World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Big language modeling in financial markets: present and future: One study overview

Ke Deng *

School of Finance, Central University of Finance and Economics，China.

World Journal of Advanced Research and Reviews, 2025, 25(03), 635-644

Publication history: Received on 30 January 2025; revised on 05 March 2025; accepted on 07 March 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.3.0754

Abstract

With the breakthrough of large language models (LLMs) represented by GPT and BERT in the field of natural language processing, they provide a new paradigm for financial prediction and analysis by processing unstructured textual data (e.g., news, financial reports, and social media opinions). Studies have shown that LLMs have demonstrated significant advantages in areas such as customer service, risk management, and investment decision. Some models (e.g., BloombergGPT) have achieved a balance between specialized capabilities and general performance through hybrid data training and vertical fine-tuning, which provides a paradigm for the development of large language models in current and future verticals. However, LLMs still face challenges such as insufficient understanding of specialized terminology, data privacy risk, lack of model interpretability and high-frequency transaction latency in financial scenarios. Through the research review, this paper further proposes to deal with the above problems through the technical paths of vertical model development, federated learning, and interpretability enhancement, and emphasizes the need for continuous exploration in the future in the areas of large model multimodalization , regulatory synergy , and financial inclusion deepening, in order to promote the scaled landing and value release of LLMs in the financial domain. By analyzing and discussing the history of LLMs, its core technology, current applications and problems in the financial industry, this paper shows that big language models have a broad development potential in the financial field, and are one of the core driving forces for innovation of the future the industry.

Keywords: Big Language Modeling; Fintech; Current Development; Future Trends

1. Introduction

In recent years, with the rapid development of artificial intelligence technology, LLMs such as GPT and BERT have made great breakthroughs in the field of Natural Language Processing (NLP), breaking the shackles of traditional language models based on Recurrent Neural Network (RNN) that are difficult to carry out parallel operations and capture long-term dependencies. These models have been widely used in tasks such as text classification, sentiment analysis, and machine translation by demonstrating powerful language understanding and generation capabilities through training on massive text data. As an important core of economic activities, the prediction and analysis of financial market has been a research hotspot in both academia and industry. Traditional financial forecasting methods mainly rely on structured data such as historical price data and financial statements, however, these methods often face the problems of sparse data, high model complexity and limited prediction accuracy. In contrast, big language models can effectively handle unstructured text data such as news reports, social media comments, company announcements, etc., providing a new perspective and tools for financial market prediction and analysis. However, the application and promotion of big language models in financial markets also face some challenges and problems, such as the privacy protection problem of data reading, the under-interpretability of models, and the high cost of pre-training and operation. The purpose of this paper is to review the history of the development of Big Language Models, explore their advantages over traditional methods, summarize the current development of Big Language Models, and introduce the core technology and core logic of their construction in order to provide financial practitioners and researchers with practical knowledge in this

* Corresponding author: Ke Deng

direction, so as to further promote the in-depth development of Big Language Models in the financial field. In addition, this paper will summarize the current applications of Big Language Models in the financial market, focusing on the latest research in the direction of stock market prediction effectiveness to explore the feasibility and prospect of Big Language Models in financial market prediction. In conclusion, Big Language Modeling brings new opportunities and challenges for the financial industry to transform into digital intelligence and build new core competitiveness, and this paper aims to provide theoretical and practical references for this purpose.

2. Trajectory of the development of a large language model

LLMs are a kind of deep learning models in the field of natural language processing , which usually have tens of billions or even more parameters at this stage (e.g., the parameters of ChatGPT-3 can be up to 175 billion parameters, and the parameters of the full-blooded version of deepseek are 671 billion parameters), and the size of the parameters represents the complexity of the model, which is usually linked to the understanding and learning ability of the model. The emergence of big language models aims at understanding and generating natural language and solving real-world problems by learning massive textual data. By reviewing the history of big language models, this paper aims to help readers have a more comprehensive understanding of big language models.

2.1. Early budding (before 2017)

2.1.1. Models based on rules and statistical methods

Early NLP tasks mainly relied on rule systems and simple statistical methods, which processed language through predefined rules or statistical probabilities, because of this, its dependence on experts is very high, both in the design of rules, and in the predefinition of the experts need to be carefully designed. In addition, it has limited processing ability to face complex tasks, lacks memory, and is difficult to capture semantics and understand the logic of the context, for example, the early n-gram model obtained statistical word sequences by counting the frequency of occurrence of text words, and then further predicted the probability of the word sequences by counting the probability of the word sequences, which is destined to have a limited precision of the results obtained from this method of processing, and the mechanism determines that it is not able to deal with the long-distance dependency relationship, and With the increase of n, the complexity of the model rises sharply, thus the running cost increases, even the hardware technology at that time could not support the calculation of too large n-value samples.

2.1.2. Introduction of neural networks

In the 1980s, Yann LeCun and others pioneered convolutional neural networks (CNNs) based on the biological visual system, and actively applied them to image processing tasks. The introduction of this technology greatly promoted the progress of machine learning technology in the direction of image recognition and processing, and laid the foundation for machine learning to move towards multimodality, and in 1998, the proposal of LeNet-5 also laid a solid foundation for the later more complex architecture.

Despite the maturing of neural networks in the image domain, their application in NLP is still in its early stages, mainly simple Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs).RNNs and LSTMs are capable of handling sequential data but are difficult to capture long distance dependencies due to the problem of vanishing gradients and are less efficient to train[1] .

2.1.3. word embedding technology

In 2013, the proposal of the Word2Vec model was an important node in the history of the development of a large language model.Word2Vec utilizes a processing method that maps words into a continuous vector space to capture the semantic relationships between words. Subsequently, the introduction of models such as GloVe and FastText further optimized the word embedding technique and provided fundamental technical support for the construction of subsequent more complex language models.

2.2. Transformer era (2017-present)

2.2.1. Breakthroughs in Transformer Architecture (2017)

In 2017, a team of Google researchers consisting of Ashish Vaswani and others disruptively proposed the Transformer architecture, which provides the basic framework paradigm for the latest big language modeling hitch from 2017 to the present day, and greatly advances the development of big language models.The core idea of the Transformer architecture is to use an attention mechanism to weight the input sequences of different The core idea of the

Transformer architecture is to weight different parts of the input sequence through an attention mechanism to capture the global contextual relationships. The architecture completely abandons the structure of RNN, innovatively uses Attention Mechanism to process sequence data, solves the long-range dependency problem, and realizes efficient parallel computation, eliminating the shortcomings of traditional modeling approaches, revolutionizing NLP and other fields[3] .

2.2.2. Birth of BERT and GPT (2018-2019)

- ①BERT (Bidirectional Encoder Representations from Transformer): In 2018, Google released the BERT model, which builds a model based on the encoder portion of Transformer and uses bidirectional processing.BERT uses the Masked Language Model (MLM) and the Next Sentence prediction task for pre-training, which enables better understanding of context and is suitable for tasks such as reading comprehension and Q&A.
- ②GPT (Generative Pre-trained Transformer):In the same year, OpenAI released the first generation GPT model (ChatGPT-1), which also builds a model based on the decoder part of the Transformer, adopts a unidirectional generative approach, and through pre-training and fine-tuning, it demonstrated to the world at that time its potential for tasks such as text generation and translation.

2.2.3. Model scale-up (2020-present)

- ① GPT-3 (2020):In 2020, OpenAI released GPT-3, which, with 175 billion parameters, was one of the largest language models at the time. It demonstrated zero-sample (Zero-Shot) and few-sample (Few-Shot) learning capabilities, i.e., it can complete multiple tasks without task-specific training, has good generalization capabilities, and shows notable capabilities in translation, code generation and Q&A.
- ② GPT-4 and its successor (2023-present):GPT-4 is based on the concept of Scaling Law proposed by its development company OpenAI, i.e., there is a power law relationship between the performance of an AI model and the number of model parameters, dataset size, and computational resources, and the performance of an AI model can be further improved by increasing the number of model parameters, dataset size, and computational resources in exchange. further improvement. Although its specific parameters are not disclosed, it is speculated that its parameters exceed 1 trillion. As a result, GPT-4 performs particularly well on multimodal tasks (combining text and images), further improving the model's generality and generalization ability.
- ③ Deepseek-V3 as well as Deepseek-R1 (2025):Deepseek-V3 and Deepseek-R1 are next-generation language models developed by Deepseek, a company incubated by Chinese quantitative private equity giant Phantom Square Quantitative. Among them, V3 lays the basic architecture, while R1 builds on it with further inference optimization and engineering innovation.The R1 model replaces traditional supervised fine-tuning (SFT) through reinforcement learning (RL), and uses a combination of hybrid model of experts (MoE) architecture and FP8 mixed-accuracy training instead of the previous big model (ChatGPT) relying on large-scale parameter stacking and traditional training methods. In addition, it also adopts hybrid architecture (Transformer+Knowledge Graph) and dynamic retrieval enhancement, which reduces the video memory consumption through MLA low-rank attention mechanism as well as knowledge distillation and data compression to replace the pure Transformer decoder architecture in order to avoid relying on massive generalized data. The model's creative model construction idea greatly saves the cost of the pre-training phase and the running phase, reduces the dependence of the large language model on advanced high-end hardware equipment, breaks the traditional thinking that the model can only improve the performance by complying with the Scaling Law, and opens up a brand new track for the model performance improvement in the future. Table 1 below shows the comparison between V3 and ChatGPT-4 in terms of various indexes to reflect the great advantages of Deepseek in reducing the cost of model training and running, as well as the hardware dependency. (Since R1 is built based on V3, the core technology is highly consistent, so the data comparison between V3 and GPT-4 is used to provide a reference for the cost of R1 and traditional inference models such as ChatGPT-o1.)

Table 1 Comparison of training, running, and hardware costs between Deepseek-V3 and GPT-4

norm	**DeepSeek-V3**	**GPT-4**	**cost gap**
Training costs	5.576 million dollars	About $100 million	1/20
Number of GPUs	2048 H100	Over 16,000 pieces of A100/H100	1/8
Reasoning costs ($/thousand tokens)	$0.0012	0.03 dollars	1/25

- ④Other General Models:In addition to the above selected and introduced several general models that can reflect the development of large language models to a certain extent, there are also LLaMA series models developed by Meta, PaLM series models developed by Google, etc., which are constantly upgrading and making breakthroughs in terms of the model scale and the performance of the tasks.

2.2.4. Financial Domain-Specific Models (2021-present)

The big language modeling models in the financial domain belong to the segmented models of the big language generic modeling verticals introduced above, and are the result of fine-tuning and redesigning the generic models. This paper focuses on the BloombergGPT model and seeks to help readers get a quick overview of this domain in order to get a glimpse of what is going on in the pipe.

Bloomberg GPT (2023)

BloombergGPT model is a finance-specific large language model released by Bloomberg on March 30, 2023, and is the first finance-specific large language model. In this paper, we will introduce the BloombergGPT model from four aspects, so that readers can have a comprehensive and deep knowledge about it.

First, in terms of technical architecture, it uses a decoder with 50 billion parameters (causal language model only) and is improved based on the BLOOM architecture, which contains 70 layers of Transformer decoder, 7680 hidden layer dimensions, and 40 multi-attention headers. and it adopts a mixed dataset (54.2% of financial data and 48.73% of general-purpose data), combining a combination of the AdamW optimizer and ZeRO optimization techniques for training, using Unigram disambiguator to improve inference efficiency, supporting dynamic retrieval enhancement and multi-Token prediction, thus reducing GPU usage by 90% to achieve the effect of saving training and running costs (the training cost is roughly $30 million).

Secondly, in terms of data construction, as mentioned before it uses a mixed dataset for pre-training, and here we introduce its dataset content components as well as methods in terms of data cleaning. Its financial dataset (FinPile) contains 363 billion tokens, covering structured and unstructured data such as company documents (42.01%), news (5.31%), and financial reports (2.04%) with a time span of 2007-2022. Its generalized dataset integrates data from THE PILE (25.9%), C4 (19.48%), and Wikipedia (3.35%) to enhance the model generalization capability. The hybrid dataset constructed from these two datasets undergoes data cleaning by optimizing Unigram splitter with partitioning idea, and then is used for pre-training, which theoretically possesses both financial expertise and good generalization ability.

Thirdly, in terms of core ability assessment, for financial task processing, it outperforms models such as GPT-NeoX and OPT in tasks such as sentiment analysis (e.g., FiQA SA), Named Entity Recognition (NER), and News Categorization (FPB), with a ConvFinQA score of 62.51 (53.01 for GPT-NeoX), which is a bright performance. And it equaled or surpassed models with the same parameters in benchmarks such as BIG-bench Hard and Knowledge Evaluation, indicating that it also performed well in terms of generalization ability.

Fourth, in terms of industry impact, the model can greatly reduce the time for analysts to write research reports, and as the first finance-specific big language model, it innovatively proposes a hybrid training method, providing a paradigm for vertical domain big model training. In summary, Bloomberg has become a benchmark for vertical domain big model development by achieving high performance in financial tasks while ensuring a high level of general-purpose capability through hybrid data training and targeted architecture optimization.

Other financial models

After the release of BloombergGPT model, financial big language models ushered in a period of mushrooming and booming development. models such as FinGPT, scholar and Endless have been developed one after another, which are trained on financial domain-specific data to make up for the short board of general-purpose models with insufficient knowledge in the financial domain, and are mainly applied to risk management, fraud detection and quantitative analysis etc.[1], which continues to promote the financialization of large language models.

3. Core technologies and concepts of the big language model

3.1. Recurrent Neural Network (RNN) architecture

The recurrent neural network is an outstanding representative of an early product of machine learning, which dates back as far as 1982-1986. It is a neural network specialized for processing sequential data that enables information to

be passed in the time dimension by introducing recurrent links, with implicit state memory (the network captures dependencies in the sequence over time by hiding the states) and parameter sharing (the same set of weight matrices is reused at all time steps, reducing the number of parameters and capturing the sequence patterns.) The characteristics of the Its network structure consists of three parts: an output layer, a hidden layer, and an output layer, and its cyclic nature can be visualized by unfolding it into a chain structure that allows the same parameters to be shared at each time step. This allows the RNN to store historical information through hidden states and capture temporal dependencies in sequential data, e.g., in speech modeling, the RNN can infer the meaning of the current word by the words that appeared in the previous text. In addition, since RNNs share weight matrices (e.g., weights from input layer to hidden layer and hidden layer to output layer) between time steps, this leads to a significant reduction in the number of its parameters, which greatly improves the training efficiency and enables the model's ability to generalize to handle sequences of different lengths. However, since RNN belongs to the results of early machine learning exploration, it also has many shortcomings : on the one hand, the gradient problem. Although RNN can realize the memory function by hiding the state, its architecture is prone to the problem of gradient explosion under long sequences, which leads to the loss of memory, and its long-term reliance on the problem of weak capture ability cannot achieve the expected theoretical effect. On the other hand, it is poor in parallelism. Since its time step needs to be processed serially, parallel computation is difficult, leading to its poor computational efficiency. However, it is undeniable that RNN still belongs to an excellent neural network architecture, which has a wide range of applications in natural language processing, time series prediction, and sequential decision-making tasks, so two improvement schemes are proposed at this stage to address its shortcomings.

3.1.1. Improved RNN architecture

Long Short-Term Memory Network (LSTM)

Compared to traditional RNN techniques, LSTM introduces gating mechanisms (input gate, forget gate, output gate) to dynamically control the information flow. Its structure is optimized to consist of six parts: forgetting gate, which determines which historical information is discarded; input gate, which is used to control the new information input; candidate memory cell, which stores the current information; output gate, which is used for the final result output; memory cell updating; and hidden state.

Gated Recycling Unit (GRU)

The gated recurrent unit (GRU) is a simplified version of the long and short-term memory network (LSTM), whose core structure consists of four parts: reset gate, update gate, candidate hidden state, and hidden state update. It structurally merges the forgetting gate and input gate into a single reset gate to reduce the number of parameters, and simplifies the LSTM's scheme of separating long-term memory from short-term output through separate cell states and hidden states to storing both uniformly in a single state in terms of memory management. These two simplification initiatives allow GRU to perform well, consume less resources, and be more efficient when dealing with short sequence tasks, but at the same time not perform as well as LSTM when dealing with long sequence scenarios (e.g., translation work).

3.1.2. Variants and extensions

In addition to improved RNN architectures, variants and extended applications of RNNs have also emerged. The following three are common: Bidirectional RNN, which traverses sequences by simultaneously traversing them forward and backward as a way to capture forward and backward dependencies, often used for sentiment analysis ; Deep RNN, which enhances expressive power by stacking multiple RNN layers, but with gradient clipping to circumvent gradient explosions ; Hybrid CNN-RNN architecture, which utilizes CNNs to extract local features, and then RNNs to model global dependencies, often used for video classification.

3.2. Convolutional Neural Networks (CNN)

Convolutional Neural Network (CNN) is a learning model designed for processing grid-like data (e.g., images, videos), which lays a solid foundation for the large language model to move towards multimodality, i.e., processing image and video samples in addition to text, through the core idea of local connectivity and weight sharing mechanism, thus efficiently extracting spatial features and lowering the number of parameters to improve the efficiency and reduce the amount of computation. Its typical structure consists of three parts: Convolutional Layer: using filters with learnable properties to continuously slide over the input samples, read the features, and further extract local features such as edges and textures. Pooling layer: reduces the data dimensions by maximum pooling or average pooling, and retains only the key features to reduce the computation. Fully connected layer: all previously extracted features are pooled and analyzed to map to the final output. Based on its typical structure, it has the property of being able to recognize the samples accurately even after they have been spatially transformed, i.e., it has the excellent property of having

robustness in recognition. And by having the parameter sharing and pooling operation characteristics, it can greatly improve the computational efficiency and reduce the computational cost. In addition, the automatic feature extraction, avoiding manual design of features has also become one of the core advantages of its technology. However, it also has the problems of long training cycle, relying on a large amount of labeled data, and poor model interpretability ("black box" problem). Overall, CNN has made a revolutionary breakthrough in the field of computer vision through hierarchical feature extraction and parameter sharing mechanism, but its application is still limited by the dependence on data labeling and computational resources, and future research should further improve the model efficiency, solve the problem of dependence on labeled data, realize the self-supervised learning, and further explore the direction of cross-modal applications under the condition of ensuring the accuracy.

3.3. Transformer architecture

Transformer is a deep learning architecture that is the cornerstone of modern large language models, revolutionizing the field of natural language processing (NLP). Its core innovation lies in the use of a self-attention mechanism that allows the model to process sequence data in parallel and does not rely on RNNs or CNNs at all, but instead utilizes the self-attention mechanism to capture long-distance dependencies in sequences, overcoming the limitations of traditional Recurrent Neural Networks (RNNs), especially the problem of gradient vanishing when processing long sequences. Its architecture mainly consists of two parts, encoder and decoder, and the following is its detailed structure.

3.3.1. Input embedding and positional coding

Input embedding: the input sequence is converted into a vector representation using a word embedding table.

Positional encoding: Since the Transformer does not have an inherent order like RNN, additional positional encoding is needed to preserve the original sequence order. Positional encoding is usually realized by adding a fixed sine-cosine function value to the embedding vector.

3.3.2. Encoder

The encoder is mainly used to process the input sequence, generating a representation that is used to capture the entire input context, which consists of several identical layers, each containing the following components:

Multi-Head Self-Attention mechanism

The self-attention mechanism will compute the attention weight of each position with respect to the other positions in the sequence. Specifically, for each position, three vectors are computed: query -Q), key (K) and value, V). The attention weights will be computed by the dot product of Q and K, scaled and normalized by softmax, and then further weighted and summed over the V vectors to generate a new representation. In contrast, multi-head attention implies running multiple self-attention processes in parallel, each with independent Q, K, and V weights, with outputs spliced and then linearly transformed to capture different types of relations (e.g., syntactic or semantic).

Feed-Forward Neural Network

The vectors at each position are processed independently through a two-layer fully connected network, usually with a ReLU activation function in the first layer. This allows the model to learn complex nonlinear transformations at each position.

Residual Connections and Layer Normalization

Each component carries out the process of adding its output to the input (residual join) followed by layer normalization. In this case, the residual join helps to train the deep network, while the layer normalization normalizes the features on each sample, further helping the model to stabilize the training process.

3.3.3. decoder

The decoder generates an output sequence based on the output of the encoder and its own input sequence, and its structure is slightly different from that of the encoder.

Multi-head self-attention mechanism (with mask)

Similar to the encoder, but uses masked self-attention to ensure that the decoder can only see previous words and not future words when generating the current word, i.e., causal attention, with which the order of sequence generation is ensured.

Multiattention (to encoder output)

This is an additional layer in the decoder that looks at the output of the encoder to get contextual information about the input sequence.

Feedforward neural network

Similar to the encoder, the vectors for each position are processed independently.

Residual connectivity and layer normalization

Consistency with the same encoder ensures consistency in training.

3.3.4. Output Layer

The final output of the decoder is mapped to the dimension of the vocabulary size through a linear layer before a softmax function generates a probability distribution for each position for predicting the next word.

The training process of Transformer uses teacher coercion and cross-entropy loss functions as well as backpropagation and optimizer to improve the accuracy of prediction and thus has good generalization ability. And during its inference process, the decoder starts from the initial input and gradually generates the output sequence and after each word is generated, it is used as input again and repeated until the end symbol is generated or the maximum length is reached.

In practice, Transformer is equipped with a variety of extensions and applications, the most common are the following three: decoder-only model, represented by the GPT series ; encoder-only model, represented by the BERT model ; multimodal tasks. That is, the Transformer is extended to the track of processing data such as images, speech, etc., demonstrating its strong versatility.

Overall, the Transformer architecture achieves global dependency modeling of sequence data through the self-attention mechanism, and its parallel computing capability breaks through the bottleneck of sequential processing in traditional RNN/LSTM. The encoder-decoder structure combined with the multi-head attention mechanism enables the model to capture multi-dimensional contextual relationships simultaneously, while the residual concatenation and layer normalization techniques significantly improve the training stability. This design not only solves the long sequence dependency problem, but also preserves sequence order information through positional encoding. As the core architecture of deep learning and big language models, Transformer has been widely used in NLP (e.g., machine translation, text generation) and cross-domain tasks (e.g., Vision Transformer for computer vision), and its efficient parallelization feature and scalability have become the key driving force for the evolution of AI technology Big Language Models in Financial Markets: Current Status and Prediction Research.

4. Financial market applications of large language models

In the field of intelligent customer service: with the development of LLMs , in view of the powerful contextual understanding and language generalization ability brought by the Transformer architecture, the in-depth integration of LLMs and intelligent customer service has become an inevitable trend of development. The systems represented by Ant Group's "Xiaobao 2.0" and Ping An Bank's "BankGPT" have realized 24/7 semantic understanding and generation through natural language processing technology, and are able to accurately answer basic questions such as financial information and business handling. Basic questions. For example, in the remote banking application of Industrial and Commercial Bank of China (ICBC), the optimization of the seat service process through big model technology has improved the efficiency by 18%[4] . In addition, in terms of personalized service, the model can generate dynamic suggestions by calling internal knowledge bases (e.g., Du Xiaoman financial knowledge base) and API interfaces. And Chifu Technology uses GPT-4 to optimize telemarketing speech, and the conversion rate is increased by more than 5%[6] . In terms of technological evolution, multimodal interaction has also become a development trend, such as Tencent enterprise point customer service access Deepseek big model, with real-time networking search function, which further models semantic understanding ability, thus realizing the oral expression of financial policy interpretation and complex product information simplification, to help its trained models better competent intelligent customer service work.

in the area of :risk management and complianceIn terms of anti-fraud, Mastercard generates synthetic data to optimize the risk control model through a large model, significantly improving the efficiency of fraud detection. In credit scenarios, Du Xiaoman's "Regulus" big model analyzes credit reports and Internet texts to assist in risk control decisions644[4] . In the field of compliance, the large model automates the processing of financial contracts and reports through natural language processing technology, which can effectively reduce the risk of manual operation and at the

same time comply with the Interim Measures for the Administration of Generative Artificial Intelligence Services and other regulatory requirements[6][7] . Goldman Sachs fine-tuned GPT's model to create a "Hawk-Dove" index that can predict Fed policy moves and analyze market impacts, which is a microcosm of the application of LLMs in risk management and compliance[6] .

In the areas of investment and asset management and market forecasting:In the investment consulting service, Ant Group provides financial selection and asset allocation suggestions based on the financial big model, and the proportion of users' frequent transactions has decreased by 60%, which can effectively reduce the number of mistakes made by users and thus improve their profitability. In the investment research segment, CITIC Securities utilizes a large model to generate industry research summaries to improve efficiency. In the field of quantitative investment, thanks to the in-depth development of CNN technology and the exploration and research of multi-modalization of large models in recent years, LLMs are able to effectively identify and extract the features of unstructured data, such as images, and use them for the next step of in-depth research (such as the analysis of the current investor sentiment of a stock). Based on ChatGPT's strong advantage of processing unstructured data (e.g., company announcements) better than traditional models, it constructs a strategy by multimodally analyzing data such as news and stock k-lines to understand the investment sentiment tendency, and achieves an annualized Sharpe ratio of 3.28% and cumulative return of more than 650% from 2021-2023, which is significantly outperforming traditional models[2] [8] [9] . J.P. Morgan fine-tuned the GPT model to create the "Hawk-Dove" index, successfully predicting Fed policy moves and Treasury rate fluctuations [6] .

In the area of payments and clearing:Payment institutions use big models to analyze transaction text and behavioral patterns to identify unusual transactions, and Mastercard uses synthetic data to train risk control models to significantly improve fraud detection accuracy[4] .

In the field of insurance business:In the underwriting stage, Lemonade's GPT-3-based robot MAYA provides personalized insurance recommendations to improve underwriting efficiency[6] . In the claims stage, the model analyzes medical reports and user descriptions to assist in generating claims conclusions, which can effectively reduce the cost of manual review.

In the field of web processing and automation:Large language models can automatically generate financial reports and meeting minutes, such as Hang Seng Electronics Light-GPT to realize intelligent research report generation in the field of investment consulting. Meanwhile, BERT and other technologies automate the processing of contracts and regulatory texts to improve compliance efficiency.

4.1. Challenges and Issues of Big Language Modeling in Finance

Despite the fact that big language modeling is very widely used in finance, there at presentare many challenges and difficulties .

First, professional capacity and knowledge limitations.The application of big language models in finance faces the problem of lack of professional ability. Zeyu Yao and Hang Su pointed out that although big models can handle general-purpose natural language tasks, they have professional shortcomings in financial scenarios, making it difficult to understand complex financial logic (e.g., derivatives pricing, risk control model construction, etc.)[4] . Zhang Xiaoyan and Wu Huihang further emphasize that the model needs to rely on the fine-tuning of the financial corpus in order to adapt to the industry terminology and business scenarios, whereas the current training data are mostly general-purpose texts, which leads to the limited performance of the model in professional scenarios[6] .

Second, data quality and privacy security.Data quality directly affects model performance. Lu Minfeng mentioned that financial data often contain noise, missing values or unstructured information (e.g., social media opinions), which may lead to model prediction bias[5] . In addition, data privacy issues are particularly prominent: Xiaoyan Zhang and Huihang Wu warn that large model training requires access to a large amount of sensitive financial data (e.g., customer transaction records, credit reports), which poses a risk of leakage[6] ; Shilei Li and Chao Guo add that domestic financial institutions need to comply with Interim Measures for Administration of Generative Artificial Intelligence Services, but data anonymization and encryption technologies still need to be improved in practical applications[7] .

Third, model interpretability and transparency.The "black box" nature of big language models conflicts with the strong compliance requirements of financial scenarios. Zeyu Yao and Hang Su pointed out that the output of the model in the core aspects of credit approval and investment decision-making lacks interpretability, which makes it difficult to satisfy the regulatory requirement of "decision logic transparency"[4] . Lu Minfeng further analyzes that existing research focuses on model performance optimization, but the exploration of explanatory techniques (e.g., SHAP values, locally

interpretable models) is still insufficient[5] . Xiaoyan Zhang and Huihang Wu take the JP Morgan "Hawk-Dove" index as an example to illustrate that the model needs to improve transparency through visualization tools (e.g., attention weight analysis) to cope with high-risk scenarios, such as the Fed's policy interpretation[6] .

Fourth, regulatory compliance and ethical risks.The regulatory framework for big models in the financial industry is not yet complete. Yao Zeyu and Su Hang suggest that big models may violate consumer protection regulations (e.g., misleading publicity) in scenarios such as financial advertisement recommendation and synthetic data generation[4] . Lu Minfeng compares international experience and points out that the hierarchical regulation of high-risk models in the U.S. Generative Artificial Intelligence Safety Executive Order and the EU's Artificial Intelligence Act is worthy of reference, but the domestic policy still needs to refine the requirements for model output content auditing, algorithmic discrimination prevention, etc.[5] . Li Shilei and Guo Chao emphasized that financial institutions need to balance innovation and compliance, such as clarifying the boundaries of decision-making authority and responsibility of models in smart investment advice scenarios, and avoiding legal disputes triggered by "algorithmic black boxes"[7] .

Fifth, the limitations of technology.Big language models face technical adaptation bottlenecks in financial scenarios. Lu Minfeng mentioned that the model is sensitive to latency in real-time trading, high-frequency risk control and other scenarios, and the computational overhead of the Transformer architecture may not be able to meet the demand of high-frequency scenarios[5] . Taking the A-share market as an example, Zhang Xiaoyan and Wu Huihang pointed out that although the big language model can predict market fluctuations through sentiment analysis, the long-term trend prediction is still limited by the depth of the model's understanding of macroeconomic policies[6] . Li Shilei and Guo Chao add that the model is weak in multimodal data processing (e.g., insurance underwriting scenarios combining images and speech) and needs to be integrated with computer vision, speech recognition and other technologies for innovation[7] .

In response to the above difficulties, experts have also proposed differentsolutions and coping strategies. Such as case ofin the model optimization and industry adaptation aspects.

Develop vertical financial domain big models (e.g., "Zhihai-Jinpan" and "Regulus big model"), and enhance professionalism through financial corpus fine-tuning ;Combine small models or traditional AI models (e.g., discriminative AI) to take advantage of their controllability and accuracy to make up for the shortcomings of large models in the core decision-making process.

In the area of ₹data security and privacy protectionData is secured using techniques such as encrypted transmission and anonymization (e.g. differential privacy) ;Establish a strict data hierarchy management system to limit access to sensitive data.

In Interpretability Enhancement and Compliance Design.Introducing methods such as Locally Interpretable Models (LIME) and SHAP value analysis to reveal model decision logic ;Embedding compliance constraints at the model development stage, e.g., filtering high-risk generated content through a rules engine.

In of termstechnological and industrial synergies.Financial institutions and technology companies cooperate to develop financial grand models and share data and arithmetic resources;Promote the construction of open-source frameworks and standardized interfaces to lower the threshold for model deployment.

4.2. Future Directions for IndustryBig Language Modeling Services for the Financial

1. Deepen technology integration and scenarios: Enhance the perception of complex financial scenarios (e.g., fraud detection, contract review) by through combining multimodal data such as image and voice multimodal integration .Deploy AI Agents such as intelligent customer service and investment advisors through to cover 7x24-hour customer service and high-frequency trading scenarios.

2. Improving the regulatory system :Establish a globally unified standard for the financial application of big language models, and implement risk-graded regulation with reference to the EUAI Act (European Union Artificial Intelligence Act).Strengthening model transparency requirements, e.g., through "black box" hacking techniques (e.g., visualization of attention mechanisms) to enhance interpretability.

3. Financial inclusion and efficiency:Enhance the customer experience by using big language models to lower the threshold of financial services and provide personalized product recommendations and risk management services to

long-tail customers.Release human resources through automated processing (e.g., summary financial report generation, contract review) to further focus on high-value-added business.

4. Ethics and Social Responsibility:In-depth study of model bias mitigation techniques (e.g., fairness constraint training) can avoid discriminatory outputs for specific groups (e.g., MSMEs, low-income people).Continuously promote green finance and sustainable development, and rationally use models to analyze environmental and social risk factors.

5. Conclusion

The application of large language models (LLMs) in the financial market has moved from the experimental stage to the practical deepening, and it has reconstructed the core business processes such as customer service, risk management, and investment decision-making through the breakthrough of unstructured data processing capability. Empirical studies show that LLMs have significant advantages in improving service efficiency (e.g., 5% increase in conversion rate of telemarketing), optimizing risk identification (e.g., increase in fraud detection accuracy), and enhancing the accuracy of market prediction (e.g., excellent performance of small-cap stock prediction), etc. Some of the vertical models (e.g., Deepseek-V3) are even optimized by leaps and bounds in terms of training cost and hardware dependency. However, technical limitations (e.g., lack of long-period prediction capability) and compliance challenges (e.g., data anonymization flaws, algorithmic black-box problems) still constrain their large-scale application. Future development should focus on the integration of multimodal technologies, lightweight model development, improvement of regulatory frameworks, and strengthening of ethical constraints, in order to balance the efficiency of innovation and risk control, and to promote the transformation and upgrading of the financial industry towards intelligence and universality. This study provides systematic references for financial institutions and technology developers to help them grasp the opportunities and meet the challenges in the transformation of digital intelligence.

References

[1] Xu Xue Chen.ChatGPT and other big language models to empower the financial industry in the digital era: based on privacy protection, algorithmic discrimination and systemic risk[J]. Journal of Jinan (Philosophy and Social Science Edition),2024,46(08):108-122.

[2] Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. arXiv, abs/2304.07619.

[3] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. neural Information Processing Systems.

[4] Yao Zeyu,Su Hang.Prospect and potential impact analysis of AI big model application in financial industry[J]. International Finance,2024,(10):36-52.DOI:10.16474/j.cnki.1673-8489.2024.10.008.

[5] Lu Minfeng. Research on the principles, challenges and landing paths of the application of big language modeling in finance[J]. Journal of Chongqing Technology and Business University (Social Science Edition),2024,41(04):1-12.

[6] Zhang Xiaoyan,Wu Huihang. Application of large language modeling in finance[J]. Tsinghua Finance Review,2024,(05):22-26.DOI:10.19409/j.cnki.thf-review.2024.05.013.

[7] Li Shilei,Guo Chao. Development research and application prospect of big language modeling technology in financial field[J]. FinTech Times,2024,32(07):12-16.

[8] Obaid, K., & Pukthuanthong, K. (2022). A picture is worth a thousand words: measuring investor sentiment by combining machine learning and photos from news. journal of financial economics,. 144(1), 273-297.

[9] Jiang, J., Kelly, B., & Xiu, D. (2023). (Re-) Imag(in)ing price trends. The Journal of Finance, 78(6), 3193-3249.