

A Comparative Analysis of Financial Large Language Models: Transition from Proprietary High-Capacity Architectures to Efficient Open-Source Frameworks

Vijay K. Vyas¹, Prashant V. Chauhan¹, Chetankumar R. Chauhan¹, Devang A. Karavadiya¹, Mit H. Dave¹

¹ Assistant Professor, Department of Information Technology, VVP Engineering College, Rajkot, India

Date of Submission: 01-04-2026

Date of Acceptance: 10-04-2026

ABSTRACT: This study analyses the evolution of financial Artificial Intelligence (AI) models with a focus on improving efficiency, accessibility, and practical usability. A comparative review approach is used to examine both early large-scale models and recent lightweight, open-source financial language models. The study evaluates their data usage, architectural design, and application-specific performance. The findings reveal that earlier models relied heavily on large proprietary datasets to achieve high accuracy, while recent approaches prioritise efficiency, cost reduction, and task-oriented optimisation. Additionally, the adoption of open-source frameworks and standardised evaluation methods has enhanced model accessibility and consistency across applications. The novelty of this work lies in highlighting the clear transition from performance-centric systems to scalable and application-driven solutions. Furthermore, it emphasises the growing importance of explainability, adaptability, and multilingual capabilities in modern financial AI systems. These advancements indicate a strong shift toward real-time, user-friendly, and practical financial intelligence systems suitable for diverse real-world scenarios.

KEYWORDS: Financial AI, Large Language Models, Open-Source, Efficiency, Scalability, Explainable AI, Financial Technology, Model Evaluation

I. INTRODUCTION

Recent advancements in artificial intelligence have led to the emergence of Large Language Models (LLMs), which are capable of understanding, generating, and processing natural language with high accuracy. LLMs are typically based on the Transformer architecture, which utilizes self-attention mechanisms to capture contextual relationships in text data, enabling improved performance across various natural language processing (NLP) tasks such as text generation, classification, and question answering [10]. A Large

Language Model (LLM) can be defined as a deep learning model trained on large-scale textual data to learn the statistical patterns of language and generate human-like responses. These models have been widely adopted across multiple domains; however, their application in the financial sector presents unique challenges due to the presence of domain-specific terminology, time-sensitive information, and the need for high accuracy.

To address these challenges, Financial Large Language Models have been developed by adapting general-purpose LLMs to financial data and tasks. These models leverage techniques such as fine-tuning, instruction tuning, and domain-specific training to improve their understanding of financial contexts. Fine-tuning involves adapting a pre-trained model to a specific domain using labeled data, while instruction tuning enhances model performance by training it on task-oriented instructions [8]. In addition, recent advancements such as Low-Rank Adaptation (LoRA) and Retrieval-Augmented Generation (RAG) have further improved the efficiency and reliability of financial LLMs. LoRA enables parameter-efficient fine-tuning by updating only a subset of model parameters, reducing computational requirements [7]. On the other hand, RAG integrates external knowledge sources during inference, allowing models to retrieve up-to-date financial information and reduce hallucination in generated outputs [9].

Overall, the development of financial LLMs represents a significant step toward enhancing decision-making processes in finance by enabling accurate analysis of large volumes of financial data. This paper explores various financial LLM models, their techniques, applications, and challenges.

II. LITERATURE SURVEY

Recent advancements in financial artificial intelligence have led to the development of specialized large language models tailored for domain-specific tasks. These models are designed to

overcome the limitations of general-purpose LLMs by incorporating financial knowledge and context.

a. Pure Financial LLM Models

BloombergGPT is a Large Language Model for Finance presents a domain-specific Large Language Model (LLM) designed to address challenges in financial natural language processing. Developed by Bloomberg, it aims to overcome the limitations of general-purpose models such as GPT in handling complex financial text, which includes specialized terminology and time-sensitive information [1]. The model is based on a Transformer architecture with approximately 50 billion parameters and is trained on a large dataset of around 700 billion tokens, combining both financial and general-domain data. This enables it to achieve strong performance across financial NLP tasks such as sentiment analysis, named entity recognition, and question answering, while remaining competitive on general tasks [1]. Despite its effectiveness, BloombergGPT requires significant computational resources and relies on proprietary data, limiting accessibility and real-time adaptability. Nevertheless, it represents a major advancement in financial AI by demonstrating the value of domain-specific LLMs [1]. The model is trained on a combination of proprietary financial data and publicly available datasets, highlighting the importance of high-quality domain-specific data in achieving strong performance.

While BloombergGPT focuses on large-scale financial modeling using predominantly English and proprietary datasets, it also highlights the need for extending such domain-specific approaches to different languages and regions. This has led to the development of models like XuanYuan 2.0, which address similar challenges in non-English financial contexts.

A Large Chinese Financial Chat Model XuanYuan 2.0 introduces a domain-specific conversational LLM designed for financial applications in the Chinese language. It addresses the limitations of general-purpose models in handling financial contexts, particularly in non-English settings with complex linguistic and domain-specific nuances [2]. The model is based on a Transformer architecture and is fine-tuned using large-scale Chinese financial datasets, including news, reports, and regulatory documents. This enables strong performance in tasks such as financial dialogue generation, sentiment analysis, and question answering, outperforming general models on Chinese financial benchmarks [2]. Despite its effectiveness, the model relies on proprietary data and requires high computational resources, limiting accessibility and

real-time adaptability. Nevertheless, it highlights the importance of localized and domain-specific LLMs for financial applications [2].

b. Fine-tuned Financial LLM Models

In addition to training large domain-specific models from scratch, recent research has focused on improving financial LLMs through fine-tuning techniques. These approaches aim to enhance performance while reducing computational cost and increasing accessibility. FinGPT an open-Source Financial Large Language Models presents an open-source framework for developing financial LLMs with a focus on accessibility, transparency, and adaptability. Unlike proprietary models such as GPT and BloombergGPT, it uses publicly available data and community-driven development to improve reproducibility and reduce cost [3]. The framework is built on existing open-source models like LLaMA and ChatGLM, combined with financial fine-tuning pipelines and real-time data sources such as news and social media. It performs effectively in tasks like financial sentiment analysis and stock prediction, achieving competitive results compared to proprietary systems [3]. Despite its flexibility, performance depends on base model quality and data preprocessing. Overall, FinGPT highlights the importance of open, scalable, and real-time adaptable financial AI systems [3].

While FinGPT emphasizes open-source development and real-time adaptability, it does not fully address the high computational cost associated with large models. To overcome this limitation, Fin-LLAMA introduces efficient fine-tuning techniques that focus on reducing resource requirements without significantly compromising performance.

Fin-LLAMA is an Efficient Finetuning of Quantized LLMs for Finance proposes an efficient approach for developing financial large language models by using quantization and parameter-efficient fine-tuning techniques. It addresses the high computational cost of large LLMs by reducing resource requirements while maintaining strong performance [4]. Built on the LLaMA architecture, the model applies quantization and techniques like LoRA to enable faster training and lower memory usage. It achieves competitive performance in financial NLP tasks with minimal performance trade-offs [4]. Overall, Fin-LLAMA highlights the importance of efficiency and scalability in financial LLMs [4]. Although Fin-LLAMA improves efficiency through quantization and parameter-efficient tuning, further improvements can be achieved by enhancing model alignment with specific tasks. This leads to the use of instruction

tuning techniques, as demonstrated in Instruct-FinGPT.

Instruct-FinGPT is a financial Sentiment Analysis by Instruction Tuning proposes an approach to improve financial sentiment analysis using instruction tuning on large language models. It addresses the limitations of traditional fine-tuning by guiding models with task-specific instructions, enabling better understanding of financial context [5]. Built on the FinGPT framework, the model uses diverse financial datasets such as news and social media to learn sentiment patterns. It achieves improved performance and robustness in financial sentiment analysis [5]. Overall, it demonstrates the effectiveness of instruction tuning in financial NLP tasks [5].

c. Framework, benchmark, evaluation

Beyond model development and fine-tuning, recent research has also focused on creating standardized frameworks and benchmarks to

evaluate financial LLMs. These approaches aim to ensure consistency, reliability, and fair comparison across different models. PIXIU is a Large Language Model, Instruction Data and Evaluation Benchmark for Finance proposes a unified framework for financial LLMs that combines model development, instruction datasets, and evaluation benchmarks. It addresses the lack of standardization in financial NLP, enabling better comparison and evaluation of models [6]. The framework provides financial instruction data and benchmark tasks such as question answering and sentiment analysis, improving model performance and generalization. Furthermore, it emphasizes instruction tuning to align models with real-world financial applications, enhancing usability in areas such as financial advisory and risk analysis [6]. Thus, PIXIU complements existing model-focused approaches by providing a structured foundation for evaluation and benchmarking, highlighting the importance of standardization in advancing financial LLM research.

III. COMPARISON OF DIFFERENT LLM MODELS

Below table compares different LLM models with respect to different parameters.

Table 1. Comparison of LLM models

Model	Type	Data Source	Key Technique	Strengths	Limitations
BloombergGPT [1]	Pure Financial LLM	Proprietary + Public Financial Data	Large-scale pretraining	High accuracy, strong domain understanding	High computational cost, not accessible
Xuan Yuan 2.0 [2]	Pure Financial LLM	Chinese financial datasets	Domain-specific fine-tuning	Strong multilingual capability, good for Chinese finance	High resource requirement, limited accessibility
FinGPT [3]	Fine-tuned / Framework	Public + Real-time data (news, social media)	Open-source fine-tuning	Accessible, adaptable, real-time capability	Depends on base model quality
Fin-LLAMA [4]	Fine-tuned Model	Financial datasets	LoRA + Quantization	Efficient, low resource usage	Slight performance trade-off
Instruct-FinGPT [5]	Instruction-tuned Model	Financial instruction datasets	Instruction tuning	Better task alignment, improved generalization	Depends on instruction quality
PIXIU [6]	Framework / Benchmark	Financial instruction datasets	Benchmark + Evaluation framework	Standardized evaluation, consistency	Not a standalone model

IV. DECISION FRAMEWORK FOR FINANCIAL LLM SELECTION

Selecting an appropriate LLM approach for financial applications depends on factors such as data

availability, computational resources, performance requirements, and data privacy constraints. It is recommended to begin with lightweight approaches such as zero-shot or few-shot learning using pre-

trained models, especially when labelled data or computational resources are limited. The choice between third-party APIs and open-source models depends on factors such as data confidentiality, deployment requirements, and cost considerations. If initial performance is not sufficient, more advanced methods such as fine-tuning with domain-specific financial data or integration of external tools can be adopted to improve accuracy and task-specific alignment. The framework follows a step-wise progression, where practitioners move from simple approaches to more complex solutions only when necessary. For highly complex applications requiring maximum performance, training a domain-specific LLM from scratch may be considered, although it requires significant computational resources, large-scale data, and expertise. Overall, this decision-making framework helps balance cost, efficiency, and performance, enabling the selection of the most suitable LLM approach for real-world financial applications [11].

V. KEY TECHNIQUES IN FINANCIAL LLM

Modern financial large language models rely on several advanced techniques to improve performance, efficiency, and reliability in domain-specific applications. These techniques play a crucial role in enabling scalable and practical deployment of LLMs in the financial sector.

a. Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning method that reduces computational cost by updating only a small subset of model parameters instead of the entire model. This significantly lowers memory usage and training time while maintaining strong performance, making it suitable for deployment in resource-constrained environments [7].

b. Instruction tuning is another important technique that enhances model alignment with specific tasks by training on instruction-based datasets. By incorporating explicit task descriptions and examples, the model becomes more capable of generating context-aware and accurate responses, particularly in financial applications such as sentiment analysis and question answering [8].

c. Retrieval-Augmented Generation (RAG) further improves model reliability by integrating external knowledge sources during response generation. Instead of relying solely on pre-trained knowledge, the model retrieves relevant financial information from external databases or documents at inference time. This approach reduces hallucination and enhances the accuracy and relevance of generated outputs [9].

VI. EVALUATION OF FINANCIAL LLMs

Evaluating financial large language models (LLMs) is essential to ensure their reliability, accuracy, and practical usability in real-world financial applications. Due to the high-stakes nature of financial decision-making, multiple evaluation techniques are required to assess different aspects of model performance [11].

a. Classification-based Evaluation

Classification-based evaluation is widely used in financial LLMs for tasks such as sentiment analysis, fraud detection, and financial news classification. This approach measures the model's ability to correctly categorize input data into predefined classes. Common evaluation metrics include precision, recall, and F1-score, which assess the accuracy and completeness of predictions. These metrics are critical in financial applications, where incorrect classification may lead to misleading insights and poor decision-making [11].

b. Regression-based Evaluation

Regression-based evaluation is applied in financial tasks such as stock price prediction, trend forecasting, and risk estimation. In this approach, the model's performance is evaluated based on how closely its predicted numerical values match actual values. Metrics such as Root Mean Square Error (RMSE) and the coefficient of determination (R^2) are commonly used. These metrics help determine the model's prediction accuracy and reliability in continuous financial scenarios [11].

c. Performance-based Evaluation (Financial Metrics)

Performance-based evaluation focuses on real-world financial outcomes rather than only predictive accuracy. This includes metrics such as profit, return, and Sharpe ratio, which measure the effectiveness of financial decisions made using the model. Such evaluation is essential in finance because even a highly accurate model may not necessarily result in profitable outcomes. So, this approach ensures that the model's predictions translate into practical financial benefits [11].

d. Backtesting and Historical Evaluation

Backtesting involves evaluating the model using historical financial data to simulate how it would have performed in past market conditions. This technique is widely used in trading strategies and investment analysis. It allows researchers to assess the model's stability and effectiveness before deploying it in real-world environments. Backtesting provides valuable insights into potential risks and performance under different historical scenarios [11].

e. Real-time (Online) Evaluation

Real-time or online evaluation measures the performance of financial LLMs in live environments where data is continuously updated. This approach is crucial for applications such as automated trading systems and financial advisory platforms. It evaluates the model's ability to adapt to rapidly changing market conditions and maintain consistent performance over time [11].

f. Robustness and Generalization Evaluation

Robustness and generalization evaluation ensures that the model performs consistently across different datasets and varying market conditions. A robust model should not be overly dependent on specific patterns in training data. This evaluation helps determine whether the model can handle unseen data effectively and maintain reliability in diverse financial scenarios [11].

g. Overfitting and Reliability Assessment

Overfitting occurs when a model performs well on training data but fails to generalize to new, unseen data. This evaluation technique focuses on identifying such issues and ensuring that the model maintains consistent performance across different datasets. It is important to ensure alignment between accuracy metrics and real-world financial performance to avoid misleading results [11].

h. Fairness and Bias Evaluation

Fairness and bias evaluation examines whether the model produces biased or unfair outputs, particularly in sensitive financial applications such as credit scoring and loan approval. Since financial decisions can have significant societal and economic impacts, it is essential to ensure that the model operates in a fair and unbiased manner [11].

VII. CHALLENGES IN FINANCIAL LLMs

Financial large language models face several challenges that affect their reliability and real-world deployment:

- **Hallucination:** Models may generate incorrect or misleading financial information, which can lead to poor decision-making and financial risks [11].
- **Bias in Outputs:** LLMs can inherit biases present in training data, potentially affecting fairness in applications such as credit scoring and risk assessment [11].
- **Lack of Explainability:** These models operate as black-box systems, making it difficult to interpret their reasoning and outputs [11].
- **Data Privacy and Security:** Financial data is highly sensitive, and ensuring compliance with privacy regulations remains a major challenge.

- **High Computational Cost:** Training and deploying large-scale financial LLMs require significant computational resources, limiting accessibility.
- **Real-time Adaptability:** Financial markets are dynamic, and models may struggle to stay updated with rapidly changing information.

VIII. CONCLUSION

This study highlights the significant evolution of financial Artificial Intelligence (AI) models from large, performance-driven systems to more efficient, scalable, and application-oriented solutions. Earlier models primarily focused on achieving high accuracy by leveraging large proprietary datasets and substantial computational resources. In contrast, recent financial language models emphasize efficiency, cost-effectiveness, and task-specific optimization, making them more practical for real-world use. The comparative analysis also shows that the adoption of open-source frameworks and standardized evaluation methods has improved accessibility, transparency, and consistency across financial AI applications. Furthermore, modern models are increasingly designed to support explainability, adaptability, and multilingual capabilities, which are essential for broader adoption in diverse financial environments. Overall, the study confirms a clear shift toward developing user-friendly, real-time, and scalable financial AI systems that can effectively address real-world challenges. Future research will focus on developing real-time, efficient, and explainable financial AI systems that can support better decision-making. There is also a need to improve multilingual capabilities, standard datasets, and low-cost model designs.

REFERENCES

- [1]. Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann, BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564, 2023.
- [2]. Xuanyu Zhang, Qing Yang, and Dongliang Xu, XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters. arXiv:2305.12002, 2023.
- [3]. Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang, FinGPT: Open-Source Financial Large Language Models. arXiv:2306.06031, 2023.
- [4]. Pedram Babaei William Todt, Ramtin Babaei, Fin-LLAMA: Efficient Finetuning of

- Quantized LLMs for Finance, 2023. <https://github.com/Bavest/fin-llama>.
- [5]. Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu, Instruct- FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. arXiv:2306.12659, 2023.
- [6]. Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang, PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. arXiv:2306.05443, 2023.
- [7]. Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685, 2021.
- [8]. Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, Training language models to follow instructions with human feedback. arXiv:2203.02155, 2022.
- [9]. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401, 2021.
- [10]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, Attention Is All You Need. arXiv:1706.03762, 2017.
- [11]. Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli, Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. arXiv:2102.02503, 2021.