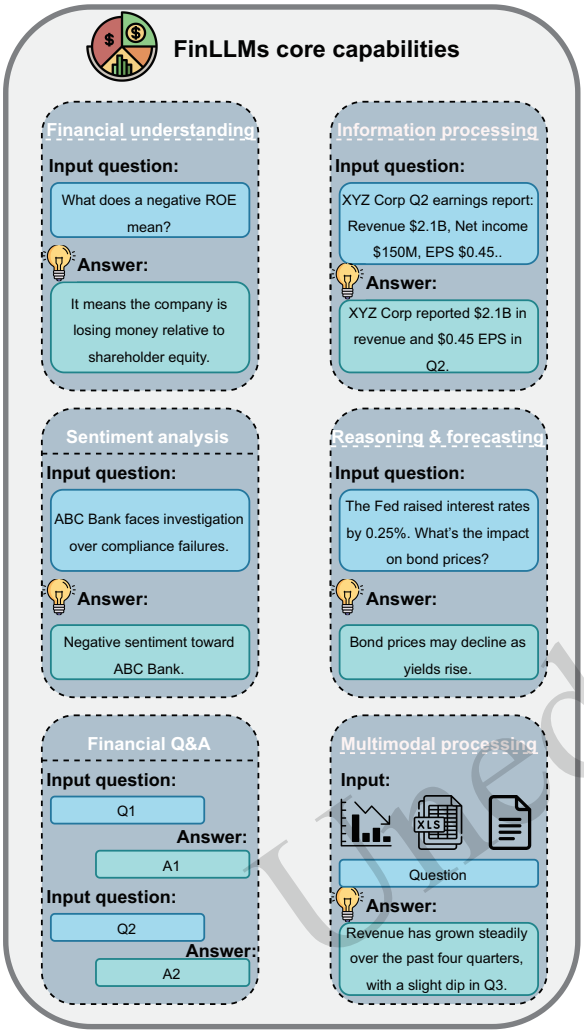


Fig. 2 FinLLMs core capabilities diagram

provision of investment suggestions which must be interpreted with caution, and functioning as intelligent customer service agents (Raza et al., 2025); logical reasoning based on available information to support risk assessment and market forecasting (FO); and processing of multimodal financial data, including text, tables, time series, and visual elements such as charts (Kong et al., 2024).

Table 1 presents a comparison of several primary construction pathways for FinLLMs. Mixed pre-training from scratch constitutes one possible route, represented by BloombergGPT (Wu et al., 2023), involving the construction of a model de novo by leveraging both general-domain sources and large-scale financial corpora. Although this approach enables deep customization, it requires substantial computational resources and is typically feasible only for large institutions.In contrast, continual pre-training involves incrementally training an existing general LLM on financial data, as demonstrated by FinTral (Bhatia et al., 2024). This method significantly reduces training costs while improving the model's understanding of financial concepts and language.Another widely adopted strategy is instruction fine-tuning, which adapts general LLMs using domain-specific instruction datasets to enhance performance on targeted financial tasks. Representative models include PIXIU (Xie et al., 2023), FinGPT (Liu et al., 2023), and InvestLM (Yang et al., 2023). This approach offers advantages such as lower cost, high flexibility, and rapid deployment; however, it may affect the model's generalization ability across broader domains.

Unedited

Table 1 Comparison of finLLMs construction approaches

Construction approach	Representative model
Mixed pre-training from scratch	BloombergGPT
Continual pre-training	FinTral
Instruction fine-tuning	PIXIU InvestLM FinGPT

2.2 KD

Operating on a teacher–student paradigm (Li et al., 2023), KD, as noted by (Li et al., 2023), is frequently employed to achieve model compression and facilitate knowledge transfer. As shown in Fig. 3, through KD, expertise is transferred from a large, high-capacity teacher model to a student model that is smaller and more computationally efficient. The primary objective is to reduce computational resource consumption and inference la-

tency, while preserving the performance of the teacher model as much as possible. The core mech-

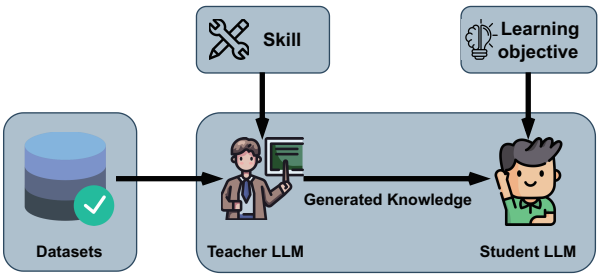


Fig. 3 An illustration of the comprehensive process for disseminating knowledge from a mentor model to a learner model is provided

anism involves training the student model to replicate various forms of information from the teacher, including output distributions, intermediate representations, and relational structures. This is typically achieved by minimizing a composite loss function that integrates both the task-specific loss and the distillation loss.

KD is extensively used in deep learning, particularly in scenarios involving large-scale models such as FinLLMs. By employing either white-box or black-box distillation methods, KD significantly improves model efficiency and deployment flexibility, enabling real-time applications. However, it inherently involves a trade-off between performance fidelity and compression effectiveness (Kong et al., 2024).

$$\mathcal{L}_{\mathrm{KD}} = (1-\alpha)\,\mathcal{L}_{\mathrm{task}}(y, P_s) + \alpha\,\mathcal{L}_{\mathrm{distill}}(P_t, P_s), \quad (1)$$

where $\mathcal{L}_{\mathrm{KD}}$ represents the overall loss function used in KD. It is formed by a weighted combination of two elements. The first term, $\mathcal{L}_{\mathrm{task}}(y, P_s)$, is the task-specific loss. It is typically computed using cross-entropy between the student's predicted distribution P_s and the true labels y. This term ensures that the student remains accurate on the primary task. The second term, $\mathcal{L}_{\mathrm{distill}}(P_t, P_s)$, is the distillation loss. This process seeks to harmonize the student' s output P_s with the teacher' s soft target distribution P_t through the reduction of their divergence. This mechanism drives the student to assimilate the teacher' s sophisticated predictive conduct. The hyperparameter $\alpha \in [0, 1]$ determines how the two loss components are balanced. It determines whether the student model focuses more on ground-truth supervision or on imitating the teacher.

3 Strategies of distilled FinLLMs

Applying KD to FinLLMs aims to overcome the limitations of the latter, enhancing their practical value and feasibility in financial scenarios. More precisely, through the conveyance of functionalities from a substantial teacher FinLLM to a compact student model. KD can significantly reduce the initially high computational costs and energy consumption for both training and inference (Xu et al., 2024), and substantially shorten inference latency, meeting the stringent response speed requirements of financial applications such as quantitative trading, real-time risk control, and online customer service. The resulting lightweight student models are easier to deploy in resource-constrained environments, effectively broadening the application scope of FinLLMs.

To systematically understand the application of KD in FinLLMs, distillation strategies are grouped into white-box and black-box methods. The white-box approach scrutinizes three independent channels for knowledge dissemination. The black-box method focuses on methods that rely on synthetic data.

3.1 KD of white-box

As shown in the lower part of Fig. 4, a white-box model implies the capability to observe and utilize the teacher model' s internal processes (Gu et al., 2024; Liang et al., 2023; Liu et al.,

Unedited

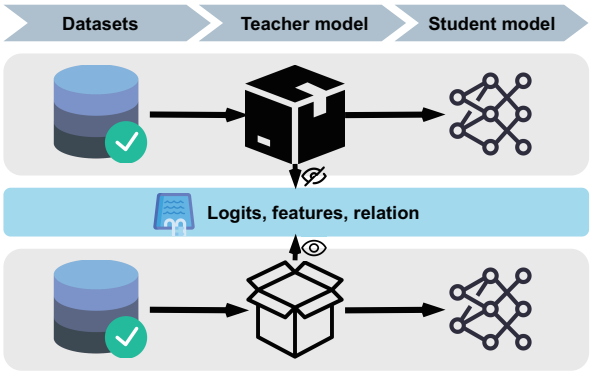


Fig. 4 White-box vs. black-box KD

2023; Timiryasov and Tastet, 2023; Agarwal et al., 2024). In addition to using the teacher's final outputs, its internal components—such as hidden layers and intermediate computations—can be employed in training the student model. This enables the student to acquire deeper knowledge, often resulting in improved performance (Qin et al., 2023; Wen et al., 2023; Zhao and Zhu, 2024; Wan et al., 2024).

3.1.1 Logit-based distillation

As shown in Fig. 5, logit-based distillation is the most established and widely adopted approach (Kim et al., 2023). The fundamental principle entails steering the student model to mirror the teacher' s logits at the output stratum. These logits contain the teacher's "dark knowledge" (Burnett and Lloyd, 2020), capturing subtle relationships between classes and reflecting its decision-making logic.

The typical loss function is the KL divergence with a temperature parameter τ(Hershey and Olsen, 2007; Van Erven and Harremos, 2014):

$$\mathcal{L}_{\text{Logit}} = \text{KL}(\sigma(\frac{z_s}{\tau}), \sigma(\frac{z_t}{\tau})), \tag{2}$$

where z_s and z_t are defined as the output scores from the student model and the teacher model, respectively. Through its role, the function σ embodies the SoftMax operation, while τ functions as a temperature hyperparameter to mitigate the sharpness of probability distributions. This loss is typically combined with a standard task loss.

3.1.2 Feature-based distillation

Feature-based distillation concentrates on conveying expertise from the teacher model' s intermediate strata to the student model (Liu et al., 2023). Rather than mimicking the final output, the student assimilates the teacher' s "cognitive framework". This is achieved by minimizing a similarity loss between feature maps f_T and f_S from corresponding layers of the teacher and student models. A A widely utilized measure is the L2 norm (mean squared error), expressed as:

$$\mathcal{L}_{textFeat} = ||Phi_T(f_T(x)) - Phi_S(f_S(x))||_2^2, \tag{3}$$

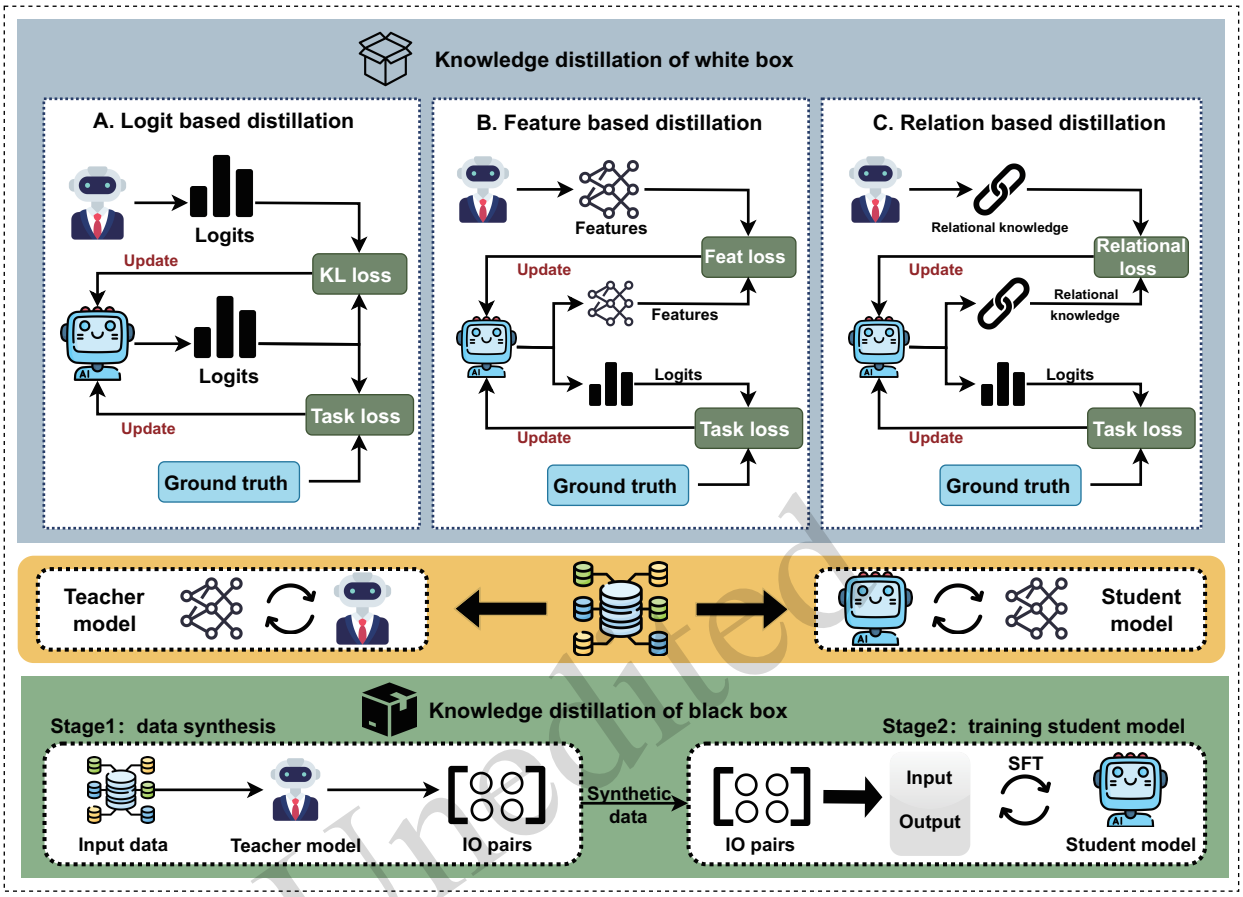
where $f_T(x)$ and $f_S(x)$ are identified as the feature maps generated by the teacher and student models for a given input x. When feature dimensions are mismatched, the optional linear transformation Φ is employed to ensure dimensional consistency. The DiFiLE (Hristova and Satani, 2025) project applied this method by aligning the word embeddings of two models.

3.1.3 Relation-based distillation

Relation-based distillation does not require the student to mimic individual outputs(Zhao et al., 2024). Instead, it transfers the structural relationships between data samples as understood by the teacher. This is achieved by defining a relational loss that minimizes discrepancies in geometric structures—such as relative distances, angles, or correlations—between the teacher and student feature spaces.

Relational knowledge is considered a deeper form of "dark knowledge." By jointly optimizing relational and task losses, the student learns

Unedited



Unedited

Fig. 5 Overview of distillation strategies

not only to solve the task but also to internalize the teacher's data organization logic, thereby improving generalization. A common way to capture relational knowledge is through the Gram matrix, which quantifies feature correlations via inner products:

$$\mathcal{L}_{\text{Relation}} = ||G(f_T(x)) - G(f_S(x))||_F^2, \quad (4)$$

where $f_T(x)$ and $f_S(x)$ are characterized as the feature maps, which are produced by the intermediate layers of the teacher and student models, respectively, for a specified input x. The function $G(\cdot)$ computes the Gram matrix from a feature map using the operation $G(F) = FF^T$, where a feature map F is multiplied by its transpose F^T. The resulting Gram matrix captures the inner products and correlations among the features, effectively representing the model's internal knowledge structure. Through its definition, the expression $\|\cdot\|_F^2$ represents the squared Frobenius norm, quantifying the total of squared elements within the disparity matrix of the teacher' s and student' s Gram matrices. This loss, when reduced, enables the student model to be trained in reflecting the teacher' s feature relationships, thus promoting the assimilation of the teacher' s sophisticated knowledge frameworks.

3.2 KD of black-box

As shown in the upper part of Fig. 4, black-box KD encompasses situations in which the internal mechanisms of the teacher model remain unreachable (Chang et al., 2019; Nguyen et al., 2022; Galichin et al., 2025). The student learns solely

from the teacher's final outputs. The teacher is regarded as a "black box", meaning its architecture, parameters, and intermediate activations are unknown (Wang, 2021; Han et al., 2024). The student is provided with identical input and strives to mimic the teacher' s output.

3.2.1 Synthetic data-based distillation

In cases where the teacher model is served via an Application Programming Interface (API), or when only the input–output interface is exposed due to privacy or intellectual property (IP) concerns, synthetic data-based distillation becomes essential (Li et al., 2023).

As shown in Fig. 5, this method bypasses internal access through two stages. First, in the data synthesis stage, input data from flexible sources is fed into the teacher model (Chen et al., 2023; Lei and Tao, 2023). Its outputs are used to create a synthetic training set (i.e., input–output pairs). Secondly, within the student training phase, this synthetically generated dataset is employed to instruct the student model via a conventional supervised learning paradigm. The ultimate aim of this procedure is to minimize the discrepancy between the student's output predictions and the teacher's corresponding pseudo-labels, thereby enabling the acquisition of the teacher's input-output mapping.

The loss function seeks to harmonize the output distributions of the student and teacher models, typically using KL divergence:

$$\mathcal{L}_{\text{Synthetic}} = \sum_{x \in D_{syn}} D_{KL}(\sigma(\frac{z_S}{\tau})||\sigma(\frac{z_T}{\tau})), \quad (5)$$

where $L_{\text{Synthetic}}$ represents the total distillation loss calculated over the synthetic dataset D_{syn}. The overarching aim is to reduce this discrepancy, thereby compelling the student model to assimilate information derived from the teacher model's performance on the synthetic dataset. The total loss is computed by accumulating the individual losses generated for each sample x within the synthetic dataset D_{syn}. Using KL divergence, symbolized as $D_{\text{KL}}(\cdot||\cdot)$, the individual loss is measured. This metric specifically calculates the divergence between a pair of probability distributions. The KL divergence specifically operates on two tempered probability distributions. The first is the student model's distribution $\sigma(\frac{z_S}{\tau})$, which is derived from its raw logits z_S. The second is the teacher model's distribution $\sigma(\frac{z_T}{\tau})$, similarly derived from its raw logits z_T. Both distributions are generated by applying the SoftMax function $\sigma(\cdot)$ and a temperature parameter τ.

Through synthetic data-based KD, the effective conveyance of knowledge is possible from a robust teacher model, which acts as a source, to a compact student model, even under black-box conditions where the inner mechanisms of the teacher model are entirely concealed.

4 Applications of distilled FinLLMs

KD makes FinLLMs more practical by significantly reducing their computational cost and inference latency, thereby enhancing deployment flexibility. To systematically analyze these applications, this section organizes use cases into a three-tier framework: "Upstream: Information Processing", "Midstream: Analysis and Prediction", and "Downstream: Decision Support and Execution". The framework is shown in Fig. 6

4.1 Upstream: financial information processing

Upstream tasks are foundational to all subsequent financial analysis. Their core objective is to convert vast amounts of unstructured information into structured, machine-readable data. In this role, distilled FinLLMs act as "information

Unedited

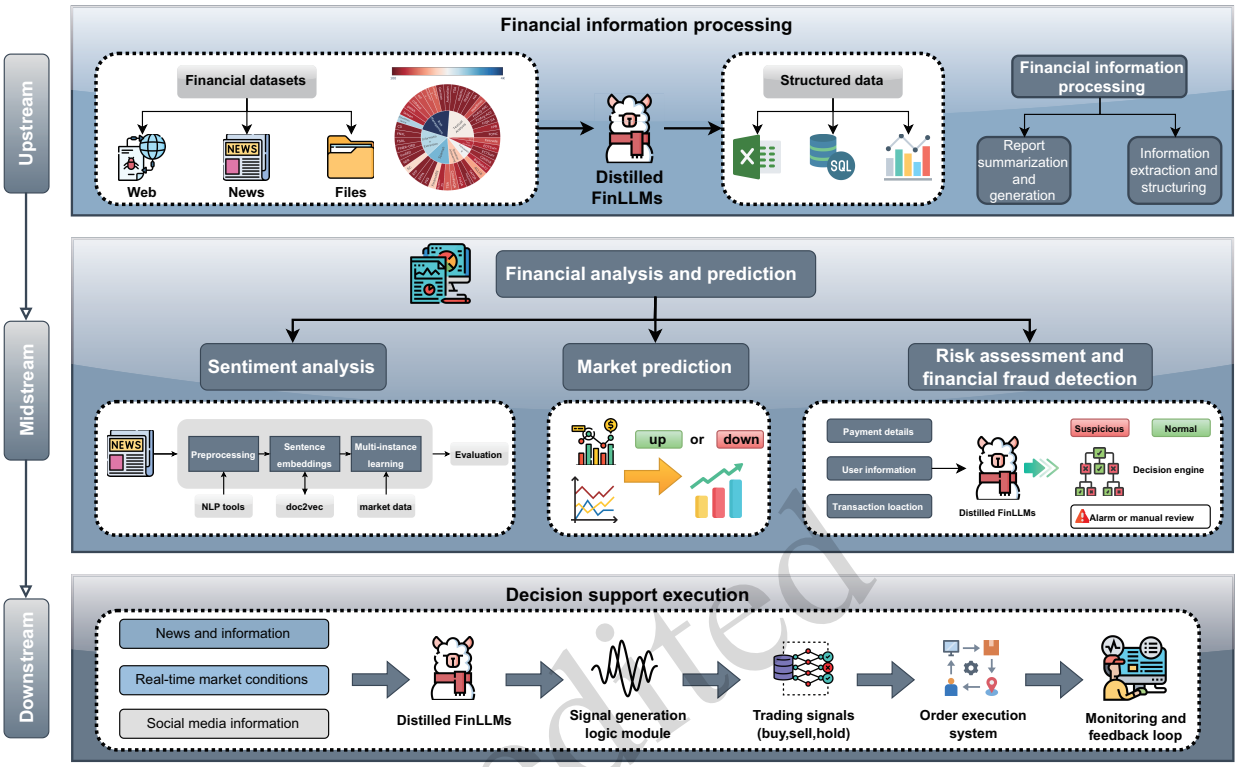


Fig. 6 Three-tier framework: upstream-midstream-downstream

Unedited

refineries" greatly improving processing efficiency and scale.

4.1.1 Report summarization and generation

Financial analysts must process numerous lengthy documents, such as annual reports and news releases. Distilled FinLLMs can generate accurate summaries for these documents, with applications extending to real-time scenarios. For instance, an end-to-end trading system uses FinGPT (Liu et al., 2023) to summarize live news feeds, which are then used for sentiment analysis to generate trading signals, demonstrating the value of low-latency models. The DiFiLE project also shows that a distilled student model, just 64% the size of its teacher, can process long 10-K reports with 30% less training time while maintaining comparable performance.

4.1.2 Information extraction and structuring

Automatically extracting key data from unstructured text is a critical, foundational task. This technology has reached commercial maturity. In academia, the FinGPT (Liu et al., 2023) project has shown that how parameter-efficient fine-tuning (PEFT) techniques(Han et al., 2024) like LoRA (Sundaram et al., 2019) can adapt general models like Llama2 (Touvron et al., 2023) into specialized experts for financial named entity recognition (NER). In practice, commercial NLP suites from companies like John Snow Labs can precisely extract specific entities from financial filings, such as company filing numbers (CFN) and trading symbols, confirming the industrial-grade readiness of this technology. The efficiency of distilled models makes them ideal for these tasks, providing robust data for downstream applications.

4.2 Midstream: financial analysis & prediction

Midstream tasks use the structured data from upstream processes to perform in-depth reasoning and prediction, linking data to decisions. In this paper, by retaining core analytical power at a lower operational cost, distilled FinLLMs show immense potential.

4.2.1 Sentiment analysis

This task aims to quantify the sentiment in texts such as news and social media posts, using it as a feature for predicting market trends. To capture market sentiment in real time, models must be both fast and accurate. The FinBERT (Huang et al., 2023) project is a prime example. Through an innovative two-stage distillation process, it produced a specialized financial sentiment model with only 14.5 million parameters. This model not only achieved 98.9% of its large teacher's accuracy on a standard benchmark but also surpassed it on an out-of-domain dataset (107.5% accuracy), demonstrating that well-designed distillation can improve both efficiency and generalization.

4.2.2 Market prediction

Distilled models can be used to predict stock prices, volatility, or other market indicators. Their lower resource requirements make it feasible to perform more frequent model updates and cover a broader range of assets.

4.2.3 Risk assessment and financial fraud detection

Financial institutions must constantly monitor various risks. For financial fraud detection, which requires real-time processing of massive transaction volumes, the low latency of distilled FinLLMs makes them an ideal choice.

The middle part of Fig. 6 shows a real-time fraud detection system that uses a distilled FinLLM. In this system, transaction data flows in and are rapidly scored by the distilled FinLLM module. A decision engine then flags suspicious transactions for review, a process whose effectiveness hinges on the model's high throughput and low latency. Underscoring this, Tang and Liu (2024) found that the distilled student model achieved an F1-score of 92.87%, significantly outperforming its teachers (81.42%) and showing that distillation can even enhance performance.

4.2.4 Compliance checks

Automation is crucial for navigating complex financial regulations. A study from the Fraunhofer Institute compared the performance of large LLMs with lightweight models for this task. It found that a lightweight approach using FinBERT (Huang et al., 2023) embeddings and a simple MLP classifier achieved an F1-score (0.74) comparable to that of a fine-tuned GPT-4o mini (0.78). This demonstrates that for highly specialized tasks, smaller, cost-effective models are often the more practical choice.

4.3 Downstream: decision support & execution

Downstream applications represent the "final mile" where analytical insights are converted into actionable decisions or user-facing services. The low latency and cost-effectiveness of distilled FinLLMs are critical for these time-sensitive, large-scale deployments.

4.3.1 Quantitative trading

In quantitative trading, millisecond latencies can determine profit or loss. Research reveals

Unedited

a "latency-profitability curve" where the fastest model is not always the best; rather, there is an optimal balance between speed and quality. The lower part of Fig. 6 shows a quantitative trading system built around a distilled FinLLM. Inputs such as market data and news are processed by the distilled FinLLM module for rapid inference. A logic module converts these insights into trading signals (buy, sell, and hold), which are then sent to an execution system. The core principle of this architecture is to leverage the distilled model for low-latency, high-throughput real-time decision support.

4.3.2 Robo-advisors

The next generation of robo-advisors aims to provide "hyper-personalized" advice, a computationally intensive task when scaled to millions of users. While industry examples like Morgan Stanley's AI platform have proven the value of AI-driven personalization (increasing client engagement by 35%), KD is the key technology to "democratize" this capability and offer it to the mass market at a low cost.

4.3.3 Intelligent customer service

Financial chatbots must understand user intent accurately and respond with low latency. In this paper, industry cases clearly show the value of distillation. For example, Snorkel AI used a large model to bootstrap the creation of a smaller, specialized bank chatbot model. Concurrently, cloud platforms like Amazon Bedrock now offer model distillation as a commercial service, specifically targeting the deployment of large-scale, low-latency chatbot systems.

In summary, these applications demonstrate that distillation technology makes it feasible to use LLMs for efficient, real-time, and large-scale financial tasks. However, the successful deployment of these systems hinges on both the model and the high-quality data, reliable execution systems, and continuous performance monitoring. Therefore, rigorous validation and responsible deployment of distilled models are paramount in the high-stakes financial sector.

5 Evaluation of distilled FinLLMs

5.1 Tasks and datasets for FinLLMs

In this section, we outline the evaluation tasks for large-scale financial models and the corresponding datasets. We have systematically organized this content into a tabular format, as shown in Table 2.

We refer to the classification method proposed in FinBen (Xie et al., 2024) to categorize financial tasks into six types: IE, textual analysis (TA), QA, TG, RM, FO and DM. IE refers to the automatic extraction of structured information (Singh, 2018). TA, sometimes called text mining or textual analytics. It uses FinLLMs to analyze text for patterns, sentiment, topics. QA is a subfield of NLP and information retrieval focused on building systems that can automatically respond to human queries posed in natural language (Pandya and Bhatt, 2021). TG, a subset of Natural Language Generation, involves producing coherent and contextually appropriate text, given some input or condition (Li et al., 2024). RM is the systematic process of identifying, assessing, and mitigating risks—events that could adversely affect organizational objectives. FO is the practice of predicting future events or trends using historical and current data, applying quantitative methods or qualitative insights to support strategic DM. DM is the process of choosing among alternatives by structuring problems, weighing cri-

Unedited

Table 2 A summary of datasets, tasks, description and metrics of FinLLMs.

Dataset	Task	Description	Metric
FINER-ORD(Shah et al., 2023)	IE	NER	Entity F1(PER,LOC,ORG)
CRA(Alvarado et al., 2015)	IE	NER	Entity F1(PER,LOC,ORG)
FINRED(Sharma et al., 2022)	IE	Relationship extraction	Recall,PPV,F1
REFinD(Kaur et al., 2023)	IE	Relationship extraction	Recall,PPV,F1
FinCausal(Mariko et al., 2020)	IE	Causal detection	Recall,PPV,F1
FNXL(Sharma et al., 2023)	IE	Numeric labeling	Recall,PPV,F1,Hits@1
FSRL(Lamm et al., 2018)	IE	Textual analogy parsing	Recall,PPV,F1
SEntFiN(Sinha et al., 2022)	TA	Sentiment analysis	Recall,Acc,PPV,F1
FinLin(Daudert, 2022)	TA	Sentiment analysis	Recall,Acc,PPV,F1
SentiEcon(Moreno-Ortiz et al., 2020)	TA	Sentiment analysis	Recall,Acc,PPV,F1
Headliness(Sinha and Khandait, 2021)	TA	News headline categorization	F1
FOMC(Shah et al., 2023)	TA	Hawkish-Dovish classification	F1,Acc
FinArg-ACC(Sy et al., 2023)	TA	Argument component categorization	F1,Acc
MultiFin(Jorgensen et al., 2023)	TA	Multi-classclassification	F1,Acc
M&A(Yang et al., 2020)	TA	Deal completeness classification	F1,Acc
MLESG(Chen et al., 2023)	TA	ESG issue identification	F1,Acc
FinQA(Chen et al., 2021)	QA	Single-turn QA	EM Accuracy, F1
TAT-QA(Zhu et al., 2021)	QA	Single-turn QA	EM Accuracy, F1
ConvFinQA(Chen et al., 2022)	QA	Multi-turn QA	EM Accuracy
ECTSum(Mukherjee et al., 2022)	TG	Text summarization	ROUGE,BERTScore,BARTScore
EDTSum(Zhou et al., 2021)	TG	Text summarization	ROUGE,BERTScore,BARTScore
LendingClub(Feng et al., 2023)	RM	Credit scoring	Acc,F1,MCC,Miss
ccfraud(Varmedja et al., 2019)	RM	Fraud detection	Acc,F1,MCC,Miss
BigData22(Soun et al., 2022)	FO	Stock trend forecasting	Acc,MCC
ACL18(Xu and Cohen, 2018)	FO	Stock trend forecasting	Acc,MCC
CIKM18(Wu et al., 2018)	FO	Stock trend forecasting	Acc,MCC

* IE, Information extraction; TA, Textual analysis; QA, Question answering; TG, Text generation; RM, Risk management; FO, Forecasting; PPV, Positive Predictive Value; Acc, Accuracy.

Unedited

teria, and selecting the most fitting option, often supported by analytic frameworks or systems.

5.2 A multi-dimensional evaluation framework for distilled FinLLMs

Although KD provides an efficient and feasible path for the deployment of FinLLMs, this optimization process also brings new challenges to the performance, reliability, and security of the models. In the high-risk financial field, relying solely on traditional NLP benchmarks is far from enough, as they cannot measure the accuracy of financial facts nor detect the potential decline in the "inference fidelity" and "robustness" of the model in a dynamic market (Brown et al., 2020). To fill this critical gap, this section aims to propose a structured, multi-dimensional evaluation framework, providing a set of systematic evaluation criteria for the research and deployment of distilled FinLLMs.

This section aims to redefine the six core dimensions required for evaluating the distillation of FinLLM. While these dimensions are designed to be as orthogonal as possible to minimize overlap and ensure focused assessment, we acknowledge potential interdependencies (Ribeiro et al., 2016). These dimensions collectively form the theoretical foundation of this evaluation framework, with

clear boundaries defined to handle any overlaps during implementation.

5.2.1 Financial accuracy and factuality

This is the most basic dimension, but its connotation goes beyond the traditional NLP concept of text similarity. It demands that the model' s output remains factually aligned with real-world financial data. This dimension is particularly concerned with auditing the inherent "hallucination" (Hallucination) problem of LLMs (Ji et al., 2023). Any statement that contains erroneous financial data, contradicts market facts, or lacks evidence should be considered "inaccurate" or "unfactual". To distinguish, "accuracy" focuses on factual correctness , while "factuality" emphasizes evidence-based consistency without fabrication (Costantino and Coletti, 2008).

For trading strategy tasks, a model' s main objective is to maximize risk-adjusted returns. This dimension is measured by the Sharpe Ratio, distinguishing accuracy (factual alignment with historical returns) from factuality (evidence-based predictions) (Jensen, 1968). As the authoritative metric for risk-adjusted returns, the Sharpe Ratio directly quantifies the strategy's ability to convert risk into effective returns, serving as the core standard for evaluating its ultimate economic value.

$$\text{Sharpe Ratio} = \frac{E[R_p - R_f]}{\sigma_p}. \tag{6}$$

As a metric for risk-adjusted return, the Sharpe Ratio assesses an investment's performance against a risk-free asset, incorporating the investment's inherent risk. R_p refers to the anticipated return of a given portfolio or trading strategy, R_f denotes the risk-free rate of return, and σ_p corresponds to the standard deviation of the strategy's excess return $(R_p - R_f)$, a measure that quantifies its overall volatility or risk. A higher Sharpe Ratio indicates a superior compensation for the level of risk undertaken, rendering it an established benchmark for evaluating the efficacy and performance of various trading approaches.

For risk assessment tasks, the evaluation focus for this category of tasks lies in precision, reliability, and compliance.This dimension is measured by Value at Risk (VaR) Accuracy, with accuracy focusing on precise loss estimates and factuality on evidence alignment (Duffie and Pan, 1997; Jorion, 1996).

$$P(\Delta V \leq -\text{VaR}) = 1 - c. \tag{7}$$

Value at Risk (VaR) is a statistical metric that estimates the possible amount of financial risk in a company or investment portfolio over a defined period. The change in a portfolio's value throughout a specific period is symbolized by ΔV, c is the confidence level (e.g., 95% or 99%), and VaR is the maximum loss expected at that confidence level. The equation thus defines VaR as the threshold for which the probability of the portfolio's loss exceeding this value is equal to $(1 - c)$. VaR accuracy, as discussed in the paper, is commonly appraised via backtesting, a procedure that entails a comparison of the number of realized losses that exceed the VaR projection with the number of exceedances forecasted by the model at the $(1 - c)$ probability threshold.

Through backtesting, the precision of the model's predictions regarding market tail risk can be verified, which is key to evaluating its core functional effectiveness.

For sentiment extraction tasks, the evaluation of these tasks focuses on the validity of the information produced and the model's processing capabilities (Loughran and McDonald, 2011).The evaluation of this dimension does not focus on linguistic correctness but is instead achieved by calculating the correlation between the model's out-

Unedited

put and key market indicators (Dow and Gorton, 1997), with accuracy on factual matches and factuality on evidence support. This relationship is usually measured by the Pearson correlation coefficient:

$$\rho_{S,M} = \frac{\mathrm{cov}(S, M)}{\sigma_S \sigma_M}, \tag{8}$$

where S represents the time series of the sentiment index generated by the model, and M represents the time series of the market indicator. $\mathrm{cov}(S, M)$ is the covariance between these two series, while σ_S and σ_M are their respective standard deviations. With values ranging from -1 to +1, the correlation coefficient $\rho_{S,M}$ shows a stronger linear relationship as its magnitude approaches 1. A statistically significant correlation serves as direct evidence of the information's financial value.

5.2.2 Reasoning fidelity

The focus of this dimension is to determine if, through KD, the student model has simply committed the teacher model's responses to memory or has genuinely grasped the foundational logical reasoning. A high-fidelity model should have a trustworthy and interpretable decision path. When evaluating, indicators such as "faithfulness" and "informativeness" should be used to quantify the quality of the explanation, ensuring that the explanation is based on evidence and meaningful (Jain and Wallace, 2019).

5.2.3 Robustness

The financial market is full of uncertainty. The goal of this dimension is to evaluate whether the model's performance will deteriorate sharply when faced with extreme market fluctuations, noisy data, or unexpected "black swan" events. Furthermore, robustness should also include two key aspects, each with specific evaluation methods: backtesting robustness and time robustness. Backtesting robustness evaluates whether the model has "lookahead bias" in the backtest, which means it has inadvertently used future information. This can be assessed using statistical tests like out-of-sample validation or rolling window analysis to detect bias (De Prado, 2018).Time robustness evaluates whether the signals generated by the model have a long-term "signal decay" problem, that is, whether the effectiveness of the strategy will weaken over time due to market adaptation. This can be quantified via time-series metrics such as exponential decay modeling or performance degradation curves over extended periods (Bollerslev, 1986).

Sometimes, Robustness is assessed through the Maximum Drawdown (MDD), supplemented by sub-metrics for backtesting and time robustness (Magdon-Ismail and Atiya, 2004).

$$MDD = \max_{t \in [0,T]} \left(\frac{\sup_{\tau \in [0,t]} X(\tau) - X(t)}{\sup_{\tau \in [0,t]} X(\tau)} \right). \tag{9}$$

A risk metric known as Maximum Drawdown (MDD) measures the biggest decline in a portfolio' s or strategy's value from its highest point to its lowest point over a specific timeframe. Where t is the total time horizon, $X(t)$ represents the portfolio's asset value at time t, and $\sup_{\tau \in [0,t]} X(\tau)$ denotes the peak (supremum) value of the portfolio in the interval from the start ($\tau = 0$) up to time t. The MDD calculation thus finds the maximum percentage loss from a running peak to a subsequent low point. This metric is crucial for evaluating robustness because it reveals the worst-case loss an investor might have experienced, providing a direct reflection of a strategy's resilience in adverse market conditions.

The MDD reveals the strategy's performance under the most unfavorable historical market conditions, providing a direct reflection of its risk re-

Unedited

silience and stability.

5.2.4 Uncertainty quantification

This is a key new dimension for the probabilistic output characteristics of LLMs. Because the outputs of LLMs are drawn from a distribution instead of being deterministic, a single prediction may be misleading. This dimension seeks to assess the model' s capability to deliver a dependable confidence interval for its predictions, which is crucial for effective risk control. Evaluation can involve metrics like calibration error or expected calibration error (ECE) to measure how well predicted probabilities align with actual outcomes (Guo et al., 2017).

5.2.5 Compliance and security

This dimension represents the rigid constraints of real-world deployment. It includes two aspects: first, whether the model and its outputs comply with the regulatory requirements of financial regulatory authorities, assessed through automated compliance audits (Barocas et al., 2023); second, whether the model can successfully safeguard the confidentiality and privacy of financial data during processing and generation, evaluated via privacy metrics like differential privacy epsilon values or data leakage tests (Dwork et al., 2006).

5.2.6 Efficiency

This is the fundamental motivation of KD. This dimension seeks to measure the model' s efficiency regarding resource usage, including its inference speed, computational cost, memory usage, energy consumption, and model compression ratio, to determine whether it truly achieves the goal of efficient deployment within specific hardware or cost budgets. Multiple indicators ensure comprehensive coverage, such as FLOPs for computational cost and peak memory usage during inference (Liebenwein et al., 2020).

6 Conclusions

This paper provides an in-depth analysis of KD as a critical solution to the deployment bottlenecks of Financial Large Language Models (FinLLMs), namely their high resource consumption and inference latency. To systematically address this issue, this paper presents three core contributions. First, it establishes a taxonomy for KD strategies, organizing existing methods along two dimensions: the "Knowledge Transfer Pathway" and the "Distillation Strategy". Second, it proposes an "upstream-midstream-downstream" application framework—encompassing information processing, analysis and prediction, and decision support—to clearly demonstrate the value of distilled models across the entire financial workflow. Finally, to address the current lack of evaluation standards, it constructs a comprehensive assessment framework with multiple dimensions, including Financial Accuracy, Reasoning Fidelity, and Robustness, designed to supplement traditional NLP benchmarks. Ultimately, this work provides a strong basis for the dependable deployment and practical value realization of distilled FinLLMs within the financial sector.

Contributors

Jiaqi SHI and Xulong ZHANG conceived the project and developed the methodology. Jiaqi SHI and Xulong ZHANG drafted the paper. Xiaoyang QU provided resources, supervised the study, and administered the project. Junfei XIE and Jianzong WANG revised the paper. Jianzong WANG also acquired funding. All authors read and approved the final paper.

Unedited

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Acharya K, Velasquez A, Song HH, 2024. A Survey on Symbolic Knowledge Distillation of Large Language Models. *IEEE Trans. Artif. Intell.*, 5(12):5928-5948. https://doi.org/10.1109/TAI.2024.3428519

Agarwal R, Vieillard N, Zhou YC, et al., 2024. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. Proc 12th Int. Conf. on Learning Representations.

Alvarado JCS, Verspoor K, Baldwin T, 2015. Domain adaption of named entity recognition to support credit risk assessment. Proc Australasian language technology association workshop, p.84–90.

Barocas S, Hardt M, Narayanan A, 2023. Fairness and machine learning: Limitations and opportunities. MIT press, Cambridge, USA.

Bhatia G, Nagoudi EMB, Cavusoglu H, et al., 2024. FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models. Proc Findings of the Association for Computational Linguistics, p.13064–13087. https://doi.org/10.18653/v1/2024.findings-acl.774

Bollerslev T, 1986. Generalized autoregressive conditional heteroskedasticity. *J econometrics*, 31(3):307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Brown T, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Proc 34th Int Conf on Neural Information Processing Systems, article 159.

Burnett S, Lloyd A, 2020. Hidden and forbidden: conceptualising Dark Knowledge. *J Doc*, 76(6):1341–1358. https://doi.org/10.1108/JD-12-2019-0234

Chang HY, Shejwalkar V, Shokri R, et al., 2019. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. *arXiv preprint arXiv:1912.11279*.

Chen XX, Yu Y, Wang ZY, et al., 2024. Data Distillation Can Be Like Vodka: Distilling More Times For Better Quality. Proc 12th Int Conf on Learning Representations.

Chen CC, Tseng YM, Kang J, et al., 2023. Multi-lingual ESG issue identification. Proc 5th Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI for Financial Forecasting, p.111–115.

Chen ZY, Li SY, Smiley C, et al., 2022. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. Proc Conf on Empirical Methods in Natural Language Processing, p.6279–6292. https://doi.org/10.18653/v1/2022.emnlp-main.421

Chen ZY, Chen WH, Smiley C, et al., 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. Proc Conf on Empirical Methods in Natural Language Processing, p.3697–3711. https://doi.org/10.18653/v1/2021.emnlp-main.300

Chen ZY, Ma J, Zhang XL, et al., 2024. A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. *Transactions on Machine Learning Research*.

Costantino M, Coletti P, 2008. Information extraction in finance. Wit Press, Billerica, USA.

Daudert T, 2022. A multi-source entity-level sentiment corpus for the financial domain: the FinLin corpus. *Lang Resour and Eval*, 56(1):333–356. https://doi.org/10.1007/s10579-021-09555-3

De Prado ML, 2018. Advances in financial machine learning. John Wiley & Sons, Hoboken, USA.

Desai AP, Ravi T, Luqman M, et al., 2024. Opportunities and challenges of generative-AI in finance. Proc 2024 IEEE International Conference on Big Data (BigData), p.4913–4920.

Dow J, Gorton G, 1997. Stock market efficiency and economic efficiency: is there a connection? *J Finance*, 52(3):1087–1129. https://doi.org/10.1111/j.1540-6261.1997.tb02726.x

Duffie D, Pan J, 1997. An overview of value at risk. *J Deriv*, 4(3):7–49. https://doi.org/10.3905/jod.1997.407971

Dwork C, McSherry F, Nissim K, et al., 2006. Calibrating noise to sensitivity in private data analysis. Proc 3rd Theory of Cryptography Conf, p.265–284. https://doi.org/10.29012/jpc.v7i3.405

Feng DY, Dai YF, Huang JM, et al., 2023. Empowering Many, Biasing a Few: Generalist Credit Scoring through Large Language Models. *arXiv preprint arXiv:2310.00566*. https://doi.org/10.48550/arXiv.2310.00566

Galichin AV, Pautov M, Zhavoronkin A, et al., 2025. GLiRA: Closed-Box Membership Inference Attack via Knowledge Distillation. *IEEE Trans on Inf Forensics Secur*, 20:3893–3906.

Gu YX, Li D, Wei FR, et al., 2024. MiniLLM: Knowledge Distillation of Large Language Models. Proc 12th Int Conf on Learning Representations.

Guo C, Pleiss G, Sun Y, et al., 2017. On calibration of modern neural networks. Proc 34th Int Conf on machine learning, p.1321–1330.

Han PC, Shi XY, Huang JW, 2024. FedAL: Black-Box Federated Knowledge Distillation Enabled by Adversarial Learning. *IEEE J. Sel. Areas Commun.*, 42(11):3064–3077. https://doi.org/10.1109/JSAC.2024.3431516

Unedited

Han ZY, Gao C, Liu JY, et al., 2024. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *Trans Mach Learn Res*.

Hershey JR, Olsen PA, 2007. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. Proc IEEE Int Conf on Acoustics, Speech, and Signal Processing, p.317–320. https://doi.org/10.1109/ICASSP.2007.366913

Hristova D, Satani N, 2025. DiFiLE: A Knowledge-Distillation Longformer Model for Finance with Ensembling. Proc 58th Annual Hawaii Int Conf on System Sciences, p.1585–1594. https://hdl.handle.net/10125/109031

Huang AH, Wang H, Yang Y, 2023. FinBERT: A large language model for extracting information from financial text. *Contemp Account Res*, 40(2):806–841. https://doi.org/10.1111/1911-3846.12832

Jain S, Wallace BC, 2019. Attention is not not Explanation. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.3543–3556. https://doi.org/10.18653/v1/N19-1357

Jensen MC, 1968. The performance of mutual funds in the period 1945-1964. *J Finance*, 23(2):389–416. https://doi.org/10.2307/2325404

Ji ZW, Lee N, Frieske R, et al., 2023. Survey of hallucination in natural language generation. *ACM Comput Surv*, 55(12):248. https://doi.org/10.1145/3571730

Jørgensen R, Brandt O, Hartmann M, et al., 2023. MultiFin: A dataset for multilingual financial NLP. Proc Findings of the Association for Computational Linguistics, p.894–909. https://doi.org/10.18653/v1/2023.findings-eacl.66

Jorion P, 1996. Risk2: Measuring the risk in value at risk. *Financ Anal J*, 52(6):47–56. https://doi.org/10.2469/faj.v52.n6.2039

Kaur S, Smiley C, Gupta A, et al., 2023. REFinD: Relation extraction financial dataset. Proc 46th Int Conf on Research and Development in Information Retrieval, p.3054–3063.

Kim M, Lee S, Lee J, et al., 2023. Token-scaled logit distillation for ternary weight generative language models. Proc 37th Int Conf on Neural Information Processing Systems, p.42097–42118.

Kong YX, Nie YQ, Dong XW, et al., 2024. Large Language Models for Financial and Investment Management: Applications and Benchmarks. *J Portfolio Manage*, 51(2):162–210.

Lamm M, Chaganty AT, Manning CD, et al., 2018. Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts. Proc Conf on Empirical Methods in Natural Language Processing, p.82–92. https://doi.org/10.18653/v1/D18-1008

Lee J, Stevens N, Han SC, 2025. Large Language Models in Finance (FinLLMs). *Neural Comput Appl*, p.1–15. https://doi.org/10.1007/s00521-024-10495-6

Lei SY, Tao DC, 2023. A comprehensive survey of dataset distillation. *IEEE Trans on Pattern Anal Mach Intell*, 46(1):17–32. https://doi.org/10.1109/TPAMI.2023.3322540

Li JY, Tang TY, Zhao WX, et al., 2024. Pre-trained language models for text generation: A survey. *ACM Comput Surv*, 56(9):230. https://doi.org/10.1145/3649449

Li Z, Li YX, Zhao PH, et al., 2023. Is Synthetic Data From Diffusion Models Ready for Knowledge Distillation? *arXiv preprint arXiv:2305.12954*.

Li LJ, Dong PJ, Li AG, et al., 2023. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. Proc 37th Int Conf on Neural Information Processing Systems, 36:69490–69504.

Li YH, Wang SF, Ding H, et al., 2023. Large Language Models in Finance: A Survey. Proc 4th ACM International Conference on AI in Finance, p.374–382.

Li YH, Wang SF, Ding H, et al., 2023. Large Language Models in Finance: A Survey. Proc 4th ACM Int Conf on AI in Finance, p.374–382. https://doi.org/10.1145/3604237.3626869

Liang C, Zuo SM, Zhang QR, et al., 2023. Less is more: Task-aware layer-wise distillation for language model compression. Proc 40th Int Conf on Machine Learning, p.20852–20867.

Liebenwein L, Baykal C, Lang H, et al., 2020. Provable Filter Pruning for Efficient Neural Networks. Proc 8th Int Conf on Learning Representations.

Liu ZC, Oguz B, Zhao CS, et al., 2024. LLM-QAT: Data-Free Quantization Aware Training for Large Language Models. Proc Findings of the Association for Computational Linguistics, p.467–484. https://doi.org/10.18653/v1/2024.findings-acl.26

Liu XYY, Wang GX, Zha DC, 2023. FinGPT: Democratizing Internet-scale Data for Financial Large Language Models. *arXiv preprint arXiv:2307.10485*.

Loughran T, McDonald B, 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J Finance*, 66(1):35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Magdon-Ismail M, Atiya AF, 2004. Maximum drawdown. *Risk Mag*, 17(10):99–102.

Moreno-Sandoval A, Porta-Zamorano J, Carbajo-Coronado B, et al., 2020. The Financial Document Causality Detection Shared Task (FinCausal 2020). Proc 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, p.23–32.

Moreno-Ortiz A, Fernandez-Cruz J, Hernández CP, 2020. Design and Evaluation of SentiEcon: a fine-grained Economic/Financial Sentiment Lexicon from a Corpus of Business News. Proc 12th Language Resources and Evaluation Conf, p.5065–5072.

Unedited

Mukherjee R, Bohra A, Banerjee A, et al., 2022. ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. Proc Conf on Empirical Methods in Natural Language Processing, p.10893–10906. https://doi.org/10.18653/v1/2022.emnlp-main.748

Nguyen D, Gupta S, Do K, et al., 2022. Black-box few-shot knowledge distillation. Proc 17th European Conf on Computer Vision, p.196–211. https://doi.org/10.1007/978-3-031-19803-812

Nie YQ, Kong YX, Dong XW, et al., 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *arXiv preprint arXiv:2406.11903*.

Nie YQ, Kong YX, Dong XW, et al., 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *arXiv preprint arXiv:2406.11903*.

Pandya HA, Bhatt BS, 2021. Question Answering Survey: Directions, Challenges, Datasets, Evaluation Matrices. *arXiv preprint arXiv:2112.03572*.

Qin CW, Xia WH, Jiao FK, et al., 2023. Improving In-context Learning via Bidirectional Alignment. *arXiv preprint arXiv:2312.17055*. https://doi.org/10.48550/arXiv.2312.17055

Raza M, Jahangir Z, Riaz MB, et al., 2025. Industrial applications of large language models. *Sci Rep*, 15(1):13755. https://doi.org/10.1038/s41598-025-98483-1

Ribeiro MT, Singh S, Guestrin C, 2016. "Why should i trust you?" Explaining the predictions of any classifier. Proc 22nd Int Conf on Knowledge Discovery and Data Mining, p.1135–1144. https://doi.org/10.1145/2939672.2939778

Shah A, Vithani R, Gullapalli A, et al., 2023. FiNER-ORD: financial named entity recognition open research dataset. *arXiv preprint arXiv:2302.11157*.

Shah A, Paturi S, Chava S, 2023. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.6664–6679. https://doi.org/10.18653/v1/2023.acl-long.368

Sharma S, Khatuya S, Hegde M, et al., 2023. Financial numeric extreme labelling: A dataset and benchmarking. Proc Findings of the Association for Computational Linguistics, p.3550–3561. https://doi.org/10.18653/v1/2023.findings-acl.219

Sharma S, Nayak T, Bose A, et al., 2022. FinRED: A dataset for relation extraction in financial domain. Proc 31st Companion of the Web Conf, p.595–597.

Singh S, 2018. Natural Language Processing for Information Extraction. *arXiv preprint arXiv:1807.02383*.

Sinha A, Khandait T, 2021. Impact of news on the commodity market: Dataset and results. In: Arai K (Ed.), Advances in Information and Communication, p.589–601.https://doi.org/10.1007/978-3-030-73103-841

Sinha A, Kedas S, Kumar R, et al., 2022. SEntFiN 1.0: Entity-aware sentiment analysis for financial news. *J Assoc for Inf Sci Technol*, 73(9):1314–1335. https://doi.org/10.1002/asi.24634

Soun YY, Yoo J, Cho MY, et al., 2022. Accurate stock movement prediction with self-supervised learning from sparse noisy tweets. Proc Int Conf on Big Data, p.1691–1700. https://doi.org/10.1109/BigData55660.2022.10020720

Sundaram JPS, Du W, Zhao Z, 2019. A survey on LoRa networking: Research problems, current solutions, and open issues. *IEEE Commun Surv Tutorials*, 22(1):371–388. https://doi.org/10.1109/COMST.2019.2949598

Sy E, Peng TC, Huang SH, et al., 2023. Fine-grained argument understanding with bert ensemble techniques: A deep dive into financial sentiment analysis. Proc 35th Conf on Computational Linguistics and Speech Processing, p.242–249.

Tang YX, Liu ZJ, 2024. A Distributed Knowledge Distillation Framework for Financial Fraud Detection Based on Transformer. *IEEE Access*, 12:62899–62911. https://doi.org/10.1109/ACCESS.2024.3387841

Timiryasov I, Tastet JL, 2023. Baby Llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. *arXiv preprint arXiv:2308.02019*.

Touvron H, Martin L, Stone K, et al., 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Van Erven T, Harremos P, 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans Inf Theory*, 60(7):3797–3820. https://doi.org/10.1109/TIT.2014.2320500

Varmedja D, Karanovic M, Sladojevic S, et al., 2019. Credit card fraud detection-machine learning methods. Proc 18th Int Symp Infoteh-Jahorina, p.1–5. https://doi.org/10.1109/INFOTEH.2019.8717766

Wan FQ, Huang XT, Cai D, et al., 2024. Knowledge Fusion of Large Language Models. Proc 12th Int Conf on Learning Representations.

Wang Z, 2021. Zero-shot knowledge distillation from a decision-based black-box model. Proc 38th Int Conf on Machine Learning, p.10675–10685.

Wen YQ, Li ZC, Du WY, et al., 2023. f-Divergence Minimization for Sequence-Level Knowledge Distillation. Proc 61st Annual Meeting of the Association for Computational, p.10817–10834. https://doi.org/10.18653/v1/2023.acl-long.605

Wu HZ, Zhang W, Shen WW, et al., 2018. Hybrid deep sequential modeling for social text-driven stock prediction. Proc 27th Int Conf on information and knowledge management, p.1627–1630. https://doi.org/10.1145/3269206.3269290

Unedited

Wu SJ, Irsoy O, Lu S, et al., 2023. BloombergGPT: A Large Language Model for Finance. *arXiv preprint arXiv:2303.17564*.

Xie QQ, Han WG, Chen ZY, et al., 2024. FinBen: A Holistic Financial Benchmark for Large Language Models. Proc 38th Int Conf on Neural Information Processing Systems, article 3033.

Xie QQ, Han WG, Zhang X, et al., 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. Proc 37th Int Conf on Neural Information Processing Systems, 36:33469–33484.

Xu XH, Li M, Tao CY, et al., 2024. A Survey on Knowledge Distillation of Large Language Models. *arXiv preprint arXiv:2402.13116*.

Xu YM, Cohen SB, 2018. Stock movement prediction from tweets and historical prices. Proc 56th Annual Meeting of the Association for Computational Linguistics, p.1970–1979. https://doi.org/10.18653/v1/P18-1183

Yang CP, Zhu Y, Lu W, et al., 2024. Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application. *ACM Trans Intell Syst Technol*, in press. https://doi.org/10.1145/3699518

Yang HY, Liu XY, Wang CD, 2023. FinGPT: Open-Source Financial Large Language Models. *FinLLM Symposium at IJCAI 2023*.

Yang LY, Kenny EM, Ng TLJ, et al., 2020. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. Proc 28th Int Conf on Computational Linguistics, p.6150–6160. https://doi.org/10.18653/v1/2020.coling-main.541

Yang Y, Tang YX, Tam KYK, 2023. InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning. *arXiv preprint arXiv:2309.13064*.

Zhang XY, Yang Q, 2023. XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters. Proc 32nd Int Conf on Information and Knowledge Management, p.4435–4439. https://doi.org/10.1145/3583780.3615285

Zhao HQ, Liu ZL, Wu ZH, et al., 2024. Revolutionizing Finance with LLMs: An Overview of Applications and Insights. *arXiv preprint arXiv:2401.11641*. https://doi.org/10.48550/arXiv.2401.11641

Zhao YX, Yu B, Hui BY, et al., 2024. Tree-Instruct: A Preliminary Study of the Intrinsic Relationship between Complexity and Alignment. Proc Joint Int Conf on Computational Linguistics, Language Resources and Evaluation, p.16776–16789.

Zhao QY, Zhu BH, 2024. Towards the Fundamental Limits of Knowledge Transfer over Finite Domains. Proc 12th Int Conf on Learning Representations.

Zhao ZH, Fan WQ, Li JT, et al., 2024. Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Trans Knowl Data Eng*, 36(11):6889–6907. https://doi.org/10.1109/TKDE.2024.3392335

Zhou ZH, Ma LQ, Liu H, 2021. Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading. Proc Findings of the Association for Computational Linguistics, p.2114–2124. https://doi.org/10.18653/v1/2021.findings-acl.186

Zhu FB, Lei WQ, Huang YC, et al., 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. Proc 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing, p.3277–3287. https://doi.org/10.18653/v1/2021.acl-long.254

Unedited