Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28450

Review on the Practice and Construction of Financial AI Large Models

Siyang Chen

Shanghai New Channel A-level International Center, Shanghai, China
Benchen1346@163.com

Abstract. The financial industry is a pillar industry of a country. Acting as an intermediary, finance provides direct or indirect capital services for various industries to meet diverse needs such as asset management, liability management, payment and settlement, and financial transaction processing. Compared with the digitalization of other industries, fintech is also characterized by massive transactions and a high degree of digitalization, giving AI numerous application scenarios and promising prospects in financial contexts. This paper investigates the practice of training large financial models based on general-purpose large models and industry data, their construction process, as well as their application cases and performance evaluation. It also analyzes and prospects the future scenarios and challenges of model development. It can be concluded that through training and fine-tuning with domain-specific financial data, large financial models can effectively enhance professional processing capabilities—for example, significantly improving the efficiency of test case generation and strengthening risk control capabilities. In the future, these models will develop towards lightweight architectures, with compressed parameter scales to meet the computational needs of small and medium-sized financial institutions, and intelligent agents will become an important application format. However, challenges such as data sharing difficulties, insufficient real-time model updates, and limited algorithm transparency still exist, requiring further exploration of secure data sharing mechanisms and dynamic model update technologies.

Keywords: General-Purpose Large Models, Industry-Specific Large Models, Intelligent Agents, Model Training

1. Introduction

In the 2020s, artificial intelligence technology has been widely used in various industries. High computing power and ease of use have made artificial intelligence an efficient productivity tool. AI large models developed based on artificial intelligence have also entered the public eye, with general-purpose large models are the most popular.

General-purpose large models refer to large-scale artificial intelligence models trained on massive data and capable of handling a wide range of tasks. These models are usually based on deep learning technology; by learning general representations through massive parameters and cross-domain data, they can adapt to a variety of downstream tasks without the need to train from scratch

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ28450

for each task. Different general-purpose large models have distinct characteristics: for example, OpenAI's GPT-4 model is known for its strong reasoning ability and low hallucination rate; Gemini 1.5 developed by Google has a huge number of parameters and excellent performance in the field of multimodal applications; and DeepSeek-MoE from DeepSeek is a commercially available open-source model, making it an extremely economical choice for small and medium-sized enterprises.

Although general-purpose large models have many advantages such as high versatility and wide task adaptability, when used in specific industries or fields, they can cover basic knowledge and common sense but may make mistakes in analyzing industry terms or professional logic, and are prone to generating AI hallucinations in practical industry applications. In contrast, industry-specific large models are more professional and accurate in vertical industry data analysis and retrieval. At the same time, general-purpose large models have extremely high deployment costs: handling the GPT-3.5 large model with 180 billion parameters requires up to 20,000 GPU chips, and the number of GPU chips required for the commercialization of future GPT large models may even exceed 30,000 [1]—a requirement that is almost unbearable for industries and enterprises that require private deployment of large models. Although a large number of lightweight solutions for general-purpose large models have emerged nowadays [2], the industry accuracy of vertical-domain large models is still incomparable to that of general-purpose large models.

Table 1. Comparison of training scenarios between general-purpose large models and industry-specific large models

Dimension	General-Purpose Large Models	Industry-Specific Large Models
Training Data	Massive cross-industry data	Vertical domain data (e.g., finance, medical care)
Parameter Scale	Over 10 billion	About 1 billion, with greater focus on industry characteristics
Cost	Approximately $100 million per training	Ranging from 25,000 to 300,000 depending on the scale
Typical Representatives	GPT-4, Gemini	BloombergGPT (financial industry), BioGPT (medical industry)

Table 1 presents the core differences between general-purpose large models and industry-specific large models in terms of training data, parameter scale, cost, and typical representatives, intuitively demonstrating the essential differences between the two in applicable scenarios and resource investment.

The financial industry is a pillar industry of a country. As an intermediary, finance provides direct or indirect capital services for various industries to meet needs such as assets, liabilities, payment and settlement, and financial transactions. Compared with the informatization of other industries, fintech features massive transactions and a high degree of digitalization, which enables AI to have broad application scenarios and promising prospects in financial contexts. Therefore, this paper will focus on discussing AI large models in the financial industry. Given the professionalism and application scenarios of the financial industry, the mainstream approaches to applying large models in the financial industry involve building financial large models based on general-purpose large models and further fine-tuning financial vertical application models based on financial large models. Table 2 presents the cooperation cases between many banks and mainstream general-purpose large models, illustrating the actual implementation paths, parameter scales, and open status of financial large models, and demonstrating the industry's demand for and application status of AI large models.

DOI: 10.54254/2755-2721/2026.TJ28450

Table 2. Cooperation between multiple banks and general-purpose large models in developing financial large models

Model Name	Developer	Cooperation Partners	Parameter Scale	Status
ERNIE Bot	Baidu	Agricultural Bank of China, SPD Bank	260B	Non-open source, commercial use allowed after application
Qwen	Alibaba	China Construction Bank, China Merchants Bank	7B, 14B, 72B, 100B	7B, 14B, 72B versions: open-source and free for commercial use; 100B version: non-open source
DeepSeek	DeepSeek	Industrial and Commercial Bank of China, Bank of China	100B	Non-open source, only available for enterprise cooperation
LLaMA 2	Meta	JP Morgan Chase, Bank of America	7B, 13B, 70B	Fully open-source and supports commercial use
BLOOM	BigScience	European Central Bank, Bank for International Settlements	176B	Fully open-source and supports commercial use

2. Construction process of financial AI large models

By analyzing the practices of industry models released by many banks, the main process of building financial AI large models is summarized as follows:

1. Select a general-purpose large model with high maturity and moderate computing power requirements as the foundation for model training;
2. Prepare domain-specific data in specific financial scenarios and complete tasks including data cleaning, preprocessing, and annotation;
3. Optimize the model training set, including training and reinforcement learning based on feedback from domain professionals;
4. Deploy and execute inference for the financial industry AI model;
5. Evaluate the financial AI model.

2.1. Preparation of general-purpose large models and basic environment for financial large models

To build a financial large model, first, a general-purpose large model must be selected as the training foundation. Since the LLaMA2 large model has low deployment costs and open-source code, the following will take LLaMA2-7B as an example for a detailed discussion. The basic environment configuration requirements for training a financial large model are relatively low; the recommended configuration is 8 NVIDIA A100 80GB or 4 NVIDIA H100 80GB (needing to support BP16 (brain floating point) or FP16 (floating point)), and the memory size should be more than 512GB.

2.2. Preparation of datasets and codes

RLHF (Reinforcement Learning from Human Feedback) is a special type of training approach, whose goal is to make the responses of artificial intelligence to customers more in line with human values. The RLHF dataset is divided into two parts: the first part is used to train the reward model, with a size ranging from 100KB to 1MB, mainly containing comparison data scored by professionals to make the answers of artificial intelligence more human-like; the second part is used

DOI: 10.54254/2755-2721/2026.TJ28450

for reinforcement learning, with a size of approximately 10KB to 100KB, including the prompts obtained when users ask questions.

After preparing the dataset, engineers also need to prepare code for pre-training, instruction fine-tuning, and model inference—including Chinese-LLaMA-Alpaca-3. This is a project for training Chinese vertical large models based on the LLaMA2 large model; these models expand and optimize the Chinese vocabulary on the basis of the original LLaMA2, and use large-scale Chinese data for incremental pre-training to further improve basic Chinese semantic understanding and instruction-following capabilities. After the dataset is expanded, the large model can use a tokenizer to tokenize the dataset; the core principle is to split text segments into tokens sequentially and gradually map the tokens to corresponding integer vectors.

2.3. Model training

The process includes secondary pre-training, supervised instruction fine-tuning, reward model training, and reinforcement learning training based on human feedback.

Secondary Pre-Training: This process will use the Transformer algorithm and train on a large amount of data based on the basic large model. After the training with basic data is completed, the Low-Rank Adaptation (LoRA) technology can be used—LoRA is a technology for efficiently fine-tuning large-scale pre-trained models, proposed by Microsoft's research team in 2021. It optimizes the model fine-tuning process by introducing low-rank matrices, which significantly reduces the demand for computing resources and storage while maintaining model performance, and is suitable for quickly fine-tuning large models in environments with limited budgets and tight schedules.

Supervised Instruction Fine-Tuning: This can better exert the model's industry capabilities in specific fields and also make it well-suited for scenarios such as industry Q&A. Unlike pre-training, supervised instruction fine-tuning does not require a large amount of data, but the quality of the data must be guaranteed—therefore, data manually screened and written by professionals is a better choice. When performing supervised instruction fine-tuning, OpenAI will incorporate the experience feedback of real users into the training process of the large model as task descriptions. After the fine-tuning data is prepared, the data can be modeled, and a Supervised Fine Tuning (SFT) model can be output after this stage is completed.

Reward Model: The SFT model can already understand user input and answer questions to a large extent, but the current model output and responses are still uncontrollable. Therefore, engineers need to use RLHF technology again to improve the model's response governance capabilities. At this stage, engineers need to create multiple completion versions of the SFT, manually score the answers, and add the scores to the original corpus to supervise the generated content of the Transformer model. After the scoring is completed, the various completion versions can be ranked in ascending order of human-likeness, and the sub-model with the highest score is retained.

Reinforcement Learning Training based on Human Feedback: The purpose of reinforcement learning is to train an intelligent agent to complete tasks within an unknown environment. The agent will learn by itself in a complex environment, constantly adapt, and provide feedback to the environment. The agent consists of a policy and a Reinforcement Learning Algorithm (RLA): the policy is an action taken based on the observation of the environment, while the RLA is an AI algorithm that continuously updates the policy based on actions, observation results, and reward data, and its goal is to find the optimal policy in the environment.

DOI: 10.54254/2755-2721/2026.TJ28450

2.4. Model inference

After the model training is completed, an intelligent agent for the financial field is developed [3]. Consider the financial testing intelligent agent of a fintech company as an example:

This company uses a financial test case generation intelligent agent based on the Tongyi Qianwen large model to improve software testing efficiency. The completeness and importance of software testing serve as important criteria for measuring test quality; traditional manual test requirement analysis and case design are time-consuming, labor-intensive, and prone to errors. For standard products such as the core banking system, this company uses the bank core system requirement assets from historical projects and the test case assets generated based on the requirements as training data. The company conducts model training on the basis of the Tongyi Qianwen large model to form a financial testing AI model. The testing intelligent agent automatically extracts requirement items and key elements from the requirement documents, and generates positive/negative test cases by leveraging the case asset library. The practice of a certain bank demonstrates that this technology reduces the time for writing a single case from 4 hours to 15 minutes, increases the coverage rate of abnormal scenarios to 98%, effectively promotes the transformation of testing work from manual labor to intelligent analysis, and allocates 70% of human resources to high-value strategy optimization activities.

2.5. Evaluation of financial AI models

To assess the correctness and compliance of financial AI models in financial scenarios, as well as their compatibility with different financial field scenarios, the Shanghai Artificial Intelligence Industry Association and the Shanghai Financial Industry Federation jointly released China's first group standard focusing on financial business capabilities—Guidelines for the Evaluation of Financial Large Model Applications.

Table 3. Evaluation dimensions and indicators of financial AI models

Dimension	Key Indicators	Application Scenario Examples
Basic Model Capabilities	Single-modal: Text classification, information extraction (11 items);Multi-modal: Image-text retrieval, video Q&A (6 items)	Research report chart analysis, financial report data extraction
Financial Security & Value Alignment	Compliance (GDPR/ Cybersecurity Regulations), cultural values, ethical norms (9 items)	Avoiding discriminatory credit suggestions, privacy protection
Financial Risk Control Capabilities	19 indicators such as interest rate/exchange rate/credit/public opinion risks, dynamic adjustment of market fluctuation responses	Stock price anomaly early warning, real-time update of default probability
Financial Professional Cognition Capabilities	23 items in accounting (financial statement analysis), finance (risk assessment), and economics; 6 types of knowledge such as IPO chart interpretation	Financial statement analysis, investment portfolio optimization
Financial Business Auxiliary Expansion Capabilities	117 indicators in 5 fields of banking/fund/insurance/securities/trust (28 scenarios including deposit and loan, quantitative trading, claim assessment, etc.)	Intelligent investment advisor response speed improvement, 20% reduction in complaint rate

DOI: 10.54254/2755-2721/2026.TJ28450

3. Challenges and future outlook of financial AI large models

Financial AI large models still face numerous challenges in the industry:

1. Financial data is highly sensitive, which limits cross-institutional data sharing and makes it difficult to expand the model training set. In addition, data collection and use may violate the principle of informed consent; even if the data is anonymized, the model may restore individual identities through correlation analysis, leading to privacy violations and legal disputes [4].

2. The financial market environment is complex and changeable, with data and business needs constantly updated. If the model cannot be updated and optimized in a timely manner, it is prone to obsolescence, resulting in diminished performance and an inability to accurately respond to new market conditions.

3. The algorithms of large models are complex and exhibit a "black box" characteristic; their internal logic is opaque, making it difficult to trace the source of risks in a timely manner, thereby reducing the transparency of high-risk decisions and potentially triggering a trust crisis between regulatory authorities and investors.

4. The financial industry is subject to strict supervision, and the application of financial AI large models needs to comply with relevant laws, regulations, and policy requirements. However, due to the complexity and innovation of large models, it is difficult for regulatory authorities to formulate clear regulatory standards and norms, and financial institutions also face significant compliance pressure [5].

In the future, AI-based financial large models will inevitably become lightweight, with parameter scales compressed from trillions to tens of billions to meet the computing power-cost balance of small and medium-sized financial enterprises. At the same time, intelligent agents will emerge as a key application form for implementing large models, evolving from simple chatbots to automated task executors, thereby enabling deeper human-machine collaboration in financial business.

4. Conclusion

As a key development direction in Fintech, financial AI large models have demonstrated significant application value in sectors such as banking and securities. This research indicates that by training and fine-tuning general-purpose large models with financial domain data, financial large models can enhance their professional processing capabilities in financial scenarios—such as significantly improving test case generation efficiency and strengthening risk control capabilities. In the future, financial large models will develop towards the direction of lightweight, with the parameter scale gradually compressed to meet the computing power needs of small and medium-sized financial institutions. At the same time, intelligent agents will become an important application form, transforming from conversational interaction to task execution and realizing in-depth human-machine collaboration.

However, current research has limitations: the sensitivity of financial data impedes data sharing and limits the scale of training datasets; the dynamic financial environment hinders the development of real-time model update mechanisms; the "black box" nature of algorithms reduces decision transparency, and compliance risks require further investigation. Future research can explore more secure data sharing mechanisms, develop dynamic model update technologies, and improve model interpretability to better adapt to the regulatory requirements and business needs of the financial industry.

DOI: 10.54254/2755-2721/2026.TJ28450

References

[1] Lin, J. (2024). AIGC reshapes finance: AI large model-driven financial transformation and practice.Machinery Industry Press.
[2] Shi, H., Liu, T., & Tang, F. (2025). Research and application of lightweight technology based on general visual large models. Guangxi Communication Technology, (2), 49–52.
[3] Wang, J., Yang, Z., & Yang, B. (2025). Exploration of the application of intelligent agents in the banking industry. Financial Review, (4), 69–74.
[4] Xiao, H. (2025, July). Research report on the application of financial large models and AI in financial business [Report].
[5] Li, P. (2025, August 19). How can the financial industry "dance" with large models. Financial Times, p. 005.