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 LLMs and domain-specific tool integration

Shijie HAN[§1,3], Jingshu ZHANG[§2,3], Yiqing SHEN[3,4], Kaiyuan YAN[3], 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

unedited

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 LLMs' financial analysis capabilities. Building on Stocksis, together with a novel integration framework and quantitative tools, we develop FinSphere, an 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: Instruction-tuned financial LLM; Real-time stock analysis; Evaluation framework & dataset
https://doi.org/10.1631/FITEE.2500414 **CLC number:**

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 et al., 2023). By leveraging their language comprehension capabilities, these models have exhibited exceptional performance in various financial applications, including sentiment analysis (Zhang et al., 2023a; Liu et al., 2024b) and information extraction from unstructured financial texts (Huang et al., 2023; Li et al., 2024). The advent of finance-specific LLMs, such as FinBERT (Yang et al., 2020; Liu et al., 2021), 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 foundation for developing more sophisticated financial analysis tools and shifted how investors interact with market data (Krause, 2023; Nie et al., 2024). These AI-powered systems have broadened access to professional financial insights, allowing retail investors to benefit from advanced analysis once reserved for institutions.

‡ Corresponding author
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
© Zhejiang University Press 2025

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 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 et al., 2024). However, LLM-based financial analysis systems still face substantial challenges in effectively interpreting and utilizing 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 (Ni et al., 2024; Bhat and Jain, 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:

- **AnalyScore**: A comprehensive evaluation framework designed to systematically assess the accuracy, relevance, and analytical depth of LLM-driven stock analysis.
- **Stocksis**: A specialized dataset curated by industry experts to enhance LLMs' capabilities in stock diagnosis and financial analysis.
- **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, validating 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 (Zhao et al., 2024; Ni et al., 2024; Bhat and Jain, 2024; Kim et al., 2024). Domain-specific models, such as InvestLM (Yang 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 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 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 (Rouge, 2004)], while recent efforts such as FinEval (Zhang et al., 2023b) 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 et al., 2023), XBRL-Agent (Han et al., 2024), and FinRobot (Yang 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 dynam-

unedited

ically 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 analyses 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 points:

- **Conclusion** (20 points): Clarity, relevance, and personalization of investment recommendations.
- **Content** (45 points): Depth, coherence, and professionalism of analytical reasoning.
- **Expression** (15 points): Organization, fluency, and linguistic clarity.
- **Data** (20 points): Breadth, accuracy, and effective use of supporting quantitative data.

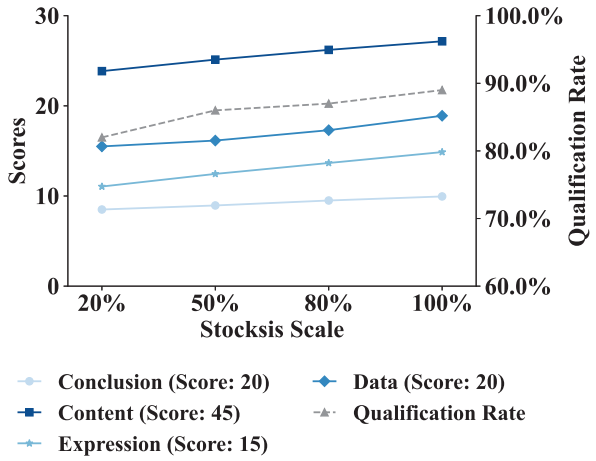

Fig. 1 Changes in scores of each sub-item as the number of Stocksis used for fine-tuning changes.

This multi-dimensional evaluation ensures comprehensive coverage of both qualitative judgment and quantitative rigor. AnalyScore is fully presented in Table 1, with the aim of promoting related research in academia and industry. Currently, AnalyScore is used exclusively by human experts to assess model-generated analyses. 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 face in synthesizing complex financial information into coherent, 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. Stocksis comprises 5,000 meticulously curated training pairs, with some of

unedited

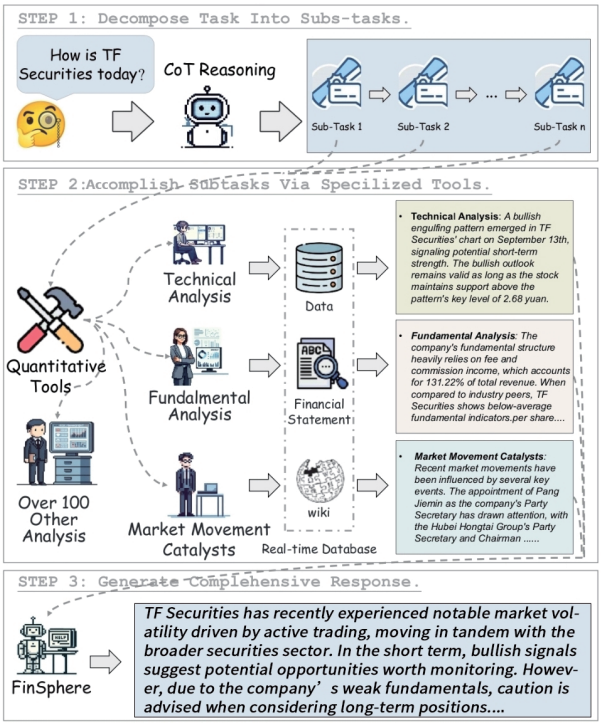

Fig. 2 This diagram illustrates the overall workflow of the FinSphere Agent, detailing how different components interact to facilitate real-time stock analysis.

them[1] available in the open-source release for research and development purposes. A complete example is detailed in Table 2. Each training sample consists of two key components:

- **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 4,000 words.

- **Expert-edited analysis (label)**: In-depth analytical reports respond to the prompt's requirements while effectively utilizing the background information, averaging 3,000 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.

[1]Open-sourced Stocksis is available at https://github.com/KirkHan0920/Stocksis

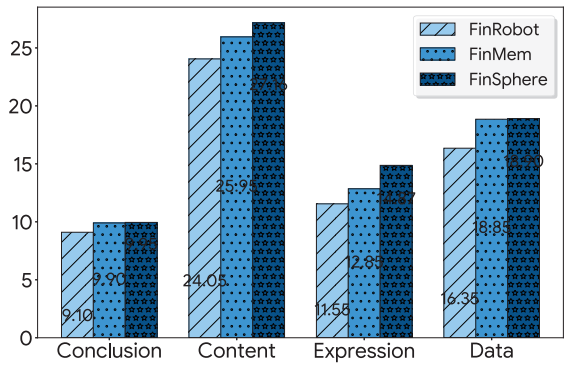

Fig. 3 Comparison of FinSphere and Agent-based systems in various dimensions.

3.2.1 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 outputs.

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 3 months to ensure high-quality outputs.

unedited

Table 1 Detailed components of AnalyScore

Evaluation dimensions	Subdimensions	Specific standards	Scores
Conclusion (Total Score: 20)		Generates personalized conclusions based on investment preferences and risk profiles Provides explicit conclusions tailored to user personas Implements differentiated investment strategies Aligns with risk tolerance and investment horizons	20
		Non-personalized but comprehensive analysis Covers diverse investment styles (conservative to aggressive) Enables user self-selection of strategies Shows broad analytical framework applicability	10
		Lacks personalization elements Single investment style analysis Limited analytical perspective Insufficient consideration of preferences	5
		No conclusive elements present	0
Content (Total Score: 45)	Analysis Dimensions	Leverages interaction history for personalization Includes $\leq$2 non-personalized dimensional analyses Shows accuracy and forward-looking insights	30
		Uses interaction history for personalization Provides accurate, targeted content Presents actionable recommendations	25
		Non-personalized analysis across $\leq$5 dimensions Demonstrates analytical accuracy and logical rigor	20
		Analysis across 4 core dimensions Includes real-time market analysis Maintains accuracy and timeliness	18
		Analysis across 4 core dimensions Maintains analytical accuracy	15
		Analysis across 3 core dimensions Demonstrates accuracy	10
		Analysis across 2 core dimensions Maintains basic accuracy	5
		Single-dimensional analysis Limited but accurate content	0
	Logical Consistency	Logical consistency across all components	15
		Exhibits logical inconsistencies	0
Expression (Total Score: 15)	Structure	Systematic organization	5
		Lacks structure	0
	Language	Clear communication with a hierarchical emphasis Logically rigorous with insightful perspectives In-depth analytical content	10
		Clear content with essential coverage Basic logical flow Superficial analysis	8
		Comprehensible but unclear expression Discontinuous logic Surface-level analysis	5
		Unclear and difficult to comprehend Logical inconsistencies Lacks substantive content	0
Data (Total Score: 20)		Extensive data (>3 dimensions)	20
		Substantial data (3 dimensions)	15
		Moderate data (2 dimensions)	10
		Limited data (1 dimension)	5
		No data utilization	0

unedited

Table 2 Complete example of Stocksis.

Prompt with background information (input)	Expert-edited analysis (label)
System Instruction. You are a proficient investment assistant, adept at stock analysis tasks. You demonstrate a robust aptitude for comprehending user-provided *[Question]* and meticulously organizing them into accurate and standardized analysis reports. You utilize the *[Background Information]* provided by the corresponding quantitative tools to respond to users in a systematic and standardized manner. **Background Information (get from quantitative tools).** • Volume and Price Analysis: TF Securities is currently trading at 4.48, showing a 0.90% increase. The stock ranks 896th overall and 8th within the securities industry. The securities sector index stands at 1580.86 with a 0.09% increase, ranging between a high of 1623.79 and a low of 1553.31, with a total sector trading value of 75.187 billion yuan. While the Shanghai Composite Index is at 3189.89, down 0.36%, TF Securities demonstrates strong trading activity with a 17.65% turnover rate (ranking 102nd overall and 1st in the securities industry). On October 16th, the stock recorded an impressive trading volume of 152.947 million shares with a total trading value of 6.970 billion yuan, leading both the overall market and the securities industry in trading volume. • Technical Analysis: A bullish engulfing pattern emerged in TF Securities' chart on September 13th, signaling potential short-term strength. The bullish outlook remains valid as long as the stock maintains support above the pattern's key level of 2.68 yuan. • Capital Flow Analysis: Recent capital flow indicators show mixed signals for TF Securities. While the 5-day cumulative DDX of -14.865 indicates significant selling pressure from major players, the stock maintains medium-level control by market forces with an optimistic outlook. The strong inflow in margin trading and increased institutional holdings from quarter to quarter suggest positive market sentiment and potential medium-term opportunities. • Financial Analysis: TF Securities released its 2024 semi-annual report on August 31st, which triggered a 23.75% stock price increase in the subsequent 10 trading days. The report revealed significant decreases in non-recurring net profit, revenue, and net profit attributable to the parent company. However, the Q2 2024 quarterly report showed substantial quarter-over-quarter improvements in these same metrics, indicating a potential turnaround in financial performance. • Fundamental Analysis: The company's fundamental structure heavily relies on fee and commission income, which accounts for 131.22% of total revenue. When compared to industry peers, TF Securities shows below-average fundamental indicators. As a member of the securities industry according to SWSI classification, the company's growth metrics raise concerns, with negative year-over-year growth rates in revenue, net profit, and earnings per share. **Question.** Please analyze TF Securities.	**TF Securities has shown unusual market movement due to active market trading, benefiting along with the broader securities sector. In the short term, there are bullish signals warranting attention. However, caution is advised for medium to long-term positions due to weak fundamentals.** Regarding volume and price, TF Securities' latest stock price is 4.48 yuan, up 0.90%. Today's turnover rate is 17.65%, ranking first in the securities industry, with a trading volume of 152.947 million shares and a trading value of 6.970 billion yuan. The stock ranks 8th in its industry for price gains and is outperforming the Shanghai Composite Index, which is down 0.36%. The volume and price performance show active market participation and high overall engagement. On the technical front, TF Securities formed a bullish engulfing pattern on September 13th, indicating short-term strength. The bullish outlook remains valid as long as the price remains above 2.68 yuan. Additionally, institutional analysis tools indicate strong institutional control, suggesting potential opportunities for investors. From a capital flow perspective, TF Securities' 5-day cumulative DDX of -14.865 indicates significant selling pressure from major players and strong bearish sentiment. Despite medium-level institutional control and determined margin trading inflows suggesting relatively active capital participation, caution is needed regarding short-term capital flows. Fundamentally, TF Securities' 2024 semi-annual report shows significant decreases in non-recurring net profit, revenue, and net profit attributable to the parent company, indicating poor overall growth. Moreover, the company's heavy reliance on fee and commission income as its main revenue source places it at a disadvantage compared to industry peers. **In conclusion, TF Securities presents short-term opportunities due to the bullish engulfing pattern in technical analysis. However, investors should maintain caution in medium to long-term positions due to weak fundamentals. Investment decisions should be made with close attention to market trends and company developments.**

The average cost of a comprehensive analysis written by experts is $10 per query.

unedited

4 FinSphere Agent

This section details the architecture and operational mechanisms of FinSphere Agent, our advanced stock analysis agent.

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.

4.1 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 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. More than 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. This category includes index quotation tools, Hong Kong stock quotation tools, fund quotation tools, A-share quotation tools, U.S. stock quotation tools, and NEEQ quotation tools, as well as dedicated analysis toolkits for *Technical Analysis*, *Fundamental Analysis*, and *Market Movement Catalysts* (together enabling FinSphere to support not only Chinese A-shares but also Hong Kong and 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, which enhance interpretability and add qualitative insights to the quantitative outputs.

This modular design ensures that analytical responses remain accurate, interpretable, and reflective of 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.

unedited

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 5,000 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 32K 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 struc-

tured, 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 responses are both 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, GPT-3.5), open-source [Qwen2-72B, Deepseek-v3 (Liu et al., 2024a)], and domain-specific models [InvestLM (Yang et al., 2023) and FinGPT (Yang et al., 2023)]; (2) agent-based systems, including FinMem (Yu et al., 2024) and FinRobot (Yang et al., 2024); and (3) our proposed FinSphere. All LLMs use CoT prompting with few-shot examples and relevant background information (Sect. "Testing promp"), while agents receive simplified prompts similar to Stocksis's inputs. FinSphere is evaluated directly via user queries, utilizing its integrated real-time data and tools. For all models, we set a maximum output length of 8K tokens and a temperature of 0.5.

5.2 Performance analysis

The expert evaluation results presented in Table ?? 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 (40.05) demonstrate relatively limited capabilities. 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 Figure 3 highlights performance differences across four critical dimensions: Conclusion, Content, Expression, and Data capabilities.

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 27.16, significantly surpassing FinMem (25.95) and Fin-

unedited

Robot (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. Forty industry experts are divided 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, reported in Table 4, 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* exhibits the highest variation in agreement across groups, with values spanning from 71.45 to 94.25, while *Expression* and *Data* 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 were able to systematically distinguish high-quality responses, ensuring that the evaluation framework accurately reflects model performance.

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 5,000 data pairs while maintaining FinSphere's framework. The detailed evaluation results are provided in Fig. 1, 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.

5.5 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**

unedited

Table 3 Ablation results for major FinSphere modules.

	Conclusion (Score: 20)	Content (Score: 45)	Expression (Score: 15)	Data (Score: 20)	Total score
FinSphere (full)	**9.95**	**27.16**	**14.87**	**18.90**	**70.88**
w/o Decision CoT	6.85	19.20	9.78	12.17	47.99
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.

Table 4 Average Kendall's Tau across 100 queries for different annotator groups rating.

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

Reported as %.

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; and (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 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 led to notable declines in *Content* and *Expression*, highlighting the gains from domain-specific instruction tuning.

5.6 Ablation on training data scale

6 Conclusions

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 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 systems across multiple evaluation dimensions, highlighting the effectiveness of our integrated approach.

Limitation

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.

Product release information

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-

unedited

jiuzhang.techgp.cn/rjhy/jzmodelsInner/#/public-login.

Testing prompt

As a professional investment advisor, you will analyze stocks based on the following:

[BACKGROUND INFORMATION - INSERT QUANTITATIVE TOOL RETURNS HERE]

[RESPONSE STANDARDS - SEE BELOW]

[WRITING GUIDELINES - SEE BELOW]

[FEW-SHOTS]

[SPECIFIC QUERY]

Response standards

Step 1: Movement Summary

- Clearly state whether the stock shows unusual movement
- Summarize the main reason from the background information in less than 20 words

Step 2: Conclusions

1. Short-term conclusion (based on technical analysis):
    - For bearish signals: suggest caution, observation, risk awareness, avoidance, and position control
    - For bullish signals: suggest appropriate attention, tracking, and validation
    - Note: state conclusion only, no explanation needed
2. Medium/long-term conclusion (based on fundamental analysis):
    - For bearish signals: suggest caution, observation, risk awareness, avoidance, and position control
    - For bullish signals: suggest appropriate attention, tracking, and validation
    - Note: state conclusion only and no explanation needed
3. Note: Technical analysis showing bullish keywords requires only direct bullish conclusions. Bullish keywords and bearish keywords are shown in the writing guidelines.

Step 3: Detailed Analysis

Provide an overview of the specific content in [background information] (such as volume and price situation, technical aspect, capital aspect, fundamental aspect, and news aspect) as a support for generating conclusions, ensuring that it does not change the financial data performance in the background information. The analysis mode that can be used for reference is as follows:

- Volume and price analysis: Current price, price movement, industry comparison, index comparison, turnover rate, trading volume/value, and market comparison. Include specific data and brief commentary.
- Technical analysis: Technical patterns, indicators with specific values. Note specialized indicators if present (e.g., "AI Top/Bottom" and "Bull Institution signals").
- Capital flow analysis: 5-day cumulative DDX and capital flow data analysis. Include specific DDX values.
- Fundamental analysis: Financial and fundamental data evaluation. Include specific values.
- News analysis: Latest news and movement causes, including specific numerical data.

Step 4: Final Summary

Restate short-term and medium/long-term conclusions with reasons.

Writing guidelines

- Provide direct conclusions and analysis, and maintain a concise response
- Avoid phrases: "according to," "information shows," "recent performance," "current situation," "comprehensive analysis," "buy," "sell," "hold," "clear position," "build position," "reduce position," and "increase position"

unedited

Table 5 Testing queries and experts' scores on FinSphere (2/3)

Query	Qual	Conc	Cont	Expr	Data	Score
Can you conduct a comprehensive analysis of Xingqiu Graphite?	1	10	35	15	20	80
How is the comprehensive analysis of Shanghai Ailu?	1	10	30	15	20	75
I want to know the details of Taihe Technology. Can you analyze it for me?	1	10	30	15	20	75
Comprehensive analysis of Dagang Holdings.	1	10	25	15	20	70
Please analyze Aorui De and provide investment advice.	1	10	25	15	20	70
Can you analyze Sunshine Real Estate?	1	10	25	15	20	70
Can you conduct a comprehensive analysis of Hongbaoli?	1	10	25	15	20	70
Can you analyze Yangzi New Materials for me?	1	10	25	15	20	70
How has Chunxing Precision performed recently? Can you analyze it?	1	10	35	15	20	80
Can you provide professional analysis of Tuoshan Heavy Industry?	1	10	25	15	20	70
I am interested in the analysis of Yayi Technology. Can you share it?	1	10	20	15	20	65
What are the key points to watch in Bofei Electric? Can you analyze it?	1	10	25	15	20	70
Can you conduct a detailed analysis of Kangliyuan?	1	10	25	15	20	70
How comprehensive is the strength of Hope Co., Ltd.?	1	10	25	15	20	70
Please conduct a comprehensive analysis of Kuntai Co., Ltd.	1	10	25	15	20	70
Can you provide comprehensive feedback on Taimusi?	1	10	25	15	20	70
What do you think about Hongming Co., Ltd.?	1	10	25	15	20	70
Can you look at Wuzhou Medical for me?	1	10	25	15	20	70
How is Zhejiang Liming recently?	1	10	25	15	20	70
Please analyze Baolijia, is it good?	1	10	25	13	15	63
Can you analyze Lvlian Technology?	1	10	25	13	15	63
Is Shanghai Hejing worth buying?	1	10	25	13	20	68
What do you think of Qiaoyuan Co., Ltd.?	1	10	25	15	20	70
How about Zhongji Huanke?	1	10	25	15	20	70
Can you talk about Kangguan Technology stock?	1	10	25	15	20	70
Tell me about Guanghe Technology.	1	10	25	15	20	70
Is Xingchen Technology doing well recently?	1	10	25	15	20	70
Please interpret the Chengdu Huawei stock.	1	10	25	15	20	70
How about Jinjiang Shipping?	1	10	30	15	20	75
What about Jinhui Co., Ltd.?	1	10	25	15	20	70
Can you provide an investment analysis of Dazhu CNC?	1	10	25	15	20	70
Is Laplace stock good?	1	5	20	15	15	55
What do you think of Shennong Group stock?	1	10	30	15	20	75

unedited

Table 6 Testing queries and experts' scores on FinSphere (1/3)

Query	Qual	Conc	Cont	Expr	Data	Score
Can you analyze Fueneng Dongfang?	1	10	35	15	15	75
Please analyze the situation of Oulai New Materials stock.	1	10	35	15	20	80
Conduct a comprehensive analysis of Kangxi Communications.	1	10	35	15	15	75
Comprehensive analysis of Chuangyitong.	0	10	35	15	20	80
Can you provide a detailed interpretation of Lanhai Huanteng stock?	1	10	20	15	20	65
Please analyze the current status of the Guixin Technology stock.	1	10	33	15	15	73
Is Zhejiang Hengwei worth investing in? Please analyze.	1	10	35	15	20	80
Conduct a comprehensive analysis of Anshuo Information.	1	10	35	13	15	73
I am interested in Chenyi Intelligence. Could you analyze it?	0	10	10	15	15	50
Diagnose Hailun Zhe.	1	10	10	15	20	55
Can you conduct an in-depth analysis of Fuguang Co., Ltd.?	1	10	35	15	15	75
Please analyze Haooubo.	1	10	5	15	20	50
How is Canxin Co., Ltd.?	1	10	30	15	20	75
Comprehensive analysis of Saiwei Intelligence.	0	10	30	15	20	75
Please provide a comprehensive evaluation of Cigu Technology.	0	10	30	15	15	70
Please comment on the overall performance of Sainuo Medical.	0	10	10	15	15	50
How is Longli Technology?	1	10	30	15	15	70
Can you provide a comprehensive analysis of Aofu Environmental Protection?	1	10	30	15	15	70
What is the comprehensive situation of Yubang New Materials?	1	10	30	15	20	75
Please conduct a comprehensive review of Zhuojin Co., Ltd.	1	10	10	15	15	50
How is the performance of Huaguang New Materials in all aspects?	1	10	5	15	20	50
Please conduct a comprehensive analysis of Jiankang.	1	10	25	15	15	65
Overall analysis of Taifu Pumps.	1	10	30	15	15	70
Comment on Zhongfu Information.	1	10	25	15	20	70
Please provide an analysis of Daoshi Technology.	1	10	30	15	20	75
Comprehensive analysis of Ruisong Technology.	1	10	30	15	20	75
Comprehensive analysis of Zhongyi Technology.	0	10	30	15	20	75
Can you provide a comprehensive evaluation and analysis of Aerospace Hongtu?	1	10	30	15	10	65
Please give specific analysis opinions on Tengjing Technology.	1	10	30	10	20	70
Comprehensive analysis of Zhenyou Technology.	1	10	25	15	10	60
How is the overall situation of Huahai Chengke?	0	10	35	15	20	80

unedited

Table 7 Testing queries and experts' scores on FinSphere (3/3)

Query	Qual	Conc	Cont	Expr	Data	Score
Comprehensive analysis of Shichuang Energy.	1	10	25	15	20	70
Comprehensive analysis of Ningbo Ocean.	1	10	25	15	20	70
Comprehensive analysis of Longqi Technology.	1	10	25	15	20	70
Please evaluate Fuerjia Co., Ltd. as a whole.	1	10	25	15	20	70
Can you provide an overall evaluation of Yongxing Co., Ltd.?	1	10	25	15	20	70
Comprehensive analysis of Hekeda Co., Ltd.	1	10	25	15	20	70
Comprehensive analysis of Craftsman Home.	1	10	25	15	20	70
Comprehensive analysis of International Composite Materials.	1	10	25	15	20	70
Please diagnose and analyze Suzhou Tianmai.	0	10	20	15	15	60
Can you diagnose the stock status of Weidian Physiology?	0	10	25	15	20	70
Please diagnose Weidao Nano.	1	10	25	15	20	70
Can you diagnose stock 6,912?	0	10	20	15	15	60
How is Hualan Vaccine stock?	1	10	25	15	20	70
Please diagnose Mousse Co., Ltd.'s stock comprehensively.	1	10	25	15	20	70
Can you conduct an in-depth analysis of Huabao New Energy stock?	1	10	25	15	20	70
Please diagnose Haikan Co., Ltd.'s stock.	1	10	25	15	20	70
Can you provide a professional analysis of Hongsheng Huayuan?	1	10	25	15	20	70
Please diagnose China Eastern Airlines.	1	10	30	15	20	75
Can you analyze the stock of Huali Group in detail?	1	10	25	15	20	70
Please analyze the Postal Savings Bank.	1	10	25	15	20	70
How to analyze the market trend of ICBC?	1	10	30	15	20	75
How about Guizhou Moutai stock?	1	10	30	15	20	75
Please analyze the Agricultural Bank of China stock.	1	10	25	15	20	70
Can you analyze China Construction Bank in detail?	1	10	30	15	20	75
China Petroleum, diagnose it.	1	10	30	15	20	75
Can you conduct a comprehensive analysis of China Mobile?	0	10	25	15	20	70
Is China Bank suitable for long-term holding?	1	10	30	15	20	75
Please analyze China Life Insurance stock.	1	10	30	15	20	75
Please research Ningde Times stock.	1	10	30	15	20	75
Please analyze Zhaosheng Micro stock.	1	10	25	15	20	70
Please analyze Xinda Securities stock.	1	10	25	15	20	70
Analyze BAIC Blue Valley.	1	10	33	15	20	78
How to view COSCO Energy stock?	1	10	25	15	20	70
Analyze Kelun Pharmaceutical stock.	1	10	25	15	20	70
Can you conduct a comprehensive diagnosis of New Industries' stock?	1	10	30	15	20	75
Comprehensive analysis of Shengyi Technology.	1	10	25	15	20	70

Unedited

- Minimize transition words: "but," "however," "then," and "finally"
- Maintain natural flow between paragraphs
- Bullish keywords: bull point, golden cross, strong uptrend, strengthening, active holding, bullish combination, and active attention
- Bearish keywords: bear point, death cross, weakening, cautious observation, bearish combination, weak adjustment, and weak decline.

LLMs' response to users' query

Here we show the responses of four LLMs (using different methods) to real user inquiries, along with the scores and comments of experts after evaluation using AnalyScore.

FinSphere

Omat Advanced Materials shows unusual movement, mainly due to the company's plan to build Mingyue Lake Industrial Park in the Shaoguan High-tech Zone. In the short term, Omat's stock price shows strong momentum, warranting active attention. In the medium to long term, given the company's weak fundamentals, caution is advised.

Regarding price and volume, Omat's latest stock price is 25.28 yuan, representing a 14.60% increase. It ranks 9th among semiconductor stocks, with the sector index up 5.91% and the Shanghai Composite Index up 0.09%, significantly outperforming both the sector and the broader market. Omat's turnover rate is 54.01%, ranking 1st in the semiconductor industry, with a total trading volume of 172,900 lots and a total trading value of 407 million yuan, ranking 63rd in the industry. While Omat shows strong price performance today with high market participation, the trading volume is relatively average.

Technical analysis shows that Omat's RSI formed a golden cross above 50 on November 5th, followed by a MACD golden cross above the zero axis on November 8th. The short-term RSI has reached 80.54, indicating overbought conditions, suggesting a possible strong adjustment in the near term.

Regarding capital flows, Omat's 5-day cumulative DDX is 6.397, indicating strong recent institutional buying. The stock shows moderate institutional control, continued margin trading inflows, and increased quarterly institutional holdings, suggesting positive capital sentiment.

Financially, Omat's Q3 2024 report shows significant declines in revenue and net profit attributable to shareholders. The company's solvency is weak, with declining current and quick ratios and a steadily rising debt-to-asset ratio, indicating poor fundamentals compared to its industry peers.

On the news front, Omat acknowledged at its performance briefing that due to economic conditions, the company's performance is under pressure, with revenue down 22.63% and net profit down 77.38% in the first three quarters. The company's plan to build the Mingyue Lake Industrial Park in the Shaoguan High-tech Zone, which aims to expand its upstream high-purity materials and copper alloy new materials business, demonstrates confidence in future development.

In conclusion, Omat shows positive short-term technical indicators, as indicated by RSI and MACD golden crosses, warranting attention. For the long term, given weak fundamentals, poor performance, and weak solvency, caution is advised. The Bull Force Index signals an entry opportunity, suggesting investors watch for potential trading opportunities.

unedited

Score and comments from experts using AnalyScore FinSphere's analysis clearly identifies the causes of unusual price movements and investment recommendations with complete data citations, including sector comparisons. The conclusions align with the data: based on strong technical indicators but weak fundamentals, it provides a short-term focus and mid-term cautious recommendations. The analysis across different dimensions is logical with sufficient data support. Specific scores are as follows:

- Conclusion (10/20):
  - Clear distinction between short-term and mid-to-long-term recommendations
  - Data-driven investment advice
  - Lacks personalized recommendations
- Content (35/45):
  - Covers five core dimensions (volume/price, technical, capital flow, financial, and news)

  - Professional and in-depth analysis across all dimensions
  - Strong logical consistency
  - Lacks personalization
- Expression (15/15):
  - Clear structure with hierarchical organization
  - Accurate use of professional terminology
  - Deep analysis with insights
- Data (20/20):
  - Comprehensive data dimensions (>3 dimensions)
  - Accurate data citations with comparative analysis
- Total: 80

GPT-4o with few-shot

Omat Advanced Materials shows unusual movement due to its planned Mingyue Lake Industrial Park project in the Shaoguan High-tech Zone, which has boosted market expectations. The stock deserves moderate attention in both the short and medium terms.

Regarding price and volume, Omat's latest price is 25.28 yuan, up 14.60%, ranking 9th in the semiconductor industry. The semiconductor sector index stands at 2,740.03, up 5.91%. The Shanghai Composite Index is at 3,455.28, up 0.09%. Omat's turnover rate of 54.01% ranks first in the industry, with a total trading volume of 172,900 lots and a trading value of 407 million yuan, ranking 63rd. The stock shows strong price gains and high market activity.

Technically, Omat formed an RSI golden cross above 50 on November 5th, followed by a MACD golden cross above the zero axis on November 8th, maintaining strong momentum. The current RSI of 80.54 indicates overbought conditions, suggesting a potential near-term adjustment.

Capital flow analysis shows a 5-day cumulative DDX of 6.397, indicating strong institutional buying and moderate institutional control, with decisive margin trading inflows.

Fundamentally, Omat's Q3 2024 report shows revenue of 297 million yuan for the first three quarters, down 22.63% year-over-year, and net profit of 9.8746 million yuan, down 77.38%. The company's solvency is weak, with a current ratio down 11.42%, a quick ratio down 12.27%, and a steadily rising debt-to-asset ratio.

Recent news indicates the company's plan to build Mingyue Lake Industrial Park in the Shaoguan High-tech Zone, expanding upstream operations, and demonstrating confidence in future development. The market views the company's multi-channel funding approach positively for steady project implementation.

In conclusion, Omat warrants moderate short-term attention given technical golden crosses and strong momentum. Medium-term moderate attention is advised, as despite the expansion plans, fundamental data remains weak, requiring careful monitoring.

Score and comments from experts using AnalyScore GPT-4o with few-shot's analysis begins with vague conclusions: "worth moderate attention in short term, should pay appropriate attention in medium to long term" lacks a clear stance; while data citations are complete, the analysis lacks depth; technical analysis is oversimplified and fails to fully interpret indicator meanings; the conclusion section lacks sufficient reasoning for investment recommendations. Specific scores are as follows:

- Conclusion (10/20):
  - Ambiguous conclusions
  - Lack of specific investment recommendations
  - Insufficient reasoning
- Content (25/45):
  - Relatively complete dimensional coverage
  - Insufficient depth of analysis
  - Simplified technical interpretation
- Expression (15/15):
  - Basic clear structure
  - Plain expression
  - Lacks insights

unedited

- Data (15/20):
  - Complete data citations
  - Insufficient data interpretation
- Total: 65

QWen2-72B with few-shot

Fundamentally, Omat Advanced Materials reported revenue of 297 million yuan for the first three quarters of 2024, down 22.63% year-over-year, with a net profit of 9.8746 million yuan, a 77.38% decrease. Profitability metrics are weak, with ROE at 1.5%, net profit margin at 3.33%, and gross margin at 18.87%. Valuation metrics show P/E at 264.02 and P/B at 4.12, ranking at 46.5% and 64.56% in the industry, respectively.

Technical analysis shows significant price volatility, with a 10-day amplitude and range of 48.21% and 19.54%. The 5-day, 10-day, and 20-day moving averages are 21.72, 21.87, and 21.87 yuan, respectively. While the price is above these averages, their bearish alignment suggests potential near-term pressure.

Capital flow analysis shows volatile institutional flows over the past 10 days, with overall net outflows. The margin trading balance stands at 7.0404 million yuan, representing 0.44% of float, indicating low leverage participation.

As a STAR Market-listed company, Omat has raised funds for projects including high-end sputtering target production facilities. The company's products serve various sectors, including semiconductor displays, touch screens, architectural glass, decorative coating, IC packaging, solar photovoltaic, and power batteries.

In conclusion, Omat requires caution in the short term given the bearish technical alignment and weak capital flows. A medium- to long-term focus should be on performance improvement and industry development opportunities. Investors should closely monitor company performance, industry trends, and market capital flows.

Score and comments from experts using AnalyScore QWen2-72B with few-shot's analysis lacks explanation of the causes behind unusual movements; technical analysis is disconnected from current trends and fails to reflect recent changes; capital flow descriptions are too general without specific data support; news analysis is superficial and fails to demonstrate a deep analysis of company strategy; and conclusions are overly conservative and do not match some positive signals. Specific scores are as follows:

- Conclusion (5/20):
  - Overly conservative conclusions
  - Mismatched with data support
  - Vague recommendations
- Content (25/45):
  - Lacks analysis of abnormal movement causes
  - Disconnected technical analysis
  - Superficial multi-dimensional analysis
- Expression (10/15):
  - Complete but weak hierarchical structure
  - Lacks professional expression
- Data (15/20):
  - Basic complete data dimensions
  - Insufficient data support for some dimensions
- Total: 55

unedited

Contributors

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

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.

References

Bhat R, Jain B, 2024. Stock price trend prediction using emotion analysis of financial headlines with distilled llm model. Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments, p.67-73.

Chen J, Zhou P, Hua Y, et al., 2024. Fintextqa: A dataset for long-form financial question answering. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), p.6025-6047.

Chen Z, Chen W, Smiley C, et al., 2021. Finqa: A dataset of numerical reasoning over financial data. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, p.3697-3711.

Ding H, Li Y, Wang J, et al., 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:240806361*, .

Gupta U, 2023. Gpt-investar: Enhancing stock investment strategies through annual report analysis with large language models. *arXiv preprint arXiv:230903079*, .

Han S, Kang H, Jin B, et al., 2024. Xbrl agent: Leveraging large language models for financial report analysis. Proceedings of the 5th ACM International Conference on AI in Finance, p.856-864.

Huang AH, Wang H, Yang Y, 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806-841.

Islam P, Kannappan A, Kiela D, et al., 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:231111944*, .

Kim A, Muhn M, Nikolaev V, 2024. Financial statement analysis with large language models. *arXiv preprint arXiv:240717866*, .

Krause D, 2023. Large language models and generative ai in finance: An analysis of chatgpt, bard, and bing ai. SSRN Working Paper, 4526071.

Lei Y, Li J, Jiang M, et al., 2023. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv preprint arXiv:231105812*, .

Li H, Gao H, Wu C, et al., 2024. Extracting financial data from unstructured sources: Leveraging large language models. *Journal of Information Systems*, :1-22.

Li Y, Wang S, Ding H, et al., 2023. Large language models in finance: A survey. Proceedings of the fourth ACM international conference on AI in finance, p.374-382.

Liu A, Feng B, Xue B, et al., 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:241219437*, .

Liu C, Arulappan A, Naha R, et al., 2024b. Large language models and sentiment analysis in financial markets: A review, datasets and case study. *IEEE Access*, .

Liu Z, Huang D, Huang K, et al., 2021. Finbert: A pre-trained financial language representation model for financial text mining. Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence, p.4513-4519.

Ni H, Meng S, Chen X, et al., 2024. Harnessing earnings reports for stock predictions: A qlora-enhanced llm approach. *arXiv preprint arXiv:240806634*, .

Nie Y, Kong Y, Dong X, et al., 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:240611903*, .

Papineni K, Roukos S, Ward T, et al., 2002. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, p.311-318.

Park T, 2024. Enhancing anomaly detection in financial markets with an llm-based multi-agent framework. *arXiv preprint arXiv:240319735*, .

Rouge LC, 2004. A package for automatic evaluation of summaries. Proceedings of Workshop on Text Summarization of ACL, Spain, 5.

Wu S, Irsoy O, Lu S, et al., 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:230317564*, .

Xie Q, Han W, Zhang X, et al., 2023. Pixiu: a large language model, instruction data and evaluation benchmark for finance. Proceedings of the 37th International Conference on Neural Information Processing Systems, p.33469-33484.

Yang H, Liu XY, Dan Wang C, 2023. Fingpt: Open-source financial large language models. *FinLLM at IJCAI*, .

Yang H, Zhang B, Wang N, et al., 2024. Finrobot: An open-source ai agent platform for financial applications using large language models. *arXiv preprint arXiv:240514767*, .

Yang Y, Uy MCS, Huang A, 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:200608097*, .

Yang Y, Tang Y, Tam KY, 2023. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:230913064*, .

Yu Y, Li H, Chen Z, et al., 2024. Finmem: A performance-enhanced llm trading agent with layered memory and character design. Proceedings of the AAAI Symposium Series, 3:595-597.

Zhang B, Yang H, Liu XY, 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:230612659*, .

Zhang L, Cai W, Liu Z, et al., 2023b. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:230809975*, .

Zhang W, Zhao L, Xia H, et al., 2024. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, p.4314-4325.

Zhao H, Liu Z, Wu Z, et al., 2024. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:240111641*, .

Zhu F, Lei W, Huang Y, et al., 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

unedited