Shijie HAN, Jingshu ZHANG, Yiqing SHEN, Kaiyuan YAN, Hongguang LI, 2025. FinSphere: a real-time stock analysis agent with instruction-tuned large language models and domain-specific tool integration. *Frontiers of Information Technology & Electronic Engineering,* 26(10):1822-1831. https://doi.org/10.1631/FITEE.2500414

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

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

Corresponding author: Hongguang LI
E-mail: harvey2@mail.ustc.edu.cn
ORCID: https://orcid.org/0009-0003-0625-8213

Motivation

1. Financial large language models (LLMs) lack a standardized, expert-grounded metric to assess the quality of stock analysis reports. This gap leads to uneven analytical depth and inconsistent reporting quality.

2. Real-time professional analysis requires continuously updated data sources and domain tools rather than static prompts or batch data. Existing agent systems often depend on pre-defined tools and stale data, which constrains responsiveness.

3. The field does not provide an expert-annotated dataset that reflects full, professional stock-diagnostic reports suitable for training and evaluation.

Main idea

1. Propose AnalyScore, an expert-grounded evaluation framework that assesses stock-analysis quality across four dimensions: conclusion (20), content (45), expression (15), and data (20).

2. Construct Stocksis, an expert-curated dataset of 5000 pairs aligning tool-enriched prompts with expert-edited analyses to provide professional-grade supervision.

3. Develop FinSphere, a real-time stock-analysis agent that integrates continuously updated financial databases, over one hundred quantitative tools and an instruction-tuned large language model.

Method

1. AnalyScore

- Purpose: establish a systematic and expert-grounded method to evaluate LLM-generated stock analyses.
- Scoring dimensions and weights: conclusion (20), content (45), expression (15), data (20). These dimensions emphasize actionability, depth and coherence, writing quality, and the breadth and accuracy of data usage.
- Implementation: human experts currently apply AnalyScore; the study further reports group-level evaluation and agreement statistics.

Method (Cont'd)

2. Stocksis

- Dataset design: each pair combines a prompt enriched with outputs from multiple tools and a corresponding expert-edited analysis, enabling training toward professional-grade reasoning.
- Scale and availability: the dataset comprises 5000 meticulously curated pairs; a subset is open-sourced for research, and full access is available upon request.
- Role in the pipeline: Stocksis provides high-quality supervision for instruction tuning and evaluation.

Method (Cont'd)

3. FinSphere

- Architecture and training: FinSphere couples continuously updated financial databases with over one hundred specialized quantitative tools, and applies full-parameter instruction tuning on the Stocksis dataset to enhance analytical depth and coherence.
- Inference: the model decomposes user queries, selects tools through decision-oriented reasoning, and synthesizes the outputs into a coherent, professional stock-analysis report.

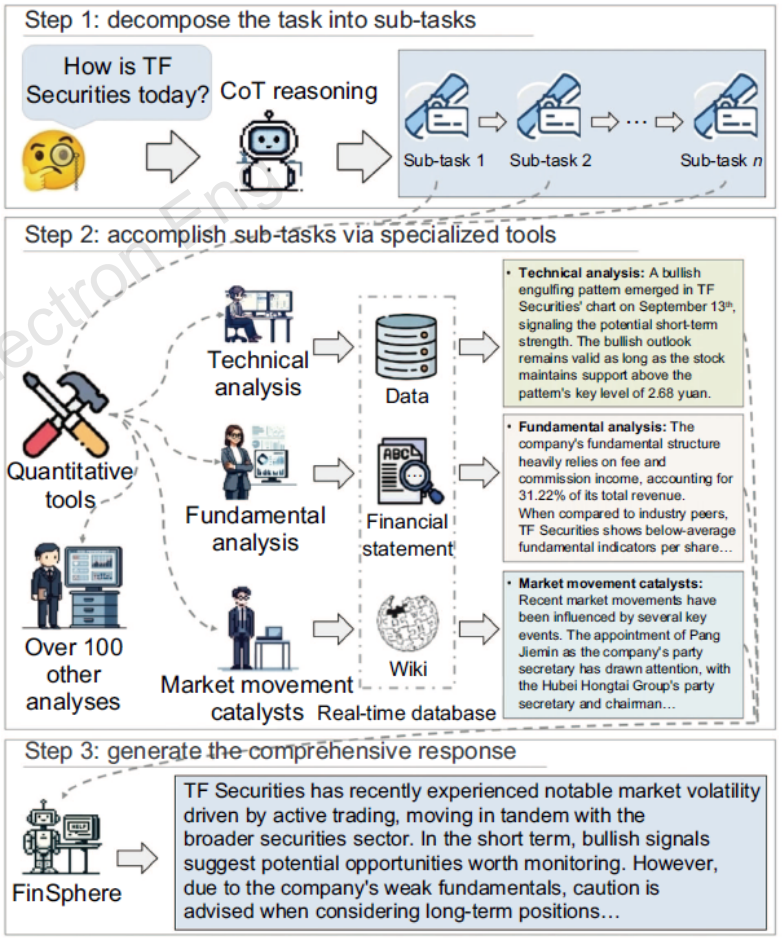

Fig. 1 Diagram of the overall workflow of the FinSphere agent. This details how different components interact to facilitate real-time stock analysis. TF: Tianfeng

Major results

Overall performance

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

FinSphere achieves an overall AnalyScore of 70.88 with sub-scores of 9.95 (conclusion), 27.16 (content), 14.87 (expression), and 18.90 (data), outperforming general-purpose LLMs, domain financial LLMs (FinLLMs), and agent baselines.

Major results (Cont'd)

Training data scale

More data from Stocksis lead to better performance, and FinSphere maintains satisfactory performance levels even with reduced training data.

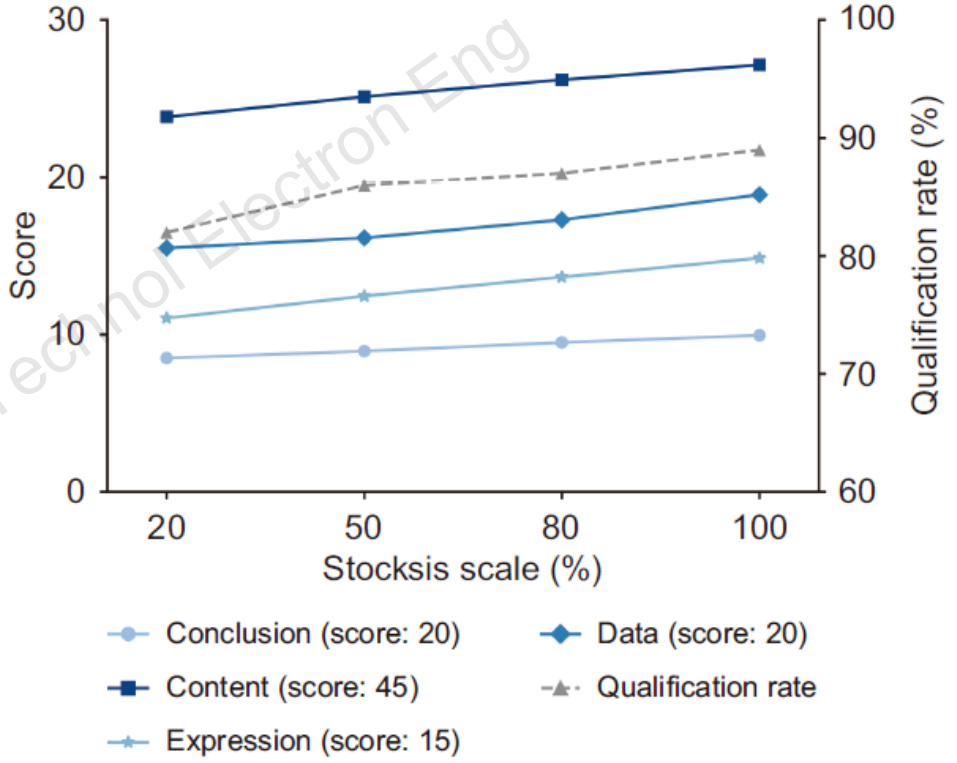

Fig. 3 Changes in scores of each sub-item as the Stocksis scale used for fine-tuning changes

Major results (Cont'd)

System modules

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

w/o decision CoT: performance drops significantly (Tool selection is crucial).

w/o tool module: performance collapses, especially in the "Data" score (Real-time data are non-negotiable).

w/o instruction tuning: performance degrades (Domain-specific fine-tuning is key).

Conclusions

1. Three-part contributions: AnalyScore (evaluation), Stocksis (expert data), and FinSphere (real-time agent) jointly address core gaps in FinLLMs.

2. Outcome: FinSphere achieves state-of-the-art expert scores with deeper content, clearer expression, and strong data usage.

3. Practicality: Stocksis comprises 5000 meticulously curated training pairs, with some of them available in the open-source release at https://github.com/KirkHan0920/Stocksis for R&D purposes. A fully functional product demo of FinSphere has been released and made freely available to the public since Dec. 2024 (see https://uatjiuzhang.techgp.cn/rjhy/jzmodelsInner/#/public-login).

Shijie HAN is a PhD student in the Department of Industrial Engineering and Decision Analytics at the Hong Kong University of Science and Technology. He holds an MS degree in Operations Research from Columbia University. His research interests include artificial intelligence (AI) in finance and business, supply chain resilience, reasoning enhancement and hallucination mitigation in LLMs, and data-driven analysis.

Jingshu ZHANG holds a PhD degree in Management from Shanghai University of Finance and Economics. Her research interests include AI, quantitative finance, and big data analytics.

Hongguang LI is head of LLM post-training in Qifu Technology, a senior expert, and a member of the LLM Committee of the Chinese Information Processing Society. He is responsible for building financial foundation models and financial AI agents that power customer-service. He was a former head of the LLM team in JF Smart Invest and a former head of Conversational AI at Xiaoice, where he led the memory-enhanced agent. He has published over 10 papers in top NLP conferences and holds six granted Chinese invention patents.