Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

FinSphere: a real-time stock analysis agent with instruction-tuned large language models and domain-specific tool integration#

Shijie HAN§1,3, Jingshu ZHANG§2,3, Yiqing SHEN3,4, Kaiyuan YAN3, Hongguang LI‡3

[1]*Department of Industrial Engineering and Operations Research, Columbia University, New York 10027, USA*

[2]*School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China*

[3]*JF SmartInvest Holdings Ltd., Shanghai 201702, China*

[4]*Department of Computer Science, Johns Hopkins University, Baltimore 21218, USA*

E-mail: sh4460@columbia.edu; zhangjingshu@mail.shufe.edu.cn; yshen92@jhu.edu; yankaiyuani@163.com; harvey2@mail.ustc.edu.cn

Received June 15, 2025; Revision accepted Aug. 25, 2025; Crosschecked Sept. 29, 2025; Published online Nov. 6, 2025

Abstract: Current financial large language models (FinLLMs) exhibit two major limitations: the absence of standardized evaluation metrics for stock analysis quality and insufficient analytical depth. We address these limitations with two contributions. First, we introduce AnalyScore, a systematic framework for evaluating the quality of stock analysis. Second, we construct Stocksis, an expert-curated dataset designed to enhance the financial analysis capabilities of large language models (LLMs). Building on Stocksis, together with a novel integration framework and quantitative tools, we develop FinSphere, an artificial intelligence (AI) agent that generates professional-grade stock analysis reports. Evaluations with AnalyScore show that FinSphere consistently surpasses general-purpose LLMs, domain-specific FinLLMs, and existing agent-based systems, even when the latter are enhanced with real-time data access and few-shot guidance. The findings highlight FinSphere's significant advantages in analytical quality and real-world applicability.

Key words: Large language model (LLM); Instruction-tuned financial LLM; Real-time stock analysis; Evaluation framework and dataset

https://doi.org/10.1631/FITEE.2500414 **CLC number:** TP391

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in comprehending and processing natural language, extending their influence across various domains, including finance (Li YH et al., 2023). By leveraging their language comprehension capabilities, these models have exhibited exceptional performance in various financial applications, including sentiment analysis (Zhang BY et al., 2023; Liu CH et al., 2024) and information extraction from unstructured financial texts (Huang et al., 2023; Li HX et al., 2025). The advent of finance-specific LLMs, such as FinBERT (Liu Z et al., 2020; Yang Y et al., 2020), BloombergGPT (Wu et al., 2023), and PIXIU (Xie et al., 2023), has further enhanced the capacity to process financial data effectively. These advancements have laid the

‡ Corresponding author
§ These two authors contributed equally to this work
# Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2500414) contains supplementary materials, which are available to authorized users
ORCID: Shijie HAN, https://orcid.org/0009-0003-3212-5625; Jingshu ZHANG, https://orcid.org/0009-0003-2233-8563; Hongguang LI, https://orcid.org/0009-0003-0625-8213

foundation for developing more sophisticated financial analysis tools and shifted how investors interact with market data (Krause, 2023; Nie et al., 2024). These artificial intelligence (AI)-powered systems have broadened access to professional financial insights, allowing retail investors to benefit from advanced analysis once reserved for institutions.

As LLM technology continues to evolve, there is a growing expectation for these models to handle more complex financial tasks, particularly in stock diagnosis and real-time financial analysis (Yang Y et al., 2023; Zhao et al., 2024). This has led to the development of tool-augmented agents that integrate LLMs' natural language understanding with specialized financial tools, significantly enhancing automated financial analysis and interactive decision support (Ding et al., 2024; Zhang WT et al., 2024). However, LLM-based financial analysis systems still face substantial challenges in effectively interpreting and using the outputs of these tools to generate high-quality analytical insights. Two primary obstacles are the absence of systematic evaluation frameworks to assess their performance in stock analysis and the lack of specialized datasets for fine-tuning LLMs' analytical reasoning capabilities. Moreover, existing research is constrained by LLMs' heavy reliance on historical data, such as GPT-4o's dependence on its pre-trained knowledge for generating responses (Bhat and Jain, 2024; Ni et al., 2024). This limitation in accessing and processing real-time financial data and domain-specific information restricts their ability to fully capture the dynamic and evolving nature of financial markets, posing a critical challenge for real-time financial analysis systems.

To address these limitations, we present three key contributions:

1. AnalyScore, a comprehensive evaluation framework designed to systematically assess the accuracy, relevance, and analytical depth of LLM-driven stock analysis;

2. Stocksis, a specialized dataset constructed by industry experts to enhance LLMs' capabilities in stock diagnosis and financial analysis;

3. FinSphere, a real-time financial analysis agent designed to produce accurate and insightful stock diagnostic reports on demand.

Our experiments demonstrate that FinSphere, by integrating real-time financial databases, specialized quantitative tools, and an instruction-tuned LLM optimized for financial analysis, significantly outperforms both general-purpose and domain-specific LLMs, as well as existing agent-based systems. This superior performance holds even when baseline models are augmented with real-time background information and few-shot prompting. This validates the effectiveness of our integrated approach to real-time financial analysis and stock market diagnostics.

2 Related works

LLMs have shown strong potential in financial tasks such as stock prediction, market analysis, and portfolio management (Bhat and Jain, 2024; Kim et al., 2024; Ni et al., 2024; Zhao et al., 2024). Domain-specific models, such as InvestLM (Yang Y et al., 2023) and GPT-InvestAR (Gupta, 2023), further highlight the benefits of financial instruction tuning. LLMs have also been applied to financial anomaly detection (Park, 2024) and financial statement understanding (Kim et al., 2024).

2.1 Financial datasets and evaluation

Existing datasets, such as FinQA (Chen ZY et al., 2021), TAT-QA (Zhu et al., 2021), and FLARE (Xie et al., 2023), focus on question–answering or numerical reasoning but lack support for comprehensive, real-time stock analysis. Others, including FinTextQA (Chen J et al., 2024), CFBenchmark (Lei et al., 2023), and FinanceBench (Islam et al., 2023), provide broader coverage but miss expert-annotated stock reports. Evaluation of financial text has typically relied on generic metrics, e.g., BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), while recent efforts such as FinEval (Guo et al., 2025) offer domain-adapted metrics but lack systematic expert-grounded scoring.

2.2 Instruction tuning and tool-augmented agents

Recent advances integrate financial LLMs with domain-specific tools, such as FinGPT (Yang HY et al., 2023), XBRL-Agent (Han et al., 2024), and FinRobot (Yang HY et al., 2024), improving task relevance and interactivity. However, most agent-style systems differ from FinSphere in three aspects. First, they rely on static or batch-updated data, limiting

real-time adaptability, whereas FinSphere integrates continuously updated databases. Second, their tool use is pre-defined, whereas FinSphere employs a decision-oriented chain-of-thought (CoT) to dynamically select from over 100 specialized tools. Third, they adopt general or lightly adapted LLMs, whereas FinSphere uses full-parameter instruction tuning on the expert-curated Stocksis dataset, achieving greater analytical depth and coherence. Together, these advances enable FinSphere's superior performance (Section 4).

3 AnalyScore and Stocksis

Stock market analysis is becoming increasingly complex, necessitating the integration of diverse data sources and sophisticated analytical approaches. While LLMs demonstrate potential in transforming financial analysis, there are two critical gaps in the current landscape: the absence of standardized evaluation frameworks for assessing the quality of AI-generated stock analysis and the lack of high-quality training data for developing LLMs' stock analysis capabilities. This section introduces two significant contributions to address these gaps: AnalyScore, a systematic evaluation framework for assessing stock analysis, and Stocksis, a comprehensive dataset specifically designed to enhance LLMs' stock analysis capabilities.

3.1 AnalyScore: a comprehensive evaluation framework for stock analysis reports

To address the lack of systematic evaluation standards in stock analysis, we introduce AnalyScore, a domain-specific framework co-developed with industry experts. It combines established financial assessment criteria with insights from LLM-generated content evaluation.

AnalyScore employs a structured scoring system across four dimensions, with a total score of 100:

1. Conclusion (20 scores), clarity, relevance, and personalization of investment recommendations;

2. Content (45 scores), depth, coherence, and professionalism of analytical reasoning;

3. Expression (15 scores), organization, fluency, and linguistic clarity;

4. Data (20 scores), breadth, accuracy, and effective use of supporting quantitative data.

This multi-dimensional evaluation ensures comprehensive coverage of both qualitative judgment and quantitative rigor. AnalyScore is fully presented in Table S1 in the supplementary materials, with the aim of promoting related research in academia and industry. Currently, AnalyScore is used exclusively by human experts to assess model-generated analysis. In future work, we aim to integrate AnalyScore into automatic evaluation pipelines by prompting LLMs to emulate expert scoring behaviors.

3.2 Stocksis: a high-quality dataset for enhancing LLMs' stock analysis

To evaluate LLMs' capabilities in stock analysis, we first assess GPT-4o's responses using AnalyScore (see Section 5.2), providing it with extensive background knowledge, including market data and quantitative indicators. Despite these supports, its outputs often exhibit reasoning inconsistencies, shallow financial insights, and occasional misinterpretations of market trends, revealing the challenges that LLMs struggle to synthesise complex financial information into coherent and actionable analysis. To address these limitations, we collaborate with industry experts to iteratively refine GPT-4o's outputs, correcting errors and enriching them with deeper reasoning. This expert-guided process results in Stocksis, a high-quality dataset that bridges automated generation and professional-grade analysis, providing structured supervision to improve LLM performance in real-world financial contexts.

A complete example is detailed in Table S2 in the supplementary materials. Each training sample consists of two key components:

1. Prompt with background information (input). A complete analytical prompt includes aggregated outputs from multiple quantitative analysis tools (averaging six tools per sample) as background information. Background information covers volume–price analysis, technical indicators, and other market metrics. Each prompt is rigorously crafted to guide the model in performing analytical tasks while leveraging the provided background information. The average length is 4000 words.

2. Expert-edited analysis (label). In-depth analytical reports respond to the prompt's requirements while effectively using the background information, averaging 3000 words per report. Due to the particularity of the stock analysis task, there is no standard answer to this task. Therefore, our industry experts

provide a high-quality reference analysis result for this task by evaluating the overall market, providing detailed reasons and demonstrating how to effectively interpret various quantitative indicators.

Here, we discuss the dataset collection and quality assurance. Stocksis is constructed through a structured pipeline grounded in our company's expertise in retail-oriented stock analysis. Data collection involves two main stages:

1. Prompt and background generation. Expert analysts select suitable quantitative tools for specific stock queries and craft prompts enriched with structured analytical output.

2. Analysis refinement. GPT-4o generates initial responses, which are then iteratively reviewed and improved by a panel of 10 senior analysts to ensure accuracy, coherence, and domain relevance. In practice, the most common types of expert corrections include: (1) restructuring the content to follow a clear introduction–analysis–conclusion format, (2) correcting factual or numerical hallucinations, especially specific market figures, (3) resolving logical inconsistencies (e.g., contradictions between different parts of the analysis), and (4) refining language expressions to enhance clarity, appropriateness, and professional tone. This refinement process spans more than three months to ensure high-quality outputs.

Stocksis addresses a key gap in financial NLP: the lack of datasets combining structured prompts with expert-refined reasoning. Unlike existing resources focused on price or sentiment data, it enables the fine-tuning of general-purpose LLMs to enhance their capability in structured, high-quality financial analysis for real-world applications.

4 FinSphere agent

This section details the architecture and operational mechanisms of FinSphere agent (Fig. 1), i.e., our advanced stock analysis agent.

4.1 Modular quantitative toolkits integrated with real-time financial data

A core advantage of FinSphere lies in its modular integration with a set of quantitative analysis toolkits, designed to support flexible and scalable financial reasoning. Each toolkit is implemented as

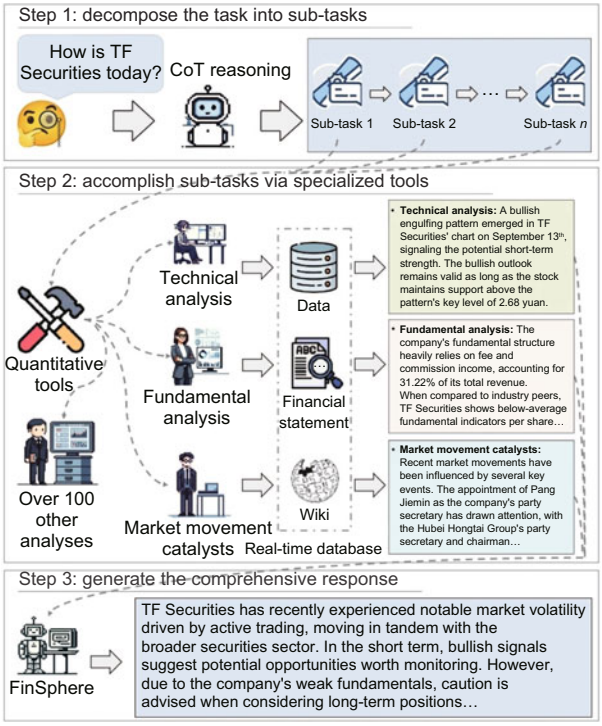

Fig. 1 Diagram of the overall workflow of the FinSphere agent. This details how different components interact to facilitate real-time stock analysis. TF: Tianfeng

an independent analysis module that interfaces directly with a continuously updated real-time financial database. This database provides comprehensive coverage of market data, including structured data (e.g., stock prices, trading volumes, and financial statements) and unstructured data (e.g., corporate disclosures, analyst commentary, and financial news).

Upon detecting the need for domain-specific computations, FinSphere dynamically invokes the appropriate toolkit module. The module automatically queries the latest relevant data, executes predefined quantitative routines—such as technical analysis signal extraction, fundamental analysis valuation modeling, or sentiment and market movement catalyst indexing—and returns context-aware results tailored to the user query. Over 100 quantitative analysis toolkits integrated into FinSphere can be broadly categorized into three functional types:

1. Market indicator tools. These tools cover real-time market data access and specialized market analysis modules (Lin, 2004). This category includes index quotation tools, Hong Kong stock quotation tools, fund quotation tools, A-shares quotation tools,

U.S. stock quotation tools, and national equities exchange and quotations (NEEQ) tools, as well as dedicated analysis toolkits for technical analysis, fundamental analysis, and market movement catalysts (together enabling FinSphere to support not only China's A-shares and the Hong Kong market, but also U.S. markets).

2. Information retrieval tools. These tools enable rapid access to relevant information via online general search, online financial search, local knowledge base vector retrieval, and stock-specific relationship graph search. Such tools facilitate the discovery and contextualization of both structured and unstructured financial information.

3. Product-oriented tools. These tools support enriched analysis through curated company expert opinions, articles, and videos, as well as proprietary product signaling tools. Together, these resources enhance interpretability and add qualitative insights to the quantitative outputs.

This modular design ensures that analytical responses remain accurate, interpretable, and reflective of the current market dynamics, while also allowing future extension with new toolkits or data modalities. Compared to static retrieval-based systems, FinSphere's architecture provides a principled path toward integrating quantitative financial reasoning into LLM agents.

4.2 Instruction tuning

We adopt full-parameter instruction tuning to adapt Qwen2-72B into a domain-specialized financial analysis agent. Using our expert-curated Stocksis dataset—comprising 5000 structured training pairs of quantitative tool outputs and corresponding expert-written analyses (see Section 3.2), we guide the model to learn not just domain knowledge but also the task format, reasoning flow, and response style required for professional stock reporting. Unlike parameter-efficient methods, full instruction tuning enables the model to internalize complex analytical patterns and generate coherent, multi-dimensional reports.

Training is performed on an NVIDIA 16×A100 GPU cluster with a 32 000-token context, using a language modeling objective with a learning rate of 1e-5, a batch size of 16, and 2 training epochs. We apply gradient clipping and mixed-precision training to ensure numerical stability and efficiency. Through this process, FinSphere acquires strong capabilities in synthesizing diverse financial signals, interpreting market data accurately, and responding with structured, human-level stock analyses tailored to real-world scenarios.

4.3 Overall workflow

FinSphere follows a structured, multi-stage workflow to generate real-time financial analyses. Upon receiving a user query, it first applies CoT reasoning to decompose the task into interpretable subtasks and determine which domain-specific quantitative tools are needed. These tools then independently access a real-time financial database to retrieve the latest market data, including technical indicators, capital flows, and fundamental metrics, each generating specialized outputs tailored to their analytical focus.

In the final stage, FinSphere's instruction-tuned LLM—trained on the Stocksis dataset—serves as an expert analyst. It synthesizes the multi-source outputs into a coherent, structured report aligned with professional financial standards. This integrated architecture enables FinSphere to combine the precision of automated quantitative analysis with the contextual depth of expert reasoning, ensuring that responses are analytically sound and up-to-date.

5 Evaluation

Given FinSphere's integration with real-time financial databases and proprietary quantitative tools, it possesses analytical capabilities that extend beyond those of general-purpose LLMs. Performance comparisons between FinSphere and general LLMs present inherent challenges, primarily due to the latter's inability to access real-time financial data and domain-specific information. For example, GPT-4o typically acknowledges its limitations with responses like "As an AI language model with knowledge cut-off in June 2024, I don't have access to real-time stock information." To demonstrate FinSphere's enhanced capabilities while ensuring a fair comparison, we have implemented a comprehensive experimental design.

5.1 Baseline

We evaluate three categories of models: (1) single LLMs, including proprietary (GPT-4o and

GPT-3.5), open-source (Qwen2-72B and DeepSeek-V3 (DeepSeek-AI, 2024)), and domain-specific models (InvestLM (Yang Y et al., 2023) and FinGPT (Yang HY et al., 2023)); (2) agent-based systems, including FinMem (Yu et al., 2024) and FinRobot (Yang HY et al., 2024); (3) our proposed FinSphere. All LLMs use CoT prompting with few-shot examples and relevant background information ("testing promp" in the supplementary materials), while agents receive simplified prompts similar to Stocksis's inputs. FinSphere is evaluated directly via user queries, using its integrated real-time data and tools. For all models, we set a maximum output length of 8000 tokens and a temperature of 0.5.

5.2 Performance analysis

The expert evaluation results (with detailed FinSphere and baseline model responses, expert ratings, and comments provided in the supplementary materials) presented in Table 1 demonstrate FinSphere's superior performance across all assessment dimensions, achieving an overall score of 70.88 out of 100. This significantly surpasses both traditional LLM-based approaches and other agent-based systems, with FinMem and GPT-4o following at 67.55 and 66.61, respectively. The evaluation reveals a clear performance hierarchy: Agent-based systems generally outperform standalone language models (except GPT-4o), while general-purpose LLMs show moderate performance, and domain-specific LLMs such as FinGPT demonstrate relatively limited capabilities (45.05). These results validate the effectiveness of FinSphere's integrated approach, which combines real-time data access, quantitative tools, and a Stocksis-tuned LLM, enabling more precise and insightful stock analysis.

5.3 Dimensional analysis and visualization

To further investigate the comparative strengths of FinSphere, we conduct a detailed analysis of its performance relative to two other leading agent-based systems, FinRobot and FinMem. The comparative visualization in Fig. 2 highlights performance differences across four critical dimensions: conclusion, content, expression, and data capabilities.

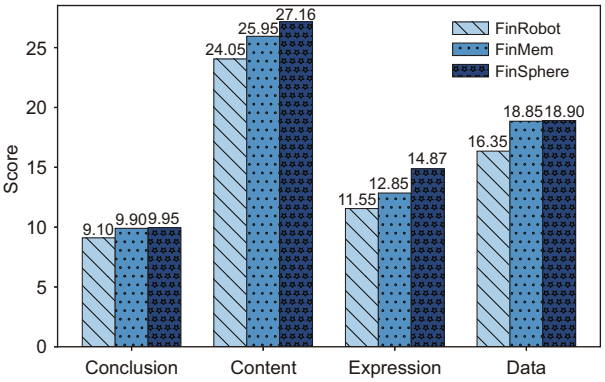

Fig. 2 Comparison of FinSphere and agent-based systems in various dimensions

From the visualization, FinSphere exhibits the highest scores across all dimensions. In the conclusion category, the three models perform relatively closely, with FinSphere slightly leading at 9.95, followed by FinMem at 9.90 and FinRobot at 9.10, demonstrating their robust ability to derive investment insights. However, in the content dimension, FinSphere shows a clear advantage, scoring

Table 1 Human experts use AnalyScore to evaluate 100 responses generated by eight models

Model	AnalyScore				Total score
	Conclusion (score: 20)	Content (score: 45)	Expression (score: 15)	Data (score: 20)	
GPT-4o	9.85	26.12	12.44	18.20	66.61
DeepSeek-V3	9.52	25.30	12.75	16.85	64.42
GPT-3.5	7.95	21.05	10.15	14.30	53.45
Qwen2-72B	8.15	22.55	10.55	14.95	56.20
InvestLM	8.40	23.10	11.25	15.75	58.50
FinGPT	6.80	18.55	8.95	10.75	45.05
FinRobot	9.10	24.05	11.55	16.35	61.05
FinMem	9.90	25.95	12.85	18.85	67.55
FinSphere	**9.95**	**27.16**	**14.87**	**18.90**	**70.88**

The scores shown are the averages across all evaluations. We disclosed 100 testing queries along with expert scores for FinSphere; see Table S3 in the supplementary materials for more details. Bold results represent the highest value of that dimension

27.16, significantly surpassing FinMem (25.95) and FinRobot (24.05), reflecting its greater analytical depth and content richness.

The most pronounced gap is observed in the expression dimension, where FinSphere achieves 14.87, noticeably higher than FinMem (12.85) and FinRobot (11.55). This highlights FinSphere's superior logical articulation in financial reporting. In terms of data utilization, FinSphere (18.90) and FinMem (18.85) exhibit comparable performance, both substantially outperforming FinRobot (16.35), reinforcing their accurate understanding and grasp of data in financial analysis.

On the contrary, one notable limitation of general-purpose LLMs is their heavy dependence on extensive in-context examples to generate accurate financial analyses. This results in a substantial increase in input token consumption, leading to higher operational costs for models such as GPT-4o and restricting the usability of small-context-window LLMs. In contrast, FinSphere's instruction-tuned architecture eliminates the need for verbose prompts, allowing it to generate high-quality outputs with significantly fewer input tokens.

The combined findings underscore FinSphere's state-of-the-art capabilities in stock analysis, driven by its robust data processing and structured analytical reasoning. These results further validate its advantage over both standalone LLMs and agent-based financial analysis systems.

5.4 Evaluation consistency

To assess the consistency of our human evaluation process, we use Kendall's tau rank correlation coefficient to measure agreement among annotators. We divide 40 industry experts into four groups of 10, with each group collaboratively generating a single consensus score for every model response. Within each group, experts first review the same set of model outputs individually, followed by a structured discussion to reconcile differences and agree on a final score, ensuring well-considered judgments over individual subjectivity. The final score for each model response is then obtained by averaging the consensus scores from the four groups.

To quantify agreement between groups, we rank all model-generated responses based on the assigned scores within each group, and compute Kendall's tau correlation coefficients for pairwise comparisons between groups. This analysis allows us to examine how consistently different groups ranked the LLM/agent responses in terms of relevance and quality. The correlation results in Table 2 represent the average Kendall's tau across 100 queries.

The results indicate a strong agreement across annotator groups, with Kendall's tau values ranging from 71.45% to 94.25%. The majority of pairwise group correlations are more than 80%, suggesting a high level of consistency in how different groups evaluate and rank model-generated responses. Notably, content value exhibits the highest variation in agreement across groups, with values spanning from 71.45% to 94.25%, while expression value and data value maintain relatively stable agreement levels, further reinforcing the reliability of the evaluation framework.

These findings suggest that despite potential subjectivity in human assessments, the evaluation process maintains a substantial level of consensus, validating its robustness. The strong Kendall's tau correlations confirm that annotators are able to systematically distinguish high-quality responses, ensuring that the evaluation framework accurately reflects model performance.

Table 2 Average Kendall's tau across 100 queries for different annotator groups

Groups	Kendall's tau rank correlation coefficient				Total (%)
	Conclusion (%)	Content (%)	Expression (%)	Data (%)	
1 and 2	79.36	93.77	88.30	84.97	78.90
1 and 3	73.90	71.45	91.65	85.03	87.70
1 and 4	79.51	94.25	90.81	75.31	74.55
2 and 3	81.59	77.61	83.12	80.80	77.28
2 and 4	85.30	73.49	77.30	79.16	81.40
3 and 4	89.63	74.99	82.86	84.81	73.16
Average	81.55	80.93	85.67	81.68	78.83

5.5 Ablation on training data scale

To investigate the impact of the training data scale on FinSphere's performance, we conduct an ablation study using different proportions of the Stocksis dataset. We fine-tune Qwen2-72B using 20%, 50%, 80%, and 100% of the 5000 data pairs while maintaining FinSphere's framework. The detailed evaluation results are provided in Fig. 3, which demonstrates a clear positive correlation between training data scale and model performance, with overall scores increasing from 58.90 (20%) to 70.88 (100%). Performance improvement shows a nonlinear pattern, with larger incremental gains observed at higher data volumes. These findings underscore the importance of comprehensive training data in achieving optimal performance, while also demonstrating the robustness of our framework, as FinSphere maintains satisfactory performance levels even with reduced training data.

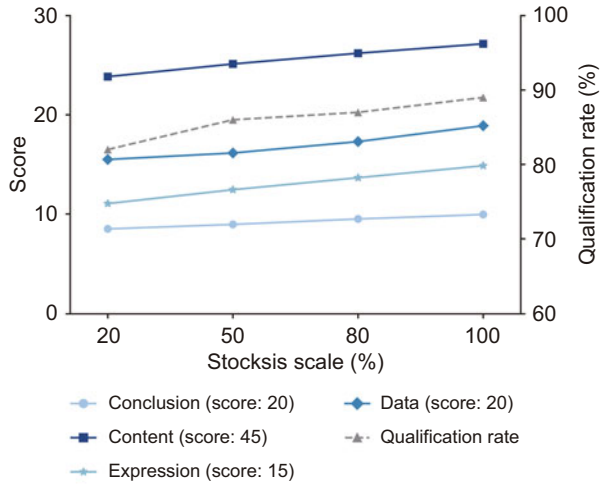

Fig. 3 Changes in scores of each sub-item as the Stocksis scale used for fine-tuning changes

5.6 Ablation on system modules

To further understand the contribution of each major component in FinSphere, we perform an additional ablation study that systematically disables one module at a time while keeping the others intact. We focus on three modules: (1) the decision CoT module, which enables the LLM to autonomously select the most relevant tools for a query; (2) the tool module, which provides domain-specific quantitative outputs from real-time financial data; (3) the instruction-tuned LLM (full fine-tuning on Stocksis), replaced by the base Qwen2-72B model without fine-tuning. Performance is evaluated with AnalyScore across 100 queries, as shown in Table 3.

The results indicate that each component plays a critical role in FinSphere's performance. Removing the decision CoT module results in a substantial drop in all dimensions, reflecting the importance of autonomous tool selection. Removing the tool module produces the most severe degradation, especially in the data dimension (score drops to zero), confirming that real-time quantitative tools are indispensable for factual accuracy and analytical depth. Replacing the instruction-tuned LLM with its base counterpart also leads to notable declines in content and expression, highlighting the gains from domain-specific instruction tuning.

6 Conclusions and future work

This paper introduces FinSphere, an innovative stock analysis agent that addresses critical gaps in the capabilities of LLMs for stock analysis. By integrating real-time financial databases, quantitative tools, and an instruction-tuned LLM, FinSphere demonstrates superior performance in generating comprehensive stock analyses. The development and release of Stocksis, a high-quality dataset for

Table 3 Ablation results for major FinSphere modules

Model	AnalyScore				Total score
	Conclusion (score: 20)	Content (score: 45)	Expression (score: 15)	Data (score: 20)	
FinSphere (full)	**9.95**	**27.16**	**14.87**	**18.90**	**70.88**
w/o decision CoT	6.85	19.20	9.78	12.17	48.00
w/o tool module	3.21	15.76	8.54	0.00	27.51
w/o instruction tuning	8.15	22.55	10.55	14.95	56.20

The scores are averaged across all evaluations using AnalyScore. Bold results represent the highest value of that dimension. w/o refers to without

enhancing LLMs' stock analysis capabilities, and AnalyScore, a systematic evaluation framework, provide valuable resources for advancing research in AI-powered financial analysis. Our experimental results indicate that FinSphere consistently outperforms general-purpose, domain-specific LLMs, and agent-based systems across multiple evaluation dimensions, highlighting the effectiveness of our integrated approach.

FinSphere's performance depends on the accuracy and availability of real-time financial data, which may impact analysis reliability. The AnalyScore framework still requires human validation, limiting full automation. Additionally, FinSphere may struggle with nuanced financial reasoning and novel market events beyond its training. Future work should focus on improving real-time adaptability, reducing reliance on curated data, and expanding domain coverage for broader financial applications.

Contributors

Shijie HAN conceived the research and drafted the paper. Jingshu ZHANG supervised the study and guided the execution. Yiqing SHEN, Kaiyuan YAN, and Hongguang LI contributed to the data collection, experiments, and paper revision. Shijie HAN and Jingshu ZHANG finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. Stocksis comprises 5000 meticulously curated training pairs, with some of them available in the open-source release for research and development purposes at https://github.com/KirkHan0920/Stocksis. As FinSphere, a powerful stock analysis agent developed by a stock investment advisory company, we have released a fully functional product demo that has been freely available to the public since December 2024. The demo can be accessed at https://uat-jiuzhang.techgp.cn/rjhy/jzmodelsInner/#/public-login.

References

Bhat R, Jain B, 2024. Stock price trend prediction using emotion analysis of financial headlines with distilled LLM model. Proc 17th Int Conf on Pervasive Technologies Related to Assistive Environments, p.67-73. https://doi.org/10.1145/3652037.3652076

Chen J, Zhou PL, Hua YN, et al., 2024. FinTextQA: a dataset for long-form financial question answering. Proc 62nd Annual Meeting of the Association for Computational Linguistics, p.6025-6047. https://doi.org/10.18653/v1/2024.acl-long.328

Chen ZY, Chen WH, Smiley C, et al., 2021. FinQA: a dataset of numerical reasoning over financial data. Proc Conf on Empirical Methods in Natural Language Processing, p.3697-3711. https://doi.org/10.18653/v1/2021.emnlp-main.300

DeepSeek-AI, 2024. DeepSeek-V3 technical report. https://arxiv.org/abs/2412.19437

Ding H, Li YH, Wang JH, et al., 2024. Large language model agent in financial trading: a survey. https://arxiv.org/abs/2408.06361

Guo X, Xia HT, Liu ZW, et al., 2025. FinEval: a Chinese financial domain knowledge evaluation benchmark for large language models. Proc Conf of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, p.6258-6292. https://doi.org/10.18653/v1/2025.naacl-long.318

Gupta U, 2023. GPT-InvestAR: enhancing stock investment strategies through annual report analysis with large language models. https://doi.org/10.48550/arXiv.2309.03079

Han SJ, Kang HQ, Jin B, et al., 2024. XBRL agent: leveraging large language models for financial report analysis. Proc 5th ACM Int Conf on AI in Finance, p.856-864. https://doi.org/10.1145/3677052.3698614

Huang AH, Wang H, Yang Y, 2023. FinBERT: a large language model for extracting information from financial text. *Contemporary Acc Res*, 40(2):806-841. https://doi.org/10.1111/1911-3846.12832

Islam P, Kannappan A, Kiela D, et al., 2023. FinanceBench: a new benchmark for financial question answering. https://doi.org/10.48550/arXiv.2311.11944

Kim A, Muhn M, Nikolaev V, 2024. Financial statement analysis with large language models. https://doi.org/10.48550/arXiv.2407.17866

Krause D, 2023. Large language models and generative AI in finance: an analysis of ChatGPT, Bard, and Bing AI. *SSRN Electr J*.

Lei Y, Li JT, Jiang M, et al., 2023. CFBenchmark: Chinese financial assistant benchmark for large language model. https://doi.org/10.48550/arXiv.2311.05812

Li HX, Gao HY, Wu CZ, et al., 2025. Extracting financial data from unstructured sources: leveraging large language models. *J Inform Syst*, 39(1):135-156. https://doi.org/10.2308/ISYS-2023-047

Li YH, Wang SF, Ding H, et al., 2023. Large language models in finance: a survey. Pro 4th ACM Int Conf on AI in Finance, p.374-382. https://doi.org/10.1145/3604237.3626869

Lin CY, 2004. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, p.74-81. https://aclanthology.org/W04-1013

Liu CH, Arulappan A, Naha R, et al., 2024. Large language models and sentiment analysis in financial markets: a review, datasets and case study. *IEEE Access*, 12:134041-134061. https://doi.org/10.1109/ACCESS.2024.3445413

Liu Z, Huang DG, Huang KY, et al., 2020. FinBERT: a pre-trained financial language representation model for financial text mining. Proc 29th Int Joint Conf on Artificial Intelligence, p.4513-4519. https://doi.org/10.24963/ijcai.2020/622

Ni HW, Meng SC, Chen XP, et al., 2024. Harnessing earnings reports for stock predictions: a QLoRA-enhanced LLM approach. https://doi.org/10.48550/arXiv.2408.06634

Nie YQ, Kong YX, Dong XW, et al., 2024. A survey of large language models for financial applications: progress, prospects and challenges. https://doi.org/10.48550/arXiv.2406.11903

Papineni K, Roukos S, Ward T, et al., 2002. BLEU: a method for automatic evaluation of machine translation. Proc 40th Annual Meeting of the Association for Computational Linguistics, p.311-318. https://doi.org/10.3115/1073083.1073135

Park T, 2024. Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework. https://doi.org/10.48550/arXiv.2403.19735

Wu SJ, Irsoy O, Lu S, et al., 2023. BloombergGPT: a large language model for finance. https://doi.org/10.48550/arXiv.2303.17564

Xie QQ, Han WG, Zhang X, et al., 2023. PIXIU: a large language model, instruction data and evaluation benchmark for finance. Proc 37th Int Conf on Neural Information Processing Systems, Article 1454. https://dl.acm.org/doi/10.5555/3666122.3667576

Yang HY, Liu XY, Wang CD, 2023. FinGPT: open-source financial large language models. https://doi.org/10.48550/arXiv.2306.06031

Yang HY, Zhang BY, Wang N, et al., 2024. FinRobot: an open-source AI agent platform for financial applications using large language models. https://doi.org/10.48550/arXiv.2405.14767

Yang Y, Uy MCS, Huang A, 2020. FinBERT: a pretrained language model for financial communications. https://doi.org/10.48550/arXiv.2006.08097

Yang Y, Tang YX, Tam KY, 2023. InvestLM: a large language model for investment using financial domain instruction tuning. https://doi.org/10.48550/arXiv.2309.13064

Yu YY, Li HH, Chen Z, et al., 2024. FinMem: a performance-enhanced LLM trading agent with layered memory and character design. *Proc AAAI Symp Series*, 3(1):595-597. https://doi.org/10.1609/aaaiss.v3i1.31290

Zhang BY, Yang HY, Liu XY, 2023. Instruct-FinGPT: financial sentiment analysis by instruction tuning of general-purpose large language models. https://doi.org/10.48550/arXiv.2306.12659

Zhang WT, Zhao LX, Xia HC, et al., 2024. A multimodal foundation agent for financial trading: tool-augmented, diversified, and generalist. Proc 30th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.4314-4325. https://doi.org/10.1145/3637528.3671801

Zhao HQ, Liu ZL, Wu ZH, et al., 2024. Revolutionizing finance with LLMs: an overview of applications and insights. https://doi.org/10.48550/arXiv.2401.11641

Zhu FB, Lei WQ, Huang YC, et al., 2021. TAT-QA: a question answering benchmark on a hybrid of tabular and textual content in finance. Proc 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing, p.3277-3287. https://doi.org/10.18653/v1/2021.acl-long.254

List of supplementary materials

1 Testing prompt

2 LLMs' response to users' queries

Table S1 Detailed components of AnalyScore

Table S2 Complete example of Stocksis

Table S3 Testing queries and expert's scores on FinSphere