Survey on Multimodal Large Language Models for Financial Applications: Enhancing Customer Service, Investment, and Compliance

Anonymous ACL submission

Abstract

This survey explores the growing role of Multimodal Large Language Models (MLLMs) in the financial sector, where their ability to process and analyze diverse data formats—text, images, audio, and time series—unlocks new potential for automating and enhancing financial decision-making. We examine key applications, including customer service, investment research, and regulatory compliance, demonstrating how MLLMs outperform traditional models in handling complex tasks such as financial chart analysis and sentiment assessment. However, challenges remain in data availability, real-time processing, and model transparency, limiting broader adoption. By identifying these challenges and suggesting future directions, we aim to bridge the gap between the potential of MLLMs and their practical use in finance, providing a comprehensive resource for both financial experts and researchers.

1 Introduction

In this article, we conduct a comprehensive survey of Multimodal Large Language Models (MLLMs) in the financial sector. Traditionally, tasks such as auditing financial documents and risk assessment in decision-making were labor-intensive processes. Since the early 2000s, machine learning (ML) and artificial intelligence (AI) have significantly automated many processes in the financial domain (Pástor, 2000; Pástor and Stambaugh, 2000; Lettau and Ludvigson, 2001). These technologies, such as supervised, unsupervised, semi-supervised, and reinforcement learning, have been applied to address various financial challenges (Avramov et al., 2022; Ross, 2013; Cong et al., 2020). For example, supervised learning models, such as pooled panel lasso regression and decision tree are used to capture the dynamics of stock returns (Avramov et al., 2022; Hong et al., 2000). Unsupervised models and reinforcement learning have been used to analyze risk-return trade-offs to optimize the portfolio management objective (Ross, 2013; Cong et al., 2020). However, most ML models rely on structured numerical data, which poses challenges when dealing with unstructured financial data, such as emails and documents (Chen et al., 2020; Ke et al., 2019).

In the past two years, the financial industry has increasingly embraced the integration of Large Language Models (LLMs) and MLLMs to enhance various functions. LLMs can efficiently handle text-based tasks such as sentiment analysis and time series forecasting for stock prediction, support personalized investment strategies, automate financial statement analysis, and improve portfolio management (Ko and Lee, 2024; Zhao et al., 2024; Li et al., 2023c). MLLMs further incorporate diverse data types such as images, and audio to improve accuracy in sentiment analysis (Srivastava et al., 2023) and surpass single-modality technologies in understanding and interpreting complex financial charts, providing deeper insights into financial applications (Wang et al., 2023; Zhang et al., 2024a).

In this survey, we review key developments in MLLMs for financial tasks. We specifically explore three areas: personal user interactions, client-facing customer service and investment research, and compliance and industry summaries. Figure 1 illustrates the structure of our paper, organized around these three areas. Each area is further broken down into specific application scenarios.

In personal user interactions, MLLMs have enhanced customer service and investment assistance by analyzing both text and images. This enables a deeper understanding of customer sentiment and allows for more personalized responses (Thakkar et al., 2024). Tools like FLLM (Chu et al., 2023) and CFGPT (Li et al., 2023a) can process multiple input formats, allowing them to handle more complex financial queries, while FinVis-GPT (Wang et al., 2023) integrates financial chart analysis with text data, empowering users to make better-

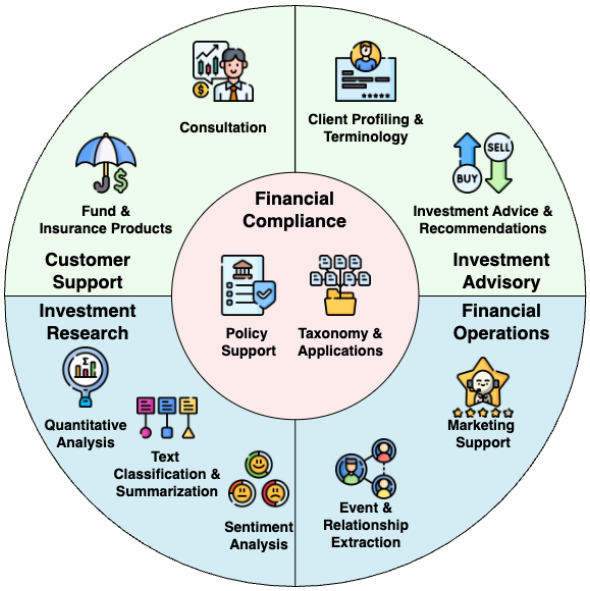

Figure 1: An overview of our paper structure, focusing on three perspectives - green for personal, blue for corporate, and red for compliance - along with specific application areas.

informed investment decisions.

From an enterprise perspective, MLLMs enhance automated customer support by incorporating sentiment analysis that combines text and audio, allowing for a more accurate understanding of customer emotions and more personalized responses (Srivastava et al., 2023). Similarly, in investment research, MLLMs surpass existing technologies in interpreting complex financial charts, offering deeper insights that improve investment strategies and portfolio management (Wang et al., 2023).

In financial compliance and industry analysis, MLLMs have raised concerns about data security and privacy due to the potential exposure of sensitive information during model training. MLLMs enhance compliance efforts by integrating various data formats to detect complex risks, such as real-time market sentiment analysis for early fraud detection and risk assessment (Delgadillo et al., 2024). Models like FinVis-GPT enable institutions to analyze complex financial data, including reports and stock charts, while FinTextQA and LongFin improve document interpretation (Chen et al., 2024a; Masry and Hajian, 2024). Collaboration between academia and industry is driving innovation, accelerating the adoption of MLLMs in finance (Wang et al., 2023; Li and Chang, 2024).

Despite recent progress, MLLMs in finance still face challenges, including limited data availability, real-time processing issues, and model interpretability. Financial data is often sensitive and complex, making it difficult to create publicly available multi-modal datasets (Zhao et al., 2024). Developing more comprehensive public and private datasets could help address this. Additionally, MLLMs can be computationally intensive, leading to latency issues when processing real-time data for high-frequency trading or real-time risk management (Zhang et al., 2024a). To mitigate this, future datasets could incorporate real-time updates, making systems more responsive to market changes. Lastly, research should focus on improving model interpretability, helping clarify how different data types influence investment decisions.

This survey comprehensively introduces the development status of MLLMs in the financial field through a progressive explanation of three application scenarios: individual, enterprise, and compliance. It presents a fresh perspective on the most widely applicable business scenarios while also exploring the development methods and geopolitical effects of large models in the financial industry. Figure 2 presents a selection of representative papers from each application area, categorized by the three perspectives and arranged in the order of their development. Finally, we address some challenges and future development directions.

The rest of this survey is structured as follows: Section 2 reviews related work, focusing on previous surveys of financial applications using LLMs. Sections 3, 4, and 5 examine MLLMs solutions across three key tasks: personal user, client-facing customer services and investment research, and compliance and industry summaries. Finally, Section 6 discusses these models' challenges and future opportunities.

2 Related Work

2.1 Existing Single-Modality Focus in Financial Applications of LLMs

Previous research have demonstrated the efficacy of LLMs in handling linguistic tasks such as textual analysis and sentiment detection. Nie et al. (2024) categorizes LLM applications across tasks like sentiment analysis, financial time series forecasting, and agent-based modeling. Despite their thorough examination of LLM applications in finance, these surveys primarily focus on text-based tasks. The integration of multimodal data, such as combining textual analysis with visual or structured financial data, remains an under-explored area. This gap

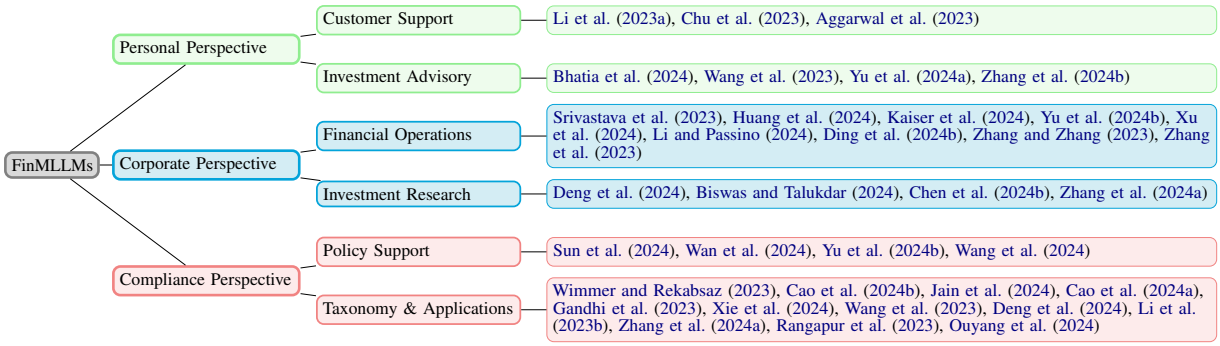

Figure 2: Selected representative papers for each application area across the three perspectives of interest, organized in order of their development.

limits the application of LLMs in more complex financial scenarios where combining multiple data modalities is essential for deeper insights.

Li et al. (2023c) also highlights the dominance of single-modality tasks such as text-based financial sentiment analysis and document classification. While the paper discusses models like FinGPT and BloombergGPT, it underscores that these models largely focus on text-based financial datasets. The challenges of integrating multimodal data, which includes visual financial charts or structured datasets like stock prices and corporate financial reports, have yet to be fully addressed. This singular focus on text neglects the potential benefits of merging text with other data formats, such as visual data from financial charts, which could provide a more comprehensive analysis of market conditions.

2.2 Limitations of Single-Modality Approaches

The focus on single-modality LLM applications limits their effectiveness in complex financial decision-making tasks. For example, relying solely on textual analysis may overlook critical insights from visual financial charts or structured data in financial statements. As highlighted in Ding et al. (2024a), LLMs used as financial trading agents primarily process text data, such as news and reports, while underutilizing numerical and visual inputs like stock price trends and K-line charts. In addition, Fan et al. (2024) illustrates that single-modality approaches often fall short in complex tasks: although Retrieval-Augmented Generation (RAG) techniques have attempted to bridge this gap by integrating external knowledge bases, their application remains predominantly text-focused.

3 Personal User

3.1 Customer Service

Existing financial MLLMs ensures that the system can cater to a wide range of user preferences, allowing for communication across various platforms and formats. They can not only handle customer queries, but also facilitate understanding of complex financial products. In the application domain of customer service, financial MLLMs can ensure that personal users gain access to trustworthy financial advice and make informed financial decisions.

CFGPT (Li et al., 2023a), a Chinese financial assistant MLLM, accepts various inputs including text, audio, and PDF files while offering multiple output formats such as raw text, templated responses, and mind maps. It also integrates Content Summary and Causal Reasoning to enhance the MLLM's conditional and open-domain question-answering capabilities. FLLM (Chu et al., 2023) integrates both tabular data and long textual content to handle complex financial scenarios requiring numerical reasoning based on formulas and diverse input formats. Since it generates QA data using a formula-driven approach, FinLLMs ensures that all questions posed by users are answerable with supporting facts, enhancing the model's reliability in customer support scenarios. NLP Chatbot (Aggarwal et al., 2023) incorporating both text and voice modalities features emotion recognition, multilingual support, speech-to-text conversion, and offline functionality for personalized customer service. Using reinforcement learning and sentiment analysis, it supports diversified user interactions.

3.2 Investment Assistant

By combining multimodal data, financial MLLMs can act as intelligent financial assistants with a more comprehensive understanding of the client's

financial situation and risk appetite and provide tailored recommendations.

FinTral (Bhatia et al., 2024) utilizing textual, numerical, tabular, and image data undergoes domain-specific pretraining, instruction fine-tuning, and Reinforcement Learning with AI Feedback (RLAIF) to optimize its performance in financial decision-making and analysis. FinTral's visual processing capabilities are enhanced through integration with the CLIP vision encoder, making it capable of interpreting financial charts and graphs. FinVis-GPT (Wang et al., 2023) provides human-like explanations and offers predictions in the financial domain through advanced chart interpretation and predictive analysis. The core strength of FinVis-GPT lies in its ability to handle both visual financial charts like candlestick charts and line charts and textual information. FINCON (Yu et al., 2024a) is built with multimodal financial data including stock price information, daily financial news, company filing reports, and Earnings Conference Call (ECC) audio recordings. FINCON also employs a multi-agent hierarchical structure to support financial decision-making, particularly in tasks like single stock trading and portfolio management. FinAgent (Zhang et al., 2024b) integrates text and visual data through a tool-enhanced approach and utilizes dual-layer reflection module enabling rapid adaptation to market changes.

4 Enterprise Perspective

4.1 Client-facing Customer Services

Financial institutions invest a considerable portion of their revenue in client-facing customer service; however, they frequently encounter challenges in delivering an optimal customer experience. This disparity arises from a significant misalignment between customer expectations and the actual services rendered. The advent of AI has catalyzed a transformative shift, redefining the paradigms of customer service and operational efficiency. Multimodal financial modeling can offer significant advantages in these contexts. By integrating diverse data types, such as text, image, audio, and video, these models can enhance the accuracy and richness of customer services, especially allowing for personalized experiences (Lee et al., 2022).

4.1.1 Customer Support

Multimodal financial intelligent customer service integrates images, voice, and text data to enhance the handling of complex business tasks. For example, customer service systems in the banking sector can extract and analyze data by recognizing insurance policies uploaded by customers. Additionally, through the integration of emojis, videos, and voice interactions, these multimodal services offer customers a more intelligent and personalized support experience.

Srivastava et al. (2023) introduce a multimodal learning framework for customer service that leverages acoustic and linguistic information from real customer service conversations, as well as survey feedback, analyzed through a decision-level fusion approach. As a result, the model outperforms monomodal methods.

4.1.2 Investment Advisory

Robo-advisers (RAs) utilize advanced algorithms and AI models to create personalized investment strategies. These automated systems provide tailored asset management services, making sophisticated financial advice more accessible (Huang et al., 2024). Trust is crucial in the financial industry, and a key question is whether RAs can build trust and influence human decisions through behaviors like making eye contact (Kaiser et al., 2024). Multimodal methods are required to enhance RAs' adoption and usability.

4.1.3 Key Techniques

Causality Extraction (CE) and Named Entity Recognition (NER) enhance AI-driven customer service in finance by improving information understanding and analysis. CE commonly relies on LLMs to extract causal relationships from text. These models process large data volumes to identify cause-and-effect pairs and benefit from knowledge graphs for enhanced causal reasoning. RC2R (Yu et al., 2024b) integrates multi-level data to analyze risk contagion, outperforming other models in prediction and generalization. FinCaKG (Xu et al., 2024) creates detailed financial knowledge graphs for causal relationship extraction, surpassing existing technologies. FinDKG (Li and Passino, 2024) tracks strategic thematic investing trends through a dynamic knowledge graph built from financial news. NER identifies financial terms in texts. GNER (Ding et al., 2024b) enhances entity recognition with negative examples and an LCS matching algorithm, improving zero-shot performance in new domains. FinBERT-MRC (Zhang and Zhang, 2023) reframes NER as a machine read-

ing comprehension task, outperforming sequence tagging models. Zhang et al. (2023) present a model for recognizing complex Chinese financial entities, outperforming other models in this task.

4.2 Investment Research

Multimodal modeling is crucial for investment research tasks in finance, such as financial sentiment analysis, text classification, and summarization, as well as quantitative analysis. It enables the integration of diverse data sources from news articles, social media, and research reports, enhancing the depth and accuracy of analyses and providing insights into market trends and investor behavior. This synthesis of qualitative and quantitative data supports more informed investment decisions, ultimately improving the effectiveness of quantitative analysis and allowing market players to respond swiftly to market dynamics.

4.2.1 Financial Sentiment Analysis

Sentiment analysis is intensively used in various financial scenarios to analyze emotions, opinions, and subjectivity (Tan et al., 2023). By incorporating diverse data types like text, images, and audio, the multimodal models can reduce information bias and enhance responsiveness to market changes in sentiment analysis, leading to better accuracy in analysis than unimodal ones (Du et al., 2024).

By conducting sentiment analysis at the sentence level and combining text and audio data, FMPAF proposed by Deng et al. (2024) generates reliable policy stance indicators, which effectively forecast financial market shifts and measure the global spillover effects of language used by the Federal Reserve Chair.

4.2.2 Financial Text Classification and Summarization

Document classification assigns category labels based on text content and predefined labels. FinEmbedDiff(Biswas and Talukdar, 2024) use pre-trained multimodal embedding models to classify financial documents that include text, tables, charts, and images. This method effectively classifies new documents by generating multimodal embedding vectors for the documents and comparing them with pre-computed class embeddings using vector similarity metrics. FinEmbedDiff significantly reduces computational costs while maintaining accuracy. Similarly, LION (Chen et al., 2024b) enhance visual information processing by integrating fine-grained spatially aware visual knowledge and high-level semantic visual evidence. LION excels in text classification by processing various types of information, including text and images.

Summarization models include extractive (select key sentences) and abstractive (predict tokens generatively) (Jin et al., 2024). MLLMs focus on abstractive summarization to incorporate non-textual information. General-purpose large models, such as GPT-4o, are capable of processing various data types, including text, images, tables, and source code, thereby providing more comprehensive summaries. However, specialized open-source large models for the emerging field of multimodal financial text summarization are still scarce.

4.2.3 Research Reports, Investment, and Quantitative Analysis

There is a growing body of research leveraging MLLMs in investment and quantitative finance. As multimodal financial macromodeling continues to evolve, research in this area has yielded significant advancements.

FinAgent (Zhang et al., 2024a) uses a deep encoder to map multimodal financial time series into a latent space that captures trends and attributes for investment analysis. Developments in MLLMs are revolutionizing financial analysis, enabling more precise predictions, better investment strategies, and more robust trading frameworks.

5 Financial Compliance and Industry Analysis

5.1 Compliance

The growing use of large language models in the financial sector has raised significant concerns about their safety and security. These concerns are not only important for businesses but also directly impact individuals' interests across the board.

5.1.1 Policy compliance

Policy compliance refers to ensuring that the development, deployment, use, and governance of large language models (LLMs) adhere to relevant laws, regulations, and ethical standards. Establishing a comprehensive policy compliance framework not only helps companies avoid legal risks and financial losses but also builds user trust, maintains a good public image, and promotes the sustainable development of the LLMs industry. For LLMs, various studies have explored related safety and privacy issues. Research has found that pre-trained

LLMs can infer personal data through output text (Staab et al., 2023), and harmful content can be easily generated by fine-tuning to remove reinforcement learning from human feedback (RLHF) protections (Zhan et al., 2024; Pelrine et al., 2024). Multimodal large language models(MLLMs) not only have the same issues but also face greater challenges in privacy protection, technical implementation, and security risk identification due to the increased complexity and diversity of the data they handle. In response to these challenges, both governments and companies need to take targeted measures to ensure the safety and compliance of MLLMs.

5.1.2 Individual Perspective

For individuals, financial multimodal large language models may come into contact with data containing a large amount of personal information during the training process. Once the model becomes aware of this personal information, malicious actors can use malicious prompts to access this private information (Sun et al., 2024). At that time, users will not only face the risk of privacy leaks and damage to their personal reputation but also the risk of financial fraud. In severe cases, it can even lead to mental health issues.

5.1.3 Corporate Perspective

For businesses, ensuring the security of customer assets and data is key to building customer trust. Wan et al. (2024) enhanced the efficiency and accuracy of data anomaly detection in the financial market. At the same time, Yu et al. (2024b) provides a new method for understanding and preventing the spread of financial risks, the RC2R mentioned in the text can align text and graph structure data at different granularities, thereby quantifying risk propagation and using Sankey diagrams to intuitively display the path and intensity of risk propagation, as well as providing detailed causal explanations, enhancing the enterprise's ability to prevent financial risks. To address privacy concerns for customers, Wang et al. (2024) enhanced the harmlessness and resistance of multimodal large language models when facing harmful inputs during the inference alignment process, ensuring that the model aligns with human values and intentions.

5.2 Industry Analysis

Under the impetus of financial technology innovation, multimodal large language models are becoming a key driving force in the financial sector. Universities, research institutions, and technology companies are investing efforts to jointly promote in-depth research and application of financial multimodal large models, injecting new vitality into the future development of financial technology.

5.2.1 Impetus for Development

The rise of multimodal analysis techniques has become indispensable in extracting actionable insights from limited and diverse data sources in finance. Increasing attention is being paid to analyzing long-text financial reports, conference audio recordings, and sentiment data, which require advanced models to address the complexity and variability of financial information.

For example, Wimmer and Rekabsaz (2023) has been able to capture richer data features by combining image and text information, thereby improving the accuracy of stock market predictions and Cao et al. (2024b); Jain et al. (2024); Cao et al. (2024a) have emphasized the multimodal fusion processing of text and audio data, extracting richer and more predictive insights and provide comprehensive financial risk predictions by combining multimodal information. At the same time, Gandhi et al. (2023) has discussed how to analyze and understand human emotions and opinions by combining text, audio, and visual data. Xie et al. (2024) have proposed FinLLaVA that based on the FinLLaMA, enhancing the handling capacity of complex financial data such as images, tables, and charts. These innovations enhance the model's ability to handle both content and structure, enabling deeper document understanding.

Moreover, LLMs have drawn increasing numbers of experts from computer science, machine learning, and related fields to contribute to advancement. This influx of diverse expertise is fueling innovation and growth in the development of financial MLLMs, accelerating progress in fintech and expanding the horizons of financial analysis. Continuous engagement of scholars ensures that new techniques and approaches are regularly introduced, driving the evolution of financial technologies.

5.2.2 Academia-Industry Overview

Research on financial LLMs is advancing globally, with China emerging as a major hub for innovation in this field. Chinese universities are producing significant breakthroughs in financial LLMs, driven by the need to process and analyze large-scale, mul-

timodal financial data.

For instance, Harbin Institute of Technology developed FinVis-GPT (Wang et al., 2023), an MLLM designed explicitly for financial chart analysis. The FMPAF (Deng et al., 2024) proposed by Peking University and Beijing University of Posts and Telecommunications can conduct fine-grained multimodal analysis of text and audio data, and reveal how the emotional changes affect the dynamics of the financial market.

In parallel, American universities are making important strides. The Stevens Institute of Technology has developed TradingGPT (Li et al., 2023b) to simulate the hierarchical structure of human memory, enabling agents to more effectively process and prioritize urgent and critical tasks.

While academic institutions are prolific in publishing research, large technology companies like OpenAI and DeepMind have published fewer papers in the financial LLM space. However, they have significant competitive strength in technology development and application. This is likely due to the industry focusing more on the practical application of methods while academia pays more attention to the theoretical development and innovation of models.

5.2.3 Cooperation methods

The development of financial LLMs is characterized by extensive collaboration, both between academic institutions and between academia and industry. One successful example is FinAgent(Zhang et al., 2024a), proposed by Nanyang Technological University, National University of Singapore, Zhejiang University, and Singapore Management University, which enables comprehensive analysis of market information and optimization of trading decisions through advanced multimodal techniques. Another example is Fin-fact (Rangapur et al., 2023), introduced by China University of Petroleum and Illinois Institute of Technology, offering a multimodal financial fact-checking platform that combines text and visual content to enhance the accuracy and transparency of fact-checking in the financial domain.

Collaboration between academia and industry helps to transform theoretical research into practical applications. Ouyang et al. (2024) introduced MANAGER, capable of integrating multimodal information such as text, video, and audio, along with relevant external financial knowledge, effectively predicts financial asset price movements and volatility from monetary policy conference calls.

6 Challenges and Future Directions

While MLLMs offer numerous advantages in the financial domain, such as integrating diverse data sources and providing more accurate analysis and decision support, they still face critical challenges, particularly in areas like data availability, real-time processing, and model interpretability.

6.1 Data Scarcity and Integration

In the financial domain, publicly available multimodal datasets are relatively scarce due to the sensitive, complex, and confidential nature of financial data. Financial institutions are often reluctant to disclose proprietary data, such as trading records, customer information, and private financial reports. The high cost of obtaining such data, combined with strict privacy regulations, further limits accessibility for researchers. Multimodal data requires high-quality annotation from multiple sources, and tasks like sentiment analysis or causal reasoning demand significant manual effort, making annotation both expensive and time-consuming. Existing financial datasets are often limited to single-modality annotations, lacking the joint labeling necessary for cross-modality analysis.

Constructing multimodal datasets is challenging due to the integration of diverse data types like text, images, time series, and audio, which complicates collection and annotation. Combining multimodal financial data—such as news, market data, charts, and investor call recordings—adds complexity due to variations in format, scale, and frequency, making it difficult to maintain data integrity while merging and modeling these inputs.

A promising direction for future research is the development of more comprehensive, publicly accessible multimodal financial datasets. Collaboration between financial institutions could also lead to joint private datasets, further enhancing data availability for research and industry use.

6.2 Real-Time Constraints on Data

In financial investment, decision-making based on multimodal data demands extremely high real-time performance, as market fluctuations mean that delayed decisions can directly affect profits. However, current multimodal datasets and systems often fail to meet these stringent real-time requirements. A key challenge lies in the fusion of heterogeneous

data: some financial data, like quarterly reports, are inherently not real-time, while others, such as news, are generated irregularly. Integrating these into a real-time stream requires specialized techniques. Additionally, the complexity of processing multimodal data requires significant computing power, making both training and inference phases time-consuming, which delays actionable insights and undermines their effectiveness in high-frequency decision-making contexts.

Future datasets could transition from static historical records to real-time data streams, incorporating mechanisms for continuous updates. For example, news and social media data could be integrated through APIs for real-time streaming, while market data and structured financial reports could be regularly pushed as updates. This shift would make multimodal systems more responsive to dynamic market conditions, reducing decision-making latency. Additionally, systems could benefit from intelligent caching and prediction mechanisms, allowing for the preloading or forecasting of potential market events, enhancing real-time decision-making. Financial institutions, within compliance and privacy regulations, could also develop standardized multimodal data-sharing platforms, enabling faster and more transparent data exchange and improving real-time analytical capabilities.

6.3 Lack of Model Interpretability

The lack of interpretability is a persistent challenge in applying MLLMs to financial decision-making. For investors, particularly those without deep financial expertise, a strategy must offer high predictive accuracy and transparency in the decision-making process. The complexity of multimodal data fusion compounds this issue, as these models often integrate diverse data types, such as text (news, reports), images (financial charts), and audio. The intricate relationships between these modalities make it difficult to explain how each contributes to the final decision. For instance, it is often unclear whether a recommendation is influenced more by a specific news event or historical stock performance. Moreover, the 'black box' nature of deep learning models, with their vast number of parameters, makes it challenging for users to understand the internal logic behind decisions. This lack of transparency can undermine trust, as investors may find it difficult to grasp how a particular recommendation was derived, especially when multiple data modalities are involved.

Future research could focus on developing new tools to improve the interpretability of multimodal models, allowing for clearer insights into how individual data types contribute to investment strategies. For example, visualization techniques could be used to highlight which trends in financial charts, combined with specific market news, influenced the model's recommendation, offering investors a clearer understanding of the decision-making process. Additionally, future multimodal investment models could be part of a human-machine collaboration framework, where machine models generate initial suggestions, and financial experts review and augment these outputs with further explanations. This hybrid approach would not only enhance interpretability but also build greater trust in the model's recommendations.

6.4 System Integration

Although various efforts are underway to tackle different aspects of multimodal large language models in the financial domain, a unified solution that comprehensively addresses the entire financial product pipeline has yet to be developed. Currently, no single system can effectively resolve all the associated challenges. From an application perspective, institutions prefer all-in-one solutions to improve operational efficiency. Similarly, from a research standpoint, establishing a unified framework would ensure consistency across specific financial domains, benefiting both industry and academic pursuits.

7 Conclusion

In this paper, we reviewed the applications of MLLMs within the financial sector. We explored their use in customer service, investment research, and compliance, demonstrating how MLLMs outperform traditional models in handling diverse data types such as text, images, and time series. We also discussed key challenges, including data scarcity, real-time processing limitations, and model interpretability. By addressing these challenges and identifying promising directions for future work, we aim to inspire further advancements and practical applications of MLLMs that meet the evolving needs of the financial industry. Although the application of multimodal large language models in the financial sector is gradually increasing, the relevant literature remains insufficient, indicating that there is still ample research space to explore in this direction.

Limitations

Although this paper provides a comprehensive and in-depth study of the development of multimodal large models in the financial domain from a scenario perspective, there are still certain limitations. Due to the complex nature of real-world applications in finance and the strong generalization capabilities of most multimodal large models, the method of categorizing their applications based on just three scenarios is not detailed enough. Therefore, how to conduct a more targeted and detailed investigation into complex financial scenarios remains a challenge. In the future, we will delve deeper into the research of financial multimodal large models, aiming to classify and summarize them from a more nuanced perspective, identify existing issues and development directions in the financial domain, and contribute to the research community.

References

Shivom Aggarwal, Shourya Mehra, and Pritha Mitra. 2023. Multi-purpose nlp chatbot: Design, methodology & conclusion. *arXiv preprint arXiv:2310.08977*.

Doron Avramov, Guy Kaplanski, and Avanidhar Subrahmanyam. 2022. Postfundamentals price drift in capital markets: A regression regularization perspective. *Management Science*, 68(10):7658–7681.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *arXiv preprint arXiv:2402.10986*.

Anjanava Biswas and Wrick Talukdar. 2024. Finembeddiff: A cost-effective approach of classifying financial documents with vector sampling using multi-modal embedding models. *arXiv preprint arXiv:2406.01618*.

Yupeng Cao, Zhi Chen, Qingyun Pei, Prashant Kumar, K.P. Subbalakshmi, and Papa Momar Ndiaye. 2024a. Ecc analyzer: Extract trading signal from earnings conference calls using large language model for stock performance prediction. *arXiv preprint arXiv:2404.18470v1*.

Yupeng Cao, Zhi Chen, Qingyun Pei, Fabrizio Dimino, Lorenzo Ausiello, Prashant Kumar, KP Subbalakshmi, and Papa Momar Ndiaye. 2024b. Risklabs: Predicting financial risk using large language model based on multi-sources data. *arXiv preprint arXiv:2404.07452*.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024a. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024b. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26540–26550.

Luyang Chen, Markus Pelger, and Jason Zhu. 2020. Deep learning in asset pricing, ssrn scholarly paper id 3350138. *Social Science Research Network, Rochester, NY*.

Zhixuan Chu, Huaiyu Guo, Xinyuan Zhou, Yijia Wang, Fei Yu, Hong Chen, Wanqing Xu, Xin Lu, Qing Cui, Longfei Li, et al. 2023. Data-centric financial large language models. *arXiv preprint arXiv:2310.17784*.

Lin William Cong, Ke Tang, Jingyuan Wang, and Yang Zhang. 2020. Alphaportfolio for investment and economically interpretable ai. *Available at SSRN*, 3554486.

Josiel Delgadillo, Johnson Kinyua, and Charles Mutigwe. 2024. Finsosent: Advancing financial market sentiment analysis through pretrained large language models. In *big data and cognitive computing*, volume 8.

Yayue Deng, Mohan Xu, and Yao Tang. 2024. Fmpaf: How do fed chairs affect the financial market? a fine-grained monetary policy analysis framework on their language. *arXiv preprint arXiv:2403.06115*.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024a. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.

Yuyang Ding, Juntao Li, Pinzheng Wang, Zecheng Tang, Bowen Yan, and Min Zhang. 2024b. Rethinking negative instances for generative named entity recognition. *arXiv preprint arXiv:2402.16602*.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, pages 6491–6499. ACM.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Harrison Hong, Terence Lim, and Jeremy C Stein. 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *The Journal of finance*, 55(1):265–295.

Zengyi Huang, Chang Che, Haotian Zheng, and Chen Li. 2024. Research on generative artificial intelligence for virtual financial robo-advisor. *Academic Journal of Science and Technology*, 10(1):74–80.

Samyak Jain, Parth Chhabra, Atula Tejaswi Neerkaje, Puneet Mathur, Ramit Sawhney, Shivam Agarwal, Preslav Nakov, Sudheer Chava, and Dinesh Manocha. 2024. Saliency-aware interpolative augmentation for multimodal financial prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14285–14297.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Carolin Kaiser, Rene Schallner, Vladimir Manewitsch, and Emma Fenne. 2024. The role of eye contact in human-robot interaction: Trust and decision-making in virtual financial advisory.

Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. 2019. Predicting returns with text data. Technical report, National Bureau of Economic Research.

Hyungjin Ko and Jaewook Lee. 2024. Can chatgpt improve investment decisions? from a portfolio management perspective. *Finance Research Letters*, 64:105433.

Chung-Hong Lee, Hsin-Chang Yang, Xuan-Qi Su, and Yao-Xiang Tang. 2022. A multimodal affective sensing model for constructing a personality-based financial advisor system. *Applied Sciences*, 12(19).

Martin Lettau and Sydney Ludvigson. 2001. Consumption, aggregate wealth, and expected stock returns. *the Journal of Finance*, 56(3):815–849.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023a. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*.

Yang Li, Yangyang Yu, Haohang Li, Zhi Chen, and Khaldoun Khashanah. 2023b. Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance. *arXiv preprint arXiv:2309.03736*.

Lezhi Li and Ting-Yu Chang. 2024. Multimodal gen-ai for fundamental investment research. *arXiv preprint arXiv:2401.06164*.

Xiaohui Victor Li and Francesco Sanna Passino. 2024. Findkg: Dynamic knowledge graphs with large language models for detecting global trends in financial markets. *arXiv preprint arXiv:2407.10909*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023c. Large language models in finance: A survey. In *4th ACM International Conference on AI in Finance (ICAIF '23)*, pages 1–9, Brooklyn, NY, USA. ACM.

Ahmed Masry and Amir Hajian. 2024. Longfin: A multimodal document understanding model for long financial domain documents. *Preprint*, arXiv:2401.15050.

Yuqi Nie et al. 2024. A survey of large language models for financial applications: Progress, prospects, and challenges. *arXiv preprint arXiv:2406.11903*.

Kun Ouyang, Yi Liu, Shicheng Li, Ruihan Bao, Keiko Harimoto, and Xu Sun. 2024. Modal-adaptive knowledge-enhanced graph-based financial prediction from monetary policy conference calls with llm. *arXiv preprint arXiv:2403.16055*.

L'uboš Pástor. 2000. Portfolio selection and asset pricing models. *The Journal of Finance*, 55(1):179–223.

L'uboš Pástor and Robert F Stambaugh. 2000. Comparing asset pricing models: an investment perspective. *Journal of Financial Economics*, 56(3):335–381.

Kellin Pelrine, Mohammad Taufeeque, Michał Zając, Euan McLean, and Adam Gleave. 2024. Exploiting novel gpt-4 apis. *arXiv preprint arXiv:2312.14302*.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2023. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *arXiv preprint arXiv:2309.08793*.

Stephen A Ross. 2013. The arbitrage theory of capital asset pricing. In *Handbook of the fundamentals of financial decision making: Part I*, pages 11–30. World Scientific.

Hardik Srivastava, Sneha Sunil, K. Shantha Kumari, and P. Kanmani. 2023. Multi-modal sentiment analysis using text and audio for customer support centers. In *Proceedings of ICACTCE'23 — The International Conference on Advances in Communication Technology and Computer Engineering*, pages 491–506, Cham. Springer Nature Switzerland.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7):4550.

Gaurish Thakkar, Sherzod Hakimov, and Marko Tadić. 2024. M2sa: Multimodal and multilingual model for sentiment analysis of tweets. *arXiv preprint arXiv:2404.01753*.

Xiangpeng Wan, Haicheng Deng, Kai Zou, and Shiqi Xu. 2024. Enhancing the efficiency and accuracy of underlying asset reviews in structured finance: The application of multi-agent framework. *arXiv preprint arXiv:2405.04294*.

Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023. Finvis-gpt: A multimodal large language model for financial chart analysis. *arXiv preprint arXiv:2308.01430*.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. 2024. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*.

Christopher Wimmer and Navid Rekabsaz. 2023. Leveraging vision-language models for granular market change prediction. *arXiv preprint arXiv:2301.10166*.

Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, et al. 2024. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*.

Ziwei Xu, Hiroya Takamura, and Ryutaro Ichise. 2024. A framework to construct financial causality knowledge graph from text. In *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pages 57–64. IEEE.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W Suchow, Rong Liu, Zhenyu Cui, Denghui Zhang, et al. 2024a. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *arXiv preprint arXiv:2407.06567*.

Guanyuan Yu, Xv Wang, Qing Li, and CYu Zhao. 2024b. Fusing llms and kgs for formal causal reasoning behind financial risk contagion. *arXiv preprint arXiv:2407.17190*.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2024. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*.

Han Zhang, Xinyu Wang, Junxiu Liu, Lei Zhang, and Lixia Ji. 2023. Chinese named entity recognition method for the finance domain based on enhanced features and pretrained language models. *Information Sciences*, 625:385–400.

Yuzhe Zhang and Hong Zhang. 2023. Finbert–mrc: financial named entity recognition using bert under the machine reading comprehension paradigm. *Neural Processing Letters*, 55(6):7393–7413.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. 2024a. Finagent: A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *arXiv preprint arXiv:2402.18485*.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. 2024b. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4314–4325.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. 2024. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

A Example Appendix