Jiaqi SHI, Xulong ZHANG, Xiaoyang QU, Junfei XIE, Jianzong WANG, 2025. Knowledge distillation for financial large language models: a systematic review of strategies, applications, and evaluation. *Frontiers of Information Technology & Electronic Engineering*, 26(10):1793-1808. https://doi.org/10.1631/FITEE.2500282

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

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

Corresponding author: Jianzong WANG
E-mail: jzwang@188.com
ORCID: https://orcid.org/0000-0002-9237-4231

Motivation

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: strategies, applications, and evaluation.

Main idea

- 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 field.
- 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.

Overview of distillation strategies

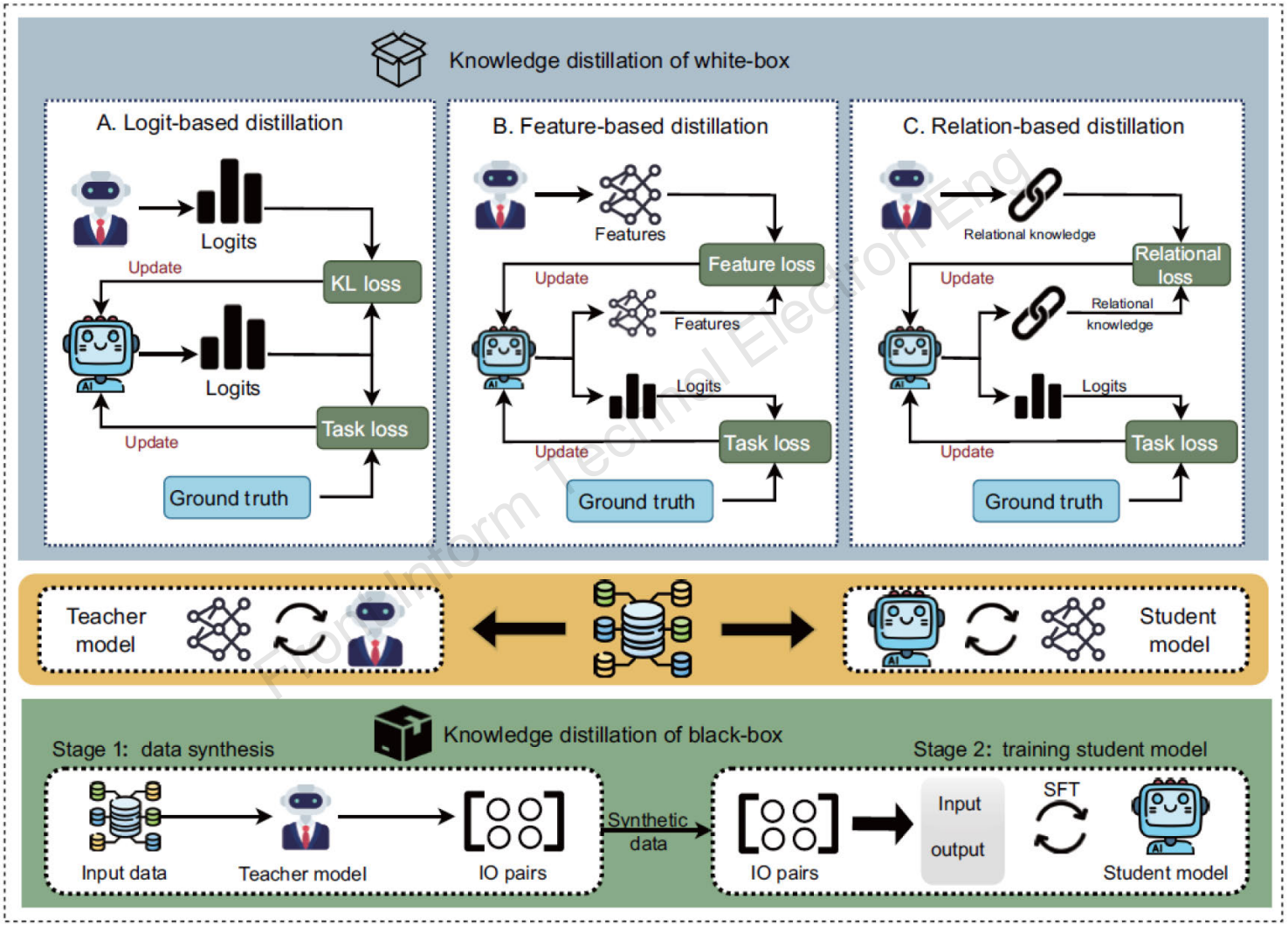


Fig. 5 Overview of the distillation strategies. IO: input/output; SFT: supervised fine-tuning

Three-tier framework

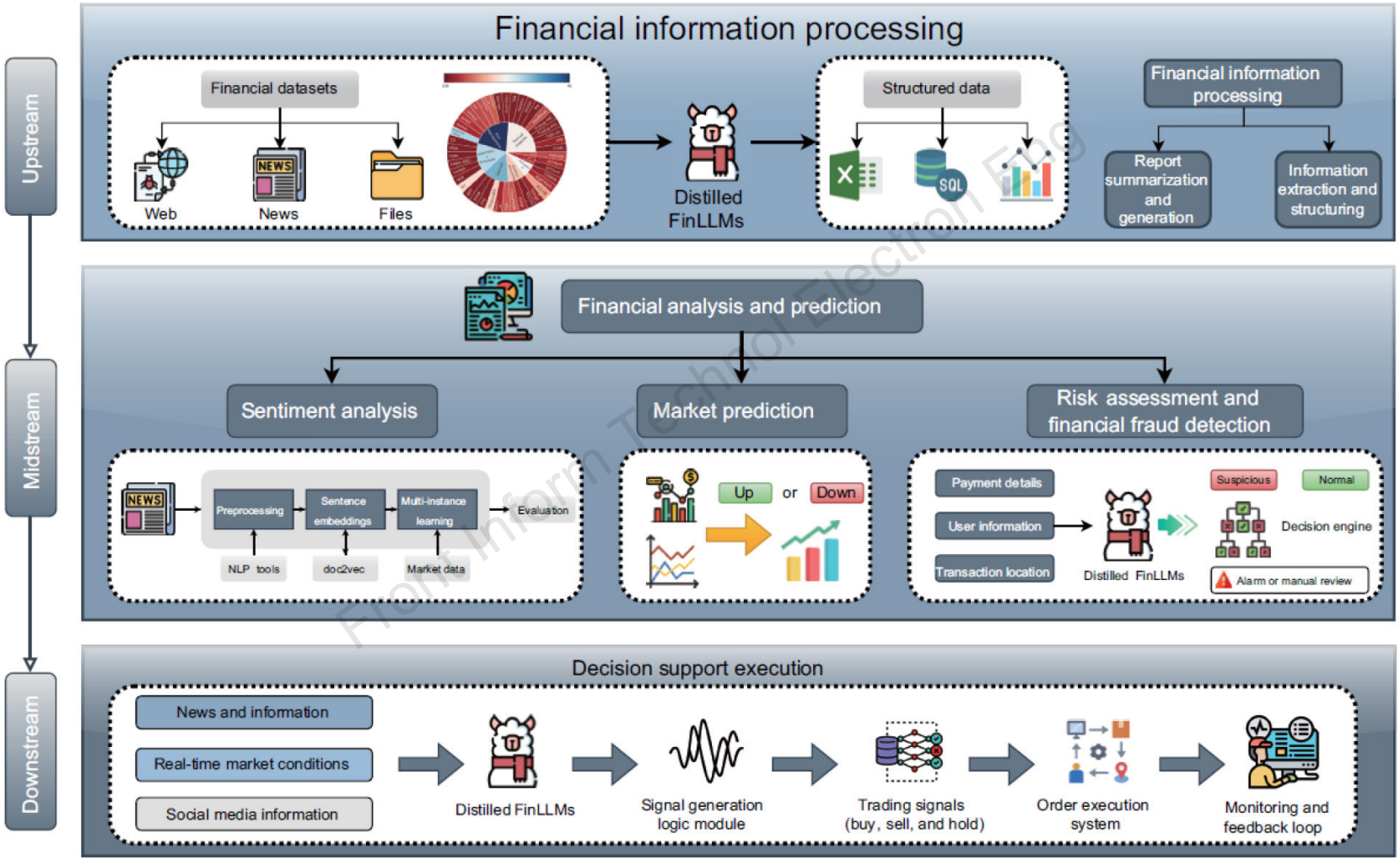


Fig. 6 Three-tier framework: upstream–midstream–downstream

Summary of existing FinLLMs

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

Dataset	Task	Description	Metric
FiNER-ORD (Shah A et al., 2023a)	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
Headlines (Sinha and Khandait, 2021)	TA	News headline categorization	F1
FOMC (Shah et al., 2023b)	TA	Hawkish–dovish classification	F1, Acc
FinArg-ACC (Sy et al., 2023)	TA	Argument component categorization	F1, Acc
MultiFin (Jørgensen et al., 2023)	TA	Multi-classclassification	F1, Acc
M&A (Yang LY et al., 2020)	TA	Deal completeness classification	F1, Acc
MLESG (Chen CC et al., 2023)	TA	ESG issue identification	F1, Acc
FinQA (Chen ZY 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 ZY 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 YM and Cohen, 2018)	FO	Stock trend forecasting	Acc, MCC
CIKM18 (Wu HZ et al., 2018)	FO	Stock trend forecasting	Acc, MCC

PPV: positive predictive value; Acc: accuracy; ESG: environmental, social, and governance; PER: person; LOC: location; ORG: organization; EM: exact match; MCC: Matthews correlation coefficient

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.

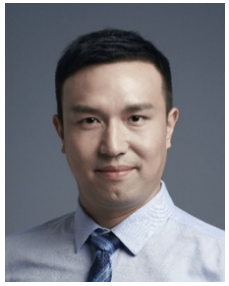
At the strategy level, this review introduces a structured taxonomy to comparatively analyze existing distillation pathways. At the application level, this review puts forward a logical upstream–midstream–downstream framework to systematically explain the practical value of distilled models in the financial 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.

Jiaqi SHI is an Algorithm Engineer at Ping An Technology (Shenzhen) Co., Ltd. She is pursuing the MS degree at the University of Science and Technology of China. Her main research interests include reinforcement learning, edge computing, and multi-modal large models.

Xulong ZHANG is a Senior Algorithm Researcher at Ping An Technology (Shenzhen) Co., Ltd. He holds a PhD degree in Computer Science from Fudan University and was selected as a Shanghai Oriental Scholar (Young Talent Program) in 2023. His main research interests include large models, embodied AI, cross-modal intelligent computing, and model context protocol.

Xiaoyang QU is the Group Lead for Advanced Machine Learning Algorithms at Ping An Technology (Shenzhen) Co., Ltd. He was a visiting scholar at the University of Central Florida and holds a PhD degree from Huazhong University of Science and Technology. He has published nearly 50 articles in international top conferences and journals in both system architecture (e.g., INFOCOM, DAC, TPDS, and IPDPS) and AI (e.g., NeurIPS, IJCAI, ICASSP, and Interspeech), including one paper nominated for a Best Student Paper Award.

Junfei XIE is an intern at Ping An Technology (Shenzhen) Co., Ltd., and a master's candidate at the University of Science and Technology of China. His main research interests include large language models, large vision-language models, visual question answering, and hallucinations.



Jianzong WANG is a Senior Engineer and the Dean of the Institute of Advanced Intelligent Finance Technology at Ping An Technology (Shenzhen) Co., Ltd. He completed his postdoctoral research in AI at the University of Florida (USA) and earned his PhD degree through a joint program between Rice University (USA) and Huazhong University of Science and Technology (China). He is a Distinguished Member of the China Computer Federation (CCF), a member of the CCF Big Data Expert Committee, and a Vice Director of the Technical Committee on Federated Data and Federated Intelligence of the Chinese Association of Automation (CAA). His main research interests include embodied AI, large models, federated learning, and deep learning.