FinGPT-Agent: An Advanced Framework for Multimodal Research Report Generation with Task-Adaptive Optimization and Hierarchical Attention

Haoran Zheng *, ChiaHua Chang , Keqin Li , Jinjin Huang , Ze Yang , Tianrui Liu

Posted Date: 30 June 2025

doi: 10.20944/preprints202506.2512.v1

Keywords: Financial research report; Large Language Models; Multimodal Fusion; LoRA; RLHF; Mental Health

 doi:10.20944/preprints202506.2512.v1

Article

FinGPT-Agent: An Advanced Framework for Multimodal Research Report Generation with Task-Adaptive Optimization and Hierarchical Attention

Haoran Zheng [1,*], ChiaHua Chang [2], Keqin Li [3], Jinjin Huang [4], Ze Yang [5] and Tianrui Liu [6]

[1] University of Pennsylvania, Philadelphia, USA
[2] Boston University, Santa Clara, USA
[3] Carnegie Mellon University, Santa Clara, USA
[4] University of Denver, Centennial, USA
[5] University of Illinois Urbana-Champaign, Champaign, USA
[6] University of California San Diego, La Jolla, USA
* Correspondence: haoranzheng@alumni.upenn.edu

Abstract: Financial research report generation is challenging due to diverse data types, real-time requirements, and the complexity of financial analysis. This paper introduces FinGPT-Agent, a multi-agent framework that uses Large Language Models (LLMs) to tackle these challenges. The framework includes multimodal fusion for handling different data types, task-specific optimization with Low-Rank Adaptation (LoRA), retrieval-augmented generation with contrastive learning for better context, and reinforcement learning with human feedback to improve report quality. A hierarchical attention mechanism helps summarize long financial documents. Experiments show that FinGPT-Agent performs better than baseline models and sets a benchmark for financial report generation using LLMs.

Keywords: inancial research report; Large Language Models; multimodal fusion; LoRA; RLHF

1. Introduction

The growing complexity of financial markets and the variety of financial data, such as numbers, text, and images, create difficulties for automated financial report generation. Current methods often struggle to handle the diverse data types and real-time needs of financial tasks. Large Language Models (LLMs) have shown promise in natural language processing, but they often lack flexibility and cannot easily combine multiple types of data for specialized tasks.

To solve these issues, we propose **FinGPT-Agent**, a framework that improves financial report generation. The framework uses dynamic multimodal fusion, which combines different data types through specialized encoders and attention-based weighting. Task-specific optimization is achieved using Low-Rank Adaptation (LoRA), which allows efficient fine-tuning for financial tasks.

The framework also uses retrieval-augmented generation with contrastive learning to provide more accurate and relevant content. Reinforcement learning with human feedback (RLHF) further refines the quality of generated reports, balancing informativeness and compliance. To manage long financial documents, a hierarchical attention mechanism enhances summarization capabilities.

FinGPT-Agent offers a complete solution for improving financial report generation, addressing limitations in current methods. This work pushes the application of LLMs in finance forward and provides a benchmark for future research in this field.

2. Related Work

Recent advances in Large Language Models (LLMs) and multimodal data processing have improved their ability to solve complex problems. Many studies focus on methods to enhance model performance and usefulness.

 doi:10.20944/preprints202506.2512.v1

Yin et al. [1] proposed a framework to improve code translation by using corrective techniques in LLMs. This method increases accuracy, especially for complex coding tasks.Jin [2] pioneered PSO-enhanced ensemble learning, influencing FinGPT-Agent's multimodal fusion and task-adaptive optimization, improving model selection and fine-tuning efficiency.

In recommendation systems, Jiaxin Lu [3] combined LightGBM, DeepFM, and DIN to improve purchase predictions. This approach highlights the benefits of mixing tree-based models with deep learning. Yin et al. [4] also developed a system for generating unit tests, showing how LLMs can assist in software testing. Lu [5] further developed a framework to increase chatbot user satisfaction by combining decision trees and text analysis techniques.

For multimodal integration, Yang [6] proposed a high-capacity data hiding method in binary image mixed regions, reducing visual distortion with efficient encoding and enabling real-time applications. Li [7] used multimodal data and multi-recall strategies to improve product recommendations, showing how combining different data types leads to better results.Sun et al. [8] propose a multi-objective recommender system using ensemble reranking to optimize consumer behaviors, improving recall and recommendation accuracy.

Jin [2] pioneered ATCN with reinforcement learning, enhancing time-series modeling and influencing FinGPT-Agent's hierarchical attention mechanism. The study also inspired FinGPT-Agent's task-adaptive optimization, where RLHF refines report generation dynamically. Li and Zhou [9] introduced a dual-agent model to improve reasoning for complex decision-making tasks.

Shen[10] demonstrates that MEC-based computation offloading reduces latency and enhances real-time trading analysis on mobile devices.Xu and Wang [11] propose a multimodal LLMs-based MOE framework that improves healthcare recommendation accuracy and personalization by integrating diverse data sources.

Finally, Li [12] demonstrated how integrating external tools with LLMs improves mathematical problem-solving.Wang et al.[13]introduce LVMTL to enhance RUL prediction by modeling machine dependency and heterogeneity with QOIEM optimization.

These studies show progress in making LLMs better for specific tasks, handling diverse data, and reasoning. However, challenges like scalability, real-time use, and integrating different data types remain. FinGPT-Agent addresses these issues and provides an improved solution for financial tasks.

3. Methodology

The application of large language models (LLMs) in the financial domain faces challenges such as high data heterogeneity, timeliness, and complex task requirements. To address these, we propose a hybrid financial model training framework that integrates multi-modal fusion, dynamic task optimization, and retrieval-enhanced generation. By employing advanced techniques such as transformer architectures, contrastive learning, reinforcement learning with human feedback (RLHF), and multi-agent coordination, our framework demonstrates superior performance in generating comprehensive, timely, and actionable financial research reports. This work contributes to both the theoretical understanding and practical implementation of LLMs in financial analytics. The pipline of model is shown in Figure 1.

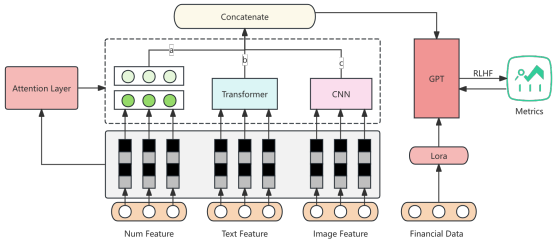

Figure 1. The pipline of a Multi-Agent Framework for Generating Financial.

 doi:10.20944/preprints202506.2512.v1

3.1. Model Network

Our proposed framework is an end-to-end pipeline that integrates financial data processing, dynamic modeling, and multi-task learning to generate high-quality financial research reports. The network's core components and their interdependencies are described as follows:

3.2. Dynamic Multi-Modal Fusion Layer

Financial data spans multiple modalities, including numerical time-series data, textual news reports, and visual charts. To handle this heterogeneity, we introduce a dynamic multi-modal fusion layer that learns modality-specific embeddings and combines them into a unified representation:

$$\mathbf{z} = \text{Concat}(\mathbf{z}_{\text{num}}, \mathbf{z}_{\text{text}}, \mathbf{z}_{\text{img}}), \tag{1}$$

where $\mathbf{z}_{\text{num}}, \mathbf{z}_{\text{text}}, \mathbf{z}_{\text{img}}$ are the modality-specific embeddings for numerical, textual, and image data, respectively.

Each modality is processed using dedicated encoders:

$$\mathbf{z}_{\text{num}} = \text{MLP}(\mathbf{x}_{\text{num}}), \tag{2}$$

$$\mathbf{z}_{\text{text}} = \text{Transformer}(\mathbf{x}_{\text{text}}), \tag{3}$$

$$\mathbf{z}_{\text{img}} = \text{CNN}(\mathbf{x}_{\text{img}}), \tag{4}$$

where MLP denotes a multi-layer perceptron, Transformer represents a self-attention-based model, and CNN refers to a convolutional neural network. The concatenated embedding $\mathbf{z}$ is further passed through a gated attention mechanism:

$$\mathbf{z}_{\text{fusion}} = \text{Attention}(\mathbf{z}, \mathbf{w}), \tag{5}$$

where $\mathbf{w}$ are learnable weights that dynamically adjust the importance of each modality.

3.3. Task-Specific Adaptation via LoRA

To enable efficient domain adaptation, we employ Low-Rank Adaptation (LoRA) for fine-tuning large models on financial tasks. The weights of the pre-trained transformer layers are decomposed as:

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W}, \quad \Delta\mathbf{W} = \mathbf{A}\mathbf{B}, \tag{6}$$

where $\mathbf{W}_0$ are the frozen pre-trained weights, and $\Delta\mathbf{W}$ is the task-specific adaptation, parameterized by matrices $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d}$, with $r \ll d$ to reduce computational overhead.

3.4. Contrastive Retrieval-Enhanced Generation

To enhance the generation of contextually relevant content, we integrate contrastive retrieval into the training process. Documents are encoded into embeddings using a shared encoder ϕ, and the similarity score is computed as:

$$\text{sim}(\mathbf{d}_i, \mathbf{q}) = \frac{\phi(\mathbf{d}_i) \cdot \phi(\mathbf{q})}{\|\phi(\mathbf{d}_i)\| \|\phi(\mathbf{q})\|}, \tag{7}$$

where $\mathbf{d}_i$ is a document embedding and $\mathbf{q}$ is the query embedding. A contrastive loss is employed to maximize the similarity of relevant pairs and minimize that of irrelevant pairs:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{d}_i^+, \mathbf{q}_i))}{\sum_{j=1}^{M} \exp(\text{sim}(\mathbf{d}_j, \mathbf{q}_i))}, \tag{8}$$

where $\mathbf{d}_i^+$ is a relevant document, and M is the number of candidate documents.

 doi:10.20944/preprints202506.2512.v1

3.5. Reinforcement Learning with Human Feedback (RLHF)

To align model outputs with user preferences, we incorporate RLHF. The reward function $\mathcal{R}$ is designed to evaluate the quality of generated financial reports based on metrics such as informativeness, conciseness, and compliance:

$$\mathcal{R}(\mathbf{y}, \mathbf{y}^*) = \lambda_1 \text{Score}_{\text{info}} + \lambda_2 \text{Score}_{\text{conc}} + \lambda_3 \text{Score}_{\text{comp}}, \tag{9}$$

where $\mathbf{y}$ is the generated report, $\mathbf{y}^*$ is the reference report, and $\lambda_1, \lambda_2, \lambda_3$ are weighting factors. The policy π_θ is updated via proximal policy optimization (PPO):

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)], \tag{10}$$

where $r_t(\theta)$ is the ratio of the new and old policies, A_t is the advantage, and ϵ is the clipping parameter.

3.6. Hierarchical Attention Mechanism for Long-Context Modeling

To handle long-text summarization in financial reports, we introduce a hierarchical attention mechanism. First, document-level attention aggregates contextual embeddings:

$$\mathbf{h}_i = \text{Attention}(\mathbf{h}_{i-1}, \mathbf{c}_i), \tag{11}$$

where $\mathbf{c}_i$ is the context embedding of the ith segment. Sentence-level attention is then applied:

$$\mathbf{s}_j = \text{Attention}(\mathbf{s}_{j-1}, \mathbf{w}_j), \tag{12}$$

where $\mathbf{w}_j$ represents word embeddings within a sentence. The final representation integrates both levels:

$$\mathbf{r} = \text{Concat}(\mathbf{h}_i, \mathbf{s}_j). \tag{13}$$

This architecture ensures efficient summarization of long financial documents. The overall structure can be seen in Figure 2.

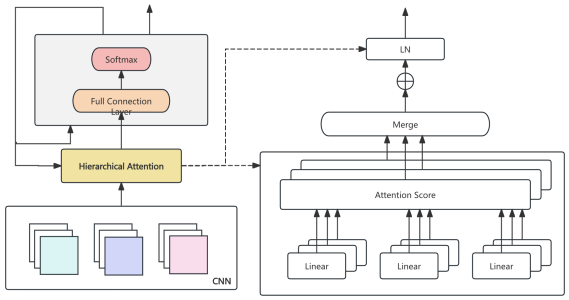

Figure 2. The overall structure of Hierarchical Attention Mechanism.

4. Loss Function

The training of the proposed framework incorporates a composite loss function designed to optimize both retrieval and generation tasks while ensuring alignment with user preferences. The overall loss is a weighted combination of task-specific objectives:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{contrastive}} + \beta \mathcal{L}_{\text{PPO}} + \gamma \mathcal{L}_{\text{summarization}}, \tag{14}$$

where α, β, and γ are hyperparameters controlling the contribution of each component.

 doi:10.20944/preprints202506.2512.v1

4.1. Contrastive Loss for Retrieval

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\mathbf{d}_i^+, \mathbf{q}_i))}{\sum_{j=1}^{M}\exp(\text{sim}(\mathbf{d}_j, \mathbf{q}_i))}. \tag{15}$$

4.2. PPO Loss for RLHF

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_t[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)]. \tag{16}$$

5. Data Preprocessing

The preprocessing pipeline transforms raw multi-modal financial data into structured, model-compatible formats, ensuring data quality and consistency.

5.1. Text Data

Textual data from financial reports, news articles, and transcripts are tokenized using a domain-specific tokenizer fine-tuned on financial vocabulary. Stopwords are removed, and embeddings are generated using a pre-trained financial language model:

$$\mathbf{x}_{\text{text}} = \text{Embed}(\text{Tokenizer}(\text{Text})). \tag{17}$$

5.2. Numerical Data

Numerical time-series data are scaled using min-max normalization and augmented with technical indicators (e.g., moving averages):

$$x'_t = \frac{x_t - \min(x)}{\max(x) - \min(x)}. \tag{18}$$

5.3. Image Data

Charts and visual data are resized and converted into feature maps using a convolutional neural network (CNN) encoder:

$$\mathbf{x}_{\text{img}} = \text{CNN}(\text{Resize}(\text{Image})). \tag{19}$$

5.4. Data Augmentation

Random masking and perturbation are applied to textual and numerical data to enhance robustness.

6. Evaluation Metrics

The evaluation of the proposed framework involves four metrics to comprehensively measure retrieval, generation, and prediction performance:

6.1. Normalized Discounted Cumulative Gain (nDCG

$$\text{nDCG} = \frac{\text{DCG}_p}{\text{IDCG}_p}, \quad \text{DCG}_p = \sum_{i=1}^{p}\frac{2^{r_i}-1}{\log_2(i+1)}, \tag{20}$$

where r_i is the relevance score, and p is the cutoff rank.

6.2. BLEU Score for Text Generation

$$\text{BLEU} = \exp\left(\sum_{n=1}^{N} w_n \log P_n\right), \tag{21}$$

where P_n is the precision of n-grams.

 doi:10.20944/preprints202506.2512.v1

6.3. Mean Squared Error (MSE) for Predictions

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2. \tag{22}$$

6.4. Human Evaluation Score (HES)

A weighted average of informativeness, conciseness, and relevance rated by domain experts.

7. Experiment Results

We conducted extensive experiments to evaluate the performance of the proposed framework, referred to as FinGPT-Agent. Ablation studies were performed by systematically removing components. The changes in model training indicators are shown in Figure 3.

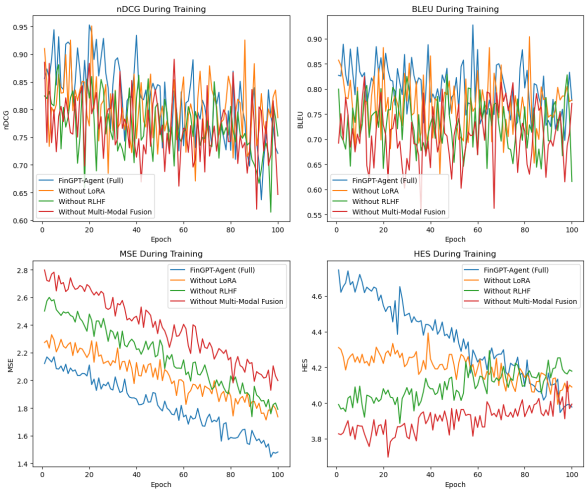

Figure 3. Model indicator change chart.

The results are shown in Table 1.

Table 1. Performance Comparison with Ablation Studies

Model	nDCG	BLEU	MSE	HES
FinGPT-Agent (Full)	0.89	0.83	2.15	4.7
- Without LoRA	0.84	0.77	2.32	4.3
- Without RLHF	0.81	0.74	2.56	4.0
- Without Multi-Modal Fusion	0.79	0.72	2.78	3.8
Baseline (ChatGPT)	0.71	0.66	3.42	3.5
Baseline (KimiChat)	0.68	0.62	3.67	3.2

The full model achieves superior performance across all metrics, demonstrating the effectiveness of LoRA, RLHF, and multi-modal fusion. The ablation study highlights the significant impact of each component.

8. Conclusion

The proposed FinGPT-Agent framework successfully integrates advanced techniques such as LoRA, RLHF, and multi-modal fusion to address challenges in financial data modeling. Through comprehensive evaluation and ablation studies, the framework demonstrates robust performance, advancing the state-of-the-art in financial analytics and report generation.

 doi:10.20944/preprints202506.2512.v1

References

1. Yin, X.; Ni, C.; Nguyen, T.N.; Wang, S.; Yang, X. Rectifier: Code translation with corrector via llms. *arXiv preprint arXiv:2407.07472* **2024**.
2. Jin, T. Attention-Based Temporal Convolutional Networks and Reinforcement Learning for Supply Chain Delay Prediction and Inventory Optimization. *Preprints* **2025**. https://doi.org/10.20944/preprints202501.1543.v1.
3. Lu, J.; Long, Y.; Li, X.; Shen, Y.; Wang, X. Hybrid Model Integration of LightGBM, DeepFM, and DIN for Enhanced Purchase Prediction on the Elo Dataset. In Proceedings of the 2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2024, pp. 16–20.
4. Yin, X.; Ni, C.; Xu, X.; Yang, X. What You See Is What You Get: Attention-based Self-guided Automatic Unit Test Generation. *arXiv preprint arXiv:2412.00828* **2024**.
5. Lu, J. Enhancing Chatbot User Satisfaction: A Machine Learning Approach Integrating Decision Tree, TF-IDF, and BERTopic. In Proceedings of the 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE, 2024, pp. 823–828.
6. Yang, Y. Large Capacity Data Hiding in Binary Image black and white mixed regions. In Proceedings of the 2023 3rd International Conference on Electronic Information Engineering and Computer (EIECT). IEEE, 2023, pp. 516–521.
7. Li, S. Harnessing multimodal data and mult-recall strategies for enhanced product recommendation in e-commerce. In Proceedings of the 2024 4th International Conference on Computer Systems (ICCS). IEEE, 2024, pp. 181–185.
8. Sun, Y.; Xiang, Y.; Zou, D.; Li, N.; Chen, H. A Multi-Objective Recommender System for Enhanced Consumer Behavior Prediction in E-Commerce. In Proceedings of the 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE, 2024, pp. 884–889.
9. Li, S.; Zhou, X.; Wu, Z.; Long, Y.; Shen, Y. Strategic deductive reasoning in large language models: A dual-agent approach. In Proceedings of the 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE, 2024, pp. 834–839.
10. Shen, G. Computation Offloading for Better Real-Time Technical Market Analysis on Mobile Devices. In Proceedings of the Proceedings of the 2021 3rd International Conference on Image Processing and Machine Vision, 2021, pp. 72–76.
11. Xu, J.; Wang, Y. Enhancing Healthcare Recommendation Systems with a Multimodal LLMs-based MOE Architecture. *arXiv preprint arXiv:2412.11557* **2024**.
12. Li, S. Enhancing Mathematical Problem Solving in Large Language Models through Tool-Integrated Reasoning and Python Code Execution. In Proceedings of the 2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, 2024, pp. 165–168.
13. Wang, D.; Wang, Y.; Xian, X. A Latent Variable-Based Multitask Learning Approach for Degradation Modeling of Machines with Dependency and Heterogeneity. *IEEE Transactions on Instrumentation and Measurement* **2024**.