

# FinGAIA: An End-to-End Benchmark for Evaluating AI Agents in Finance

Lingfeng Zeng<sup>1</sup>, Fangqi Lou<sup>1</sup>, Zixuan Wang<sup>1</sup>, Jiajie Xu<sup>1</sup>, Jinyi Niu<sup>2</sup>, Mengping Li<sup>1</sup>, Yifan Dong<sup>1</sup>, Qi Qi<sup>1</sup>, Wei Zhang<sup>1</sup>, Ziwei Yang<sup>1</sup>, Jun Han<sup>1</sup>, Ruilun Feng<sup>1</sup>, Ruiqi Hu<sup>1</sup>, Lejie Zhang<sup>1</sup>, Zhengbo Feng<sup>1</sup>, Yicheng Ren<sup>1</sup>, Xin Guo<sup>1</sup>, Zhaowei Liu<sup>1</sup>, Dongpo Cheng<sup>1</sup>, Weige Cai<sup>1</sup>, Liwen Zhang<sup>1\*</sup>

<sup>1</sup>Shanghai University of Finance and Economics

<sup>2</sup>Fudan University

zhang.liwen@shufe.edu.cn

## Abstract

The booming development of AI agents presents unprecedented opportunities for automating complex tasks across various domains. However, their multi-step, multi-tool collaboration capabilities in the financial sector remain under-explored. This paper introduces FinGAIA, an end-to-end benchmark designed to evaluate the practical abilities of AI agents in the financial domain. FinGAIA comprises 407 meticulously crafted tasks, spanning seven major financial sub-domains: securities, funds, banking, insurance, futures, trusts, and asset management. These tasks are organized into three hierarchical levels of scenario depth: basic business analysis, asset decision support, and strategic risk management. We evaluated 10 mainstream AI agents in a zero-shot setting. The best-performing agent, ChatGPT, achieved an overall accuracy of 48.9%, which, while superior to non-professionals, still lags financial experts by over 35 percentage points. Error analysis has revealed five recurring failure patterns: Cross-modal Alignment Deficiency, Financial Terminological Bias, Operational Process Awareness Barrier, among others. These patterns point to crucial directions for future research. Our work provides the first agent benchmark closely related to the financial domain, aiming to objectively assess and promote the development of agents in this crucial field. Partial data is available at <https://github.com/SUFE-AIFLM-Lab/FinGAIA>

## Introduction

The rapid evolution of AI agents is revolutionizing how various industries tackle complex and intricate problems (Yao et al. 2023; Sumers et al. 2023; Ahn et al. 2022; Schick et al. 2023). Their application mode, which involves autonomously planning and achieving goals based on large language models (Zhang et al. 2024; Li, Zhang, and Sun 2023; Rafailov et al. 2023; Jing et al. 2019; Nijkamp et al. 2022), has garnered widespread public attention. The empowering role of agents is particularly significant in the financial sector, where they are crucial for automating the processing of vast amounts of data, analyzing market trends, and optimizing decision-making processes. This ultimately helps financial institutions enhance efficiency and mitigate risks. Such integration necessitates that agents possess a comprehensive suite of core capabilities, including a deep

understanding of industry knowledge, proficient tool utilization, and integrated reasoning abilities to handle complex tasks. However, due to the stringent requirements and specialized nature of the financial industry, effectively evaluating the true capabilities of AI agents presents unique challenges.

While researchers have gradually introduced specialized benchmarks for the financial domain, such as FinEval (Guo et al. 2024), FinQA (Chen et al. 2021), CFLUE (Zhu et al. 2024), CFinBench (Nie et al. 2024) and MME-Finance (Gan et al. 2024), these benchmarks remain predominantly text-centric or image-centric. They generally lack systematic testing for multi-file, multi-tool collaboration, and multi-step execution flows, which are critical capabilities required for agents to empower the financial sector. Similarly, most existing AI benchmarks exhibit significant shortcomings when evaluating agents' capabilities in the financial domain. Common general-purpose benchmarks often focus on text-based question answering, function synthesis, or tool-use tasks (Starace et al. 2025; Zheng et al. 2025; Kokane et al. 2025; Yao et al. 2022; Jimenez et al. 2023; Ruan et al. 2023; Kokane et al. 2024; Chang et al. 2024; Liu et al. 2024; Yao et al. 2025). Although effective in measuring a model's language understanding and coding abilities, they fail to meet the assessment demands of complex business processes within financial scenarios.

To address the gaps in current work and advance the development of financial AI agents, we propose FinGAIA the first end-to-end agent evaluation benchmark specifically tailored for financial scenarios. FinGAIA systematically integrates three core dimensions: industry knowledge, tool utilization, and task complexity, covering the entire financial workflow from basic information retrieval and multi-modal document analysis to code-based computation and the coordination of multi-tool, multi-step decision-making. FinGAIA comprises a total of 407 tasks, encompassing seven major financial sub-domains: securities, funds, banking, insurance, futures, trusts, and asset management. These tasks are designed based on extensive discussions with financial domain experts and constructed using real-world financial data, undergoing professional screening and structuring to ensure their authenticity, relevance, and evaluation utility. We adopt a hierarchical evaluation framework, categorizing tasks into basic business analysis, asset decision support,

\*Corresponding Author.

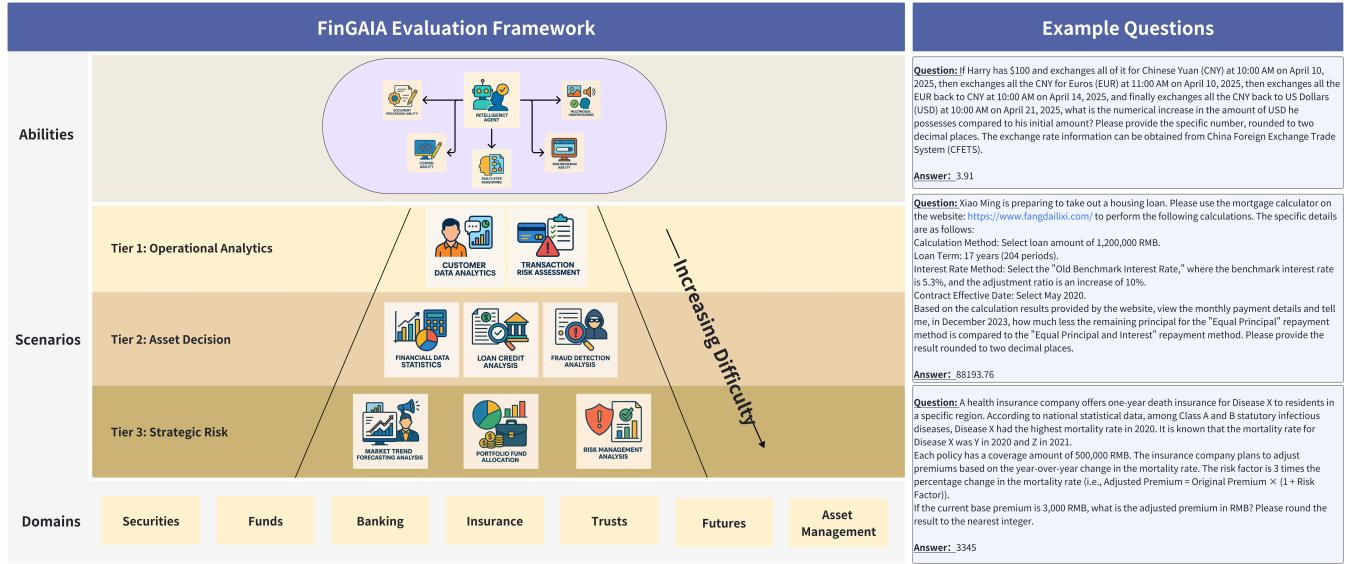


Figure 1: FinGAIA offers an AI Agent evaluation framework for full-process financial operations. From the perspective of business depth, it designs three major scenarios: Basic Business Analysis, Asset Decision Support, and Strategic Risk Control Management. Each primary scenario corresponds to 2, 3, and 3 sub-scenarios respectively, accurately reflecting the diverse business scenarios within the financial industry. The upper part of the image displays the 7 common financial sub-domains covered by the questions FinGAIA has constructed. The middle part illustrates the overall structure of FinGAIA, where business depth increases with the change in color. Concurrently, the demands on the model's understanding and analytical capabilities regarding financial business operations progressively increase. The side panel shows specific examples corresponding to the sub-scenarios, while the lower part details the specific capabilities required for the AI Agent to solve problems.

and strategic risk management, corresponding to different cognitive complexities and operational requirements. In our experimental evaluation, we tested 10 mainstream AI Agent frameworks, including both closed-source and locally deployed open-source agents. The evaluation employed a strict “zero-shot prompting” paradigm, supplemented by manual review and LLM-as-Judge for result verification.

In summary, our work makes the following contributions:

- **The First AI Agent Benchmark in the Financial Domain** We constructed FinGAIA, which includes 407 meticulously designed tasks based on real-world scenario data, covering seven major financial sub-domains and three levels of scenario depth. This fills the evaluation gap for agents in the financial sector, providing a comprehensive and in-depth assessment system for evaluating agents’ financial capabilities.
- **High-Quality and Practice-Oriented Dataset** All tasks in FinGAIA were formulated through discussions with financial experts, and the entire question creation process was executed manually. On average, each question required approximately 90 minutes to complete the full design, annotation, and verification process from extracting data from real financial scenarios. This rigorous construction process and quality control ensure the dataset’s authority while providing a practical, workflow-oriented evaluation framework for eight major scenarios.
- **Multi-Level Comparative Analysis** Our work includes a comparison of agent capabilities against those of ordinary financial undergraduate students and Ph.D. students

in finance, offering valuable insights into agents’ abilities in professional domains. The results indicate that agents still have room for improvement in specialized areas.

The remainder of this paper is structured as follows. Section 2 reviews related work on AI Agents and Agent benchmarks. Section 3 details the construction of FinGAIA, including data collection, question design, and quality control. Sections 4 and 5 present the experimental setup and results under different difficulty levels, followed by comparative and error analyses. Section 6 concludes the paper and discusses future directions for financial intelligence agents.

## Related Work

### Agent Development History

In recent years, Agents have evolved from early rule-based expert systems (Weizenbaum 1966; Barker et al. 1989) into end-to-end execution frameworks capable of autonomous learning and multi-tool invocation. In the late 1990s, data-driven decision-making was introduced via collaborative filtering (Resnick et al. 1994) and Q-Learning (Sutton, Barto et al. 1998). With the advent of deep learning, visual pre-training (Krizhevsky, Sutskever, and Hinton 2012) and large-scale language pre-training (Devlin et al. 2019; Radford et al. 2019) were integrated into Agent architectures, greatly enhancing their multimodal perception and understanding. Methodologically, the “reasoning–execution–feedback” loop proposed by ReAct (Yao et al. 2023) and Toolformer (Schick et al. 2023), together with the closed-loop toolchain orchestration of Auto-GPT

Benchmarks	Question Type	MLD	SD	MT	DE	MS	FD	NOA	NoQ
ToolBench	MC	✗	✗	✓	✗	✓	✗	5	126,486
API-Bank	OE	✗	✗	✗	✗	✓	✗	4	274
WebArena	MC	✗	✓	✗	✓	✓	✗	1	812
AgentBench	MC	✗	✓	✗	✓	✓	✗	27	1,091
GAIA	MC+OE	✓	✓	✓	✓	✓	✗	6	466
<b>FinGAIA (ours)</b>	<b>MC+OE</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>	<b>10</b>	<b>407</b>

Table 1: Comparison of Agent Benchmarks. Abbreviations in the header are: MC (Multiple-Choice questions), OE (Open-Ended questions), T/F (True/False questions), NoQ (Number of Questions), NOA (Number of Agents), MLD (Multi-level Difficulty), SD (Scenario Depth), FD (Financial Domains), MT (Multi-tool Support), DE (Dynamic Environment), MS (Multi-step Reasoning).

Scenario Depth	Financial Scenario	Questions
<b>Operational Analytics</b>	Customer Data Analytics	47
	Transaction Risk Assessment	42
	All	<b>89</b>
<b>Asset Decision</b>	Financial Data Statistics	101
	Loan Credit Analysis	43
	Fraud Detection Analysis	41
	All	<b>185</b>
<b>Strategic Risk</b>	Risk Management Analysis	42
	Portfolio Fund Allocation	40
	Market Trend Forecasting	51
	All	<b>133</b>
<b>FinGAIA</b>	All	<b>407</b>

Table 2: Distribution of Financial Scenarios. This table systematically presents the distribution of FinGAIA tasks across three progressively complex financial scenario categories. Specifically, the Basic Business Analysis category contains 89 tasks, Asset Decision Support includes 185 tasks, and Strategic Risk Management covers 133 tasks, resulting in a total of 407 carefully annotated tasks. This structured presentation enables precise evaluation of the benchmark’s ability to simulate the increasing complexity of real-world financial environments through tiered task design.

(Significant-Gravitas 2023), endowed Agents with high-level autonomy in scheduling and invoking multiple tools. At the same time, modular frameworks such as LangChain (Chase 2023) and AgentVerse (Chen et al. 2023) have substantially lowered the barrier to building and deploying multi-tool Agents.

In the financial domain, Agents have followed a similarly rapid trajectory: beginning with the first Robo-Advisor platforms in 2008 that automated portfolio allocation (Betterment 2008), progressing through the emergence of domain-specific pre-trained models tailored to financial corpora (Yang, Uy, and Huang 2020), and culminating in recent end-to-end decision systems that integrate real-time market data retrieval, model-based reasoning, and explainable report generation—thereby driving full-process intelligent financial decision-making (Huang et al. 2024).

### Agent Benchmark Analysis

The advent of agentic architectures—capable of multi-step reasoning, dynamic tool chaining, and multimodal interac-

tion—has outstripped the capabilities of traditional benchmarks built for static QA or simple API calls. Early platforms such as ToolBench (Qin et al. 2023) and API-Bank (Li et al. 2023) validate a model’s proficiency in predefined API invocations but rely on highly structured, closed environments that may encourage pattern memorization over genuine reasoning. WebArena (Zhou et al. 2023) and AgentBench (Liu et al. 2023) evaluate robustness in simulated web and system-interaction scenarios yet depend heavily on familiarity with specific DOM structures or command-line syntaxes. GAIA (Mialon et al. 2023) represents a significant advance by driving agents to interact end-to-end with live web tools—encompassing web browsing, document parsing, and code execution—to assess tool collaboration, long-chain reasoning, and self-correction. However, GAIA’s scope is largely confined to general web and document operations and does not address finance-specific requirements such as regulatory review, real-time market data retrieval, multimodal report parsing, or proprietary system API integration.

In order to bridge these gaps, we propose FinGAIA, the first end-to-end agent evaluation benchmark explicitly tailored to financial scenarios. FinGAIA systematically integrates three core dimensions—industry knowledge, tool utilization, and task complexity—and spans the full financial workflow: from basic information retrieval and multimodal document analysis to Python-based computation and orchestrated multi-tool, multi-step decision making. By capturing the intricacies of real-world financial processes, FinGAIA effectively addresses the limitations of existing benchmarks and provides a comprehensive framework for assessing agentic performance in professional financial contexts.

## FinGAIA

### Overview

We propose FinGAIA, the first benchmark specifically designed to evaluate the capabilities of intelligent agents within the financial domain. FinGAIA aims to systematically measure agents' task execution performance and professional competence in realistic financial workflows. Centered on highly faithful simulations of real-world financial scenarios, the benchmark comprehensively spans tasks ranging from basic business comprehension to complex strategic decision-making. It focuses on assessing agents' abilities in financial language understanding, data integration and analysis, tool-based collaboration, and multi-step reasoning. The overall architecture is illustrated in Figure 1. In terms of data generation and quality assurance, FinGAIA follows a rigorous construction pipeline. Four finance professors designed the task scenarios and authored the initial questions. These were annotated by six professionally trained undergraduate students with a solid background in finance, and then reviewed across multiple dimensions by four domain experts. The full dataset comprises 407 questions, covering a wide range of subdomains including securities, funds, and banking. These tasks are designed to reflect typical user demands in asset allocation, policy interpretation, risk assessment, and market analysis.

FinGAIA establishes a three-tiered capability evaluation framework, encompassing three difficulty levels and eight representative financial business scenarios. Each level corresponds to a different cognitive requirement:

**Level 1 – Basic Business Analysis:** evaluates agents' understanding of fundamental financial knowledge and their ability to process multimodal financial information;

**Level 2 – Asset Decision Support:** focuses on tasks involving moderate complexity, requiring information integration, logical reasoning, and flexible tool use;

**Level 3 – Strategic Risk Management:** targets highly complex tasks requiring multi-tool coordination and strategic planning, to comprehensively assess reasoning ability and domain expertise under realistic financial conditions.

Overall, FinGAIA provides a structured, diverse, and contextually grounded benchmark for evaluating intelligent agents in the financial vertical, enabling comprehensive and objective assessment of their professional competencies, mainly including web browsing capability, document pro-

cessing capability, multimodal understanding, coding ability, and computational ability. The distribution of tasks across business scenarios and dataset statistics is shown in Table 2, and representative task examples are included in the Appendix 1.

### Question Generation and Quality Control

The FinGAIA dataset encompasses data derived from a wide range of real-world financial scenarios. Major data sources include legal and regulatory text, publicly available market transaction data, financial news and analytical reports, and key indicators of financial products. All real-world data used in the benchmark have been carefully verified to ensure they are free of copyright restrictions, with explicit source attribution provided in each task.

During the task construction process, to more accurately simulate the real-world application scenarios of intelligent agents in the financial domain, four professors with financial expertise designed eight representative financial business scenarios distributed across three progressively difficult levels. Tasks were manually created under each scenario to reflect the corresponding complexity. The resulting tasks were annotated by six professionally trained undergraduate students majoring in finance, all of whom possessed a solid foundation in financial knowledge. The training program covered key aspects such as verifying whether each task aligned with its designated difficulty level and scenario, annotating the correct answer, estimating task difficulty, identifying the tools involved, and assessing the time required for completion. Large-scale annotation was only conducted after the annotators passed a qualification test and evaluation, ensuring both consistency and high quality in the labeling process.

In the expert review stage, each task was thoroughly inspected by four industry professionals, each with over a decade of financial experience. Review criteria included alignment with real industry practice, logical soundness of the task structure, regulatory compliance, precision of financial terminology, and contextual relevance to the business scenario. A task was accepted into the dataset only if unanimously approved by all four experts, ensuring that each question was well-designed, unambiguous, and reflective of authentic financial operations. Through this rigorous data generation and quality control pipeline, we constructed FinGAIA, a high-quality benchmark for evaluating financial AI agents. The final dataset includes 407 expert-validated tasks. The distribution of tasks across the eight financial scenarios is presented in Table 2.

### FinGAIA Question Architecture

The FinGAIA benchmark is built upon representative financial business scenarios and establishes a three-tiered evaluation framework designed to systematically assess the overall performance of intelligent agents in realistic financial environments. The benchmark consists of 407 high-quality tasks, each rigorously validated by experts. Tasks are categorized into three ascending difficulty levels based on task complexity and the depth of required capabilities: Level 1 (Basic Business Analysis), Level 2 (Asset Decision Support), and

Level 3 (Strategic Risk Management). These three levels respectively cover 2, 3, and 3 distinct financial scenarios. As the difficulty level increases, both the types of questions and the complexity of tasks intensify, enabling a comprehensive evaluation of an agent's abilities in financial language understanding, information integration, tool utilization, and complex decision reasoning. Representative examples of each level are provided in the Appendix 1. Estimates of task difficulty and the average number of tool invocations per scenario are provided in the Appendix Table 5.

**Level 1: Basic Business Analysis** This level primarily evaluates an agent's grasp of fundamental financial knowledge and its ability to quickly process multimodal information. Tasks at this level are structurally simple, typically requiring no more than five steps and the use of only one or two tools. The focus is on reading, interpreting, and performing basic calculations over essential financial inputs such as customer data and transaction charts. Suitable for entry-level tasks like customer service or preliminary transaction screening, Level 1 includes two representative scenarios: Customer Data Analytics and Transaction Risk Assessment, comprising a total of 89 tasks.

**Level 2: Asset Decision Support** Level 2 focuses on medium-complexity analysis and decision-making tasks, requiring agents to demonstrate stronger information synthesis and strategy formulation capabilities. Tasks typically involve multiple data sources, increased reasoning steps from 5 to 7, and the integration of more than two tools. This level emphasizes logical reasoning, flexibility in tool usage, and deeper domain understanding, making it suitable for mid-level financial operations such as loan evaluation, asset assessment, and fraud detection. It covers three scenarios: Loan Credit Analysis, Financial Data Statistics, and Fraud Detection Analysis, with a total of 185 tasks.

**Level 3: Strategic Risk Management** Designed for high-stakes decision-making and advanced risk management, Level 3 features tasks that closely reflect real-world financial practices. These tasks involve a greater number of steps around 10 and require coordinated use of multiple tools, including sequential tool invocation and parameter tuning. Agents are expected to exhibit precise semantic comprehension and integration of financial expertise. This level comprehensively evaluates the agent's peak performance under complex financial conditions, with a particular focus on accuracy, robustness, and consistency in chain-of-thought reasoning. It includes three core business scenarios: Portfolio Fund Allocation, Risk Management Analysis, and Market Trend Forecasting, comprising 133 tasks.

By constructing a tiered evaluation system grounded in realistic financial workflows, FinGAIA offers a scientific, comprehensive, and practically relevant benchmark for assessing the capabilities of financial AI agents.

## Experiment

### Agents

We evaluated 10 currently mainstream Agent frameworks, consisting of 9 closed-source agents and 1 locally deployed open-source framework OWL (CAMEL-AI 2025). 9 closed-

source agents were tested via their web interfaces, including ChatGPT DeepResearch (OpenAI 2025), Perplexity DeepResearch (Perplexity AI 2025), Kimi (Moonshot AI 2023), Casheat DeepResearch (FinStep 2025), Grok DeeperSearch (xAI 2025), Gemini-2.5-pro (Google 2025), Reportify (Beijing Jisha Chengta Technology 2024), and Kompas AI (Kompas AI 2024). And AutoGLM (Zhipu AI 2024) was integrated via a browser plugin. Detailed information is provided in Appendix 2.

### Evaluation Methods

In the actual evaluation process, we designed a differentiated testing process for each Agent's tool usage capability, aiming to comprehensively assess its performance in end-to-end financial scenarios. For closed-source agents provided as web or application-based services, we conducted interactive tests through manual operation, strictly adhering to the "zero-shot prompt" paradigm, providing only the problem text and necessary attachments (e.g., Excel files, PDF reports), without offering any examples or guidance. For frameworks that support API calls, we submitted all test questions in bulk through a unified API interface and embedded a preset System Prompt within the request to guide the Agent to output standardized structured answers.

To ensure the accuracy and consistency of the output, we primarily relied on manual review, supplemented by LLM-as-Judge's automatic interpretation, to perform item-by-item validation of the system's output. Only when the output results are completely consistent with the standard answer in terms of value, format, and semantics is it considered correct (1 point). Otherwise, it is considered incorrect (0 points). In cases of file processing failure or unanswered questions, they are marked as unassessable (-1 point) to ensure the uniqueness and rigor of the evaluation results.

After completing all data collection and scoring, we computed each agent's accuracy—both across the full set of tasks and within the Level 1–3 stratified tasks—by dividing its total number of correct responses by the total number of questions. This method ensures a fair and reproducible comparison of each framework's end-to-end performance and tool-coordination capabilities in financial scenarios.

## Results

### Main Results

We evaluated 10 mainstream Agents, as shown in Table 3. Due to certain models' inability to read specific file types, some questions could not be effectively answered by certain models. Therefore, we excluded the results of models that could not be evaluated from the final assessment. Overall, the results showed no significant changes. For detailed information, please refer to Appendix Table 6.

The performance of financial AI agents varies significantly across different scenarios, demonstrating clear stratification. ChatGPT (DeepResearch) ranked first with an average weighted score (WA) of 48.9%, exhibiting exceptional cross-level stability, particularly excelling in Risk Management Analysis (RMA 60.0) within strategic planning scenarios and Loan Credit Analysis (LCA 58.1%) in tactical

Agent	Operational Analytics		Asset Decision			Strategic Risk			WA
	CDA	TRA	FDS	LCA	FRA	RMA	PFA	MKT	
ChatGPT(DeepResearch)	<b>44.7</b>	<b>57.1</b>	<b>50.5</b>	<b>58.1</b>	<b>39.0</b>	<b>57.1</b>	<b>47.5</b>	<b>37.3</b>	<b>48.9</b>
Perplexity DeepResearch	31.9	52.4	30.7	48.8	26.8	42.9	35.0	27.5	37.0
Cashcat DeepResearch	29.8	52.4	21.8	48.8	31.7	26.2	35.0	19.6	33.2
Kimi	25.5	38.1	23.8	46.5	24.4	35.7	17.5	17.7	28.6
Gemini-2.5-pro(DeepResearch)	34.0	38.1	26.7	39.5	12.2	38.1	25.0	11.8	28.2
Grok DeeperSearch	23.4	35.7	15.8	34.9	2.4	31.0	35.0	19.6	24.7
Reportify	14.9	26.2	25.7	30.2	26.8	26.2	20.0	19.6	23.7
OWL	19.2	31.0	5.9	34.9	14.6	31.0	22.5	15.7	21.8
Kompas AI	25.5	9.5	19.8	20.9	14.6	16.7	22.5	5.9	16.9
AutoGLM	6.4	11.9	11.9	14.0	12.2	16.7	20.0	11.8	13.1
<b>Avg. Score</b>	25.5	35.2	23.2	37.7	20.5	32.2	28.0	18.7	—
Human Expert	82.5	88.3	85.1	90.2	79.6	92.7	83.7	75.8	84.7

Table 3: Main Results. The higher the value in the table, the higher the accuracy of the financial AI model. The evaluation assesses performance across three capability tiers: Operational Analytics covering Customer Data Analytics(CDA) and Transaction Risk Assessment(TRA); Asset Decision comprising Financial Data Statistics(FDS), Loan Credit Analysis(LCA), and Fraud Detection Analysis(FDA); and Strategic Risk evaluating Risk Management Analysis(RMA), Portfolio Fund Allocation(PFA), and Market Trend Forecasting(MTF), concluding with the calculation of the Weighted Average (WA) score for each model.

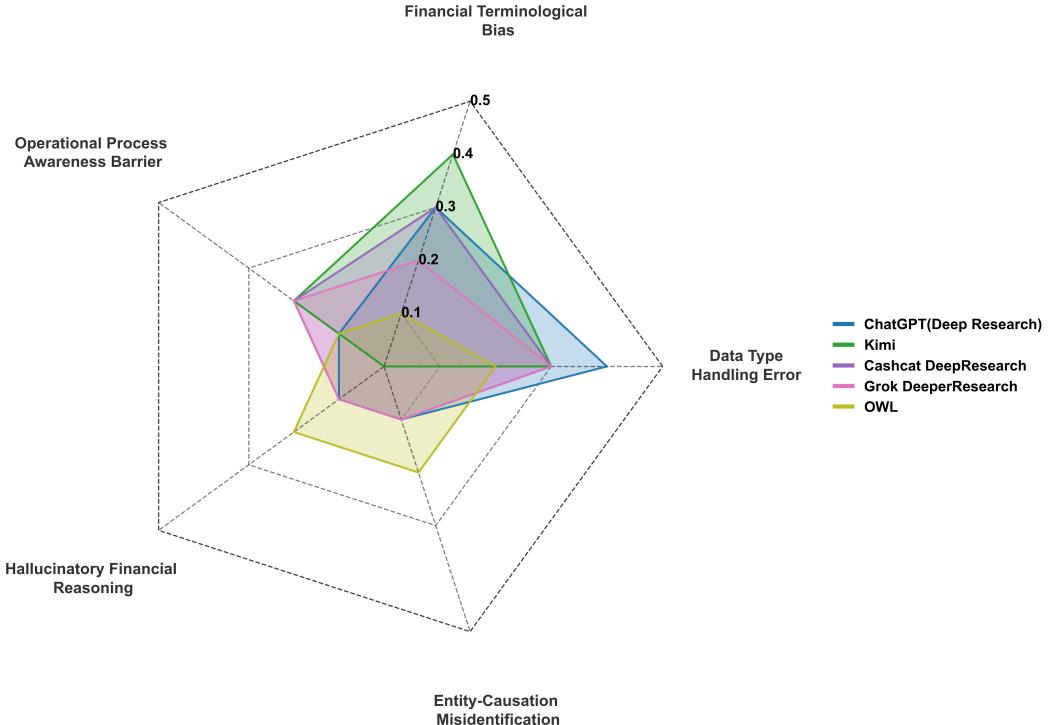


Figure 2: Error types' distribution across different Agents in FinGAIA tasks.

advisory scenarios. The second tier, consisting of Perplexity DeepResearch (WA 37%) and Cashcat DeepResearch(WA 33.2%), achieved an average score of 52.4% in Transaction Risk Assessment (TRA) for basic operational tasks, significantly exceeding the industry average by 17.2 points.

Key scenario analysis reveals that market forecasting and fraud detection remain common industry weaknesses, with only ChatGPT demonstrating breakthrough performance in fraud detection (FRA 39.0%). Further analysis indicates that the poorest-performing model, Grok DeeperSearch (FRA

2.4%), suffers from severe cross-modal alignment deficiencies and anomaly pattern recognition blind spots.

Although ChatGPT (DeepResearch) achieved the highest accuracy in this evaluation, this score remains significantly lower than that of human experts, particularly in Strategic Risk tasks, where agent performance was generally weak. This highlights the current limitations of large models in complex reasoning and deep comprehension tasks. It also suggests that while existing agents have made considerable progress in simpler or moderately complex problems, their performance in more challenging high-difficulty tasks still has substantial room for improvement. Future research and development must further enhance these models' reasoning capabilities, depth of problem-solving, and understanding of complex scenarios.

## Comparative Analysis

To better compare agent capabilities and make substantive contributions to model research, we conducted a competition among agents, non-experts, and financial experts by stratifying and sampling 50 questions from the FinGAIA dataset. Considering differences in domain knowledge and to more accurately reflect the current development stage of agents, we selected the top-performing models from both open-source and closed-source categories for comparison. For human participants, we recruited undergraduate students without financial backgrounds to represent the non-expert group, while the role of financial experts was filled by PhD candidates specializing in finance. All participants were independent of any data annotation or evaluation processes related to this study, and all responses were completed without AI assistance.

As shown in Table 4, unlike Table 3, we calculated the average results across three major scenarios and the overall score to compare human and model performance. The results demonstrate that the best-performing Agent models have surpassed non-experts in all three scenarios and in the composite average score. However, there remains a performance gap of over 32% between current agents and financial experts, indicating that agents still require continuous iterative optimization.

Source	Category	OA	AD	SR	Average
Human	Non-experts	58.3	47.1	35.2	<b>46.9</b>
	Experts	87.6	82.4	76.5	<b>82.2</b>
Closed-Source	ChatGPT (DeepResearch)	68.2	52.4	28.2	<b>49.6</b>
	Perplexity DeepResearch	51.7	40.3	22.9	<b>38.3</b>
Open-Source	OWL	25.5	20.8	15.2	<b>20.4</b>

Table 4: Performance comparison across non-experts, experts and Agents. OA refers to Operational Analytics. AD means Asset Decision. SR represents Strategic Risk.

## Error Analysis

For error analysis, we employed stratified sampling to select 50 of incorrect responses across all evaluated agents, systematically investigating deficiencies in agents' financial

business capabilities. Through comprehensive error categorization, we identified six fundamental limitations in financial domain competence: Data Type Handling Error, Financial Terminological Bias, Operational Process Awareness Barrier, Hallucinatory Financial Reasoning, and Entity-Causation Misidentification. These deficiencies collectively constrain model performance in specialized financial scenarios.

As evidenced by the error distribution analysis, most agents demonstrate relatively balanced error patterns. However, certain models (notably Kimi, ChatGPT(Deep Research), and OWL) exhibit statistically significant error concentrations in specific dimensions: Financial Terminological Bias , Data Type Handling Error , and Entity-Causation Misidentification, respectively. These systematic biases result in disproportionately higher error rates in these categories compared to baseline expectations.Detailed methodological explanations, including sampling procedures and significance testing protocols, along with representative error case studies, are documented in Appendix 2.

## Conclusion

This paper introduces FinGAIA, the first end-to-end benchmark meticulously designed to evaluate the capabilities of AI agents in the financial domain. The benchmark is structured around seven major financial sub-areas and organized into a three-tiered complexity hierarchy to systematically explore various agent capabilities.

Zero-shot evaluation results show that the best-performing model, ChatGPT (DeepResearch), achieved an overall accuracy of 48.9%. While this performance surpasses that of human finance undergraduates, there remains a significant gap when compared to finance PhDs. This highlights a substantial room for improvement, especially in highly complex, multi-step tasks. Furthermore, our comprehensive error analysis reveals five major capability deficiencies in agents when applied to real-world financial scenarios. As the first agent benchmark rooted in actual financial industry workflows, FinGAIA provides a structured, multi-level evaluation framework for measuring agents' end-to-end capabilities in financial practice.

We hope that the release of FinGAIA, along with its high-quality and diverse tasks, will advance agent research, foster a deeper understanding of real-world financial scenarios, and lay a solid foundation for the future development of more powerful and reliable financial intelligent agents. We plan to make the FinGAIA benchmark and its detailed evaluation results publicly available to encourage broader research community participation in exploring and enhancing financial agent capabilities.

## Limitations

Despite the significant progress FinGAIA has made in evaluating AI Agents in the financial domain, some limitations remain, requiring further research and development. Although certain FinGAIA tasks may inherently involve dynamic elements, there is a need for more explicit and in-depth exploration of time-series analysis, real-time market fluctuations,

and rapidly evolving economic indicators. Secondly, the current evaluation primarily focuses on zero-shot performance. While this provides a baseline understanding of an agent's out-of-the-box capabilities, it does not fully capture the potential for AI Agents to adapt and improve with limited examples. Future research should further consider the potential for AI Agents to adapt to new or changing financial tasks and scenarios through few-shot learning. Lastly, although FinGAIA covers a wide range of financial sub-domains and business scenarios, different business scenarios have varying importance. Therefore, it is necessary to design more appropriate business scenario weights to truly evaluate the actual performance of AI Agents in the financial domain.

## References

- Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.
- Barker, V. E.; O'Connor, D. E.; Bachant, J.; and Soloway, E. 1989. Expert systems for configuration at Digital: XCON and beyond. *Communications of the ACM*, 32(3): 298–318.
- Beijing Jisha Chengta Technology. 2024. Reportify. <https://reportify.ai/>. Web interface.
- Betterment. 2008. Betterment (company). [https://en.wikipedia.org/wiki/Betterment\\_\(company\)](https://en.wikipedia.org/wiki/Betterment_(company)).
- CAMEL-AI. 2025. OWL. <https://github.com/camel-ai/owl>. Locally deployed open-source framework.
- Chang, M.; Zhang, J.; Zhu, Z.; Yang, C.; Yang, Y.; Jin, Y.; Lan, Z.; Kong, L.; and He, J. 2024. Agentboard: An analytical evaluation board of multi-turn llm agents. *Advances in neural information processing systems*, 37: 74325–74362.
- Chase, H. 2023. LangChain: framework for developing applications powered by language models.
- Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Qian, C.; Chan, C.-M.; Qin, Y.; Lu, Y.; Xie, R.; et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4): 6.
- Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.-H.; Routledge, B.; et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.
- FinStep. 2025. Cashcat DeepResearch. <https://cashcat.cn/?mode=DeepResearch>. Cashcat Pro.
- Gan, Z.; Lu, Y.; Zhang, D.; Li, H.; Liu, C.; Liu, J.; Liu, J.; Wu, H.; Fu, C.; Xu, Z.; et al. 2024. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *arXiv preprint arXiv:2411.03314*.
- Google. 2025. Gemini-2.5-pro. <https://deepmind.google/models/gemini/pro/>. Web interface.
- Guo, X.; Xia, H.; Liu, Z.; Cao, H.; Yang, Z.; Liu, Z.; Wang, S.; Niu, J.; Wang, C.; Wang, Y.; Liang, X.; Huang, X.; Zhu, B.; Wei, Z.; Chen, Y.; Shen, W.; and Zhang, L. 2024. FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models. *arXiv preprint arXiv:2308.09975*.
- Huang, J.; Xiao, M.; Li, D.; Jiang, Z.; Yang, Y.; Zhang, Y.; Qian, L.; Wang, Y.; Peng, X.; Ren, Y.; et al. 2024. Openfinllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*.
- Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.
- Jing, M.; Ma, X.; Huang, W.; Sun, F.; and Liu, H. 2019. Task transfer by preference-based cost learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2471–2478.
- Kokane, S.; Zhu, M.; Awalgaonkar, T.; Zhang, J.; Hoang, T.; Prabhakar, A.; Liu, Z.; Lan, T.; Yang, L.; Tan, J.; et al. 2024. Spectool: A benchmark for characterizing errors in tool-use llms. *arXiv preprint arXiv:2411.13547*.
- Kokane, S.; Zhu, M.; Awalgaonkar, T. M.; Zhang, J.; Prabhakar, A.; Hoang, T. Q.; Liu, Z.; RN, R.; Yang, L.; Yao, W.; et al. 2025. ToolScan: A Benchmark For Characterizing Errors In Tool-Use LLMs. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Kompas AI. 2024. Kompas AI. <https://website.kompas.ai/>. Web interface.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Li, M.; Zhao, Y.; Yu, B.; Song, F.; Li, H.; Yu, H.; Li, Z.; Huang, F.; and Li, Y. 2023. Api-bank: A comprehensive benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*.
- Li, Y.; Zhang, Y.; and Sun, L. 2023. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.
- Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.
- Liu, Z.; Yao, W.; Zhang, J.; Yang, L.; Liu, Z.; Tan, J.; Choubey, P. K.; Lan, T.; Wu, J.; Wang, H.; et al. 2024. Agentlite: A lightweight library for building and advancing task-oriented llm agent system. *arXiv preprint arXiv:2402.15538*.
- Mialon, G.; Fourrier, C.; Wolf, T.; LeCun, Y.; and Scialom, T. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.

- Moonshot AI. 2023. Kimi. <https://moonshotai.github.io/Kimi-K2/>. Web interface.
- Nie, Y.; Yan, B.; Guo, T.; Liu, H.; Wang, H.; He, W.; Zheng, B.; Wang, W.; Li, Q.; Sun, W.; et al. 2024. CFinBench: A Comprehensive Chinese Financial Benchmark for Large Language Models. *arXiv preprint arXiv:2407.02301*.
- Nijkamp, E.; Pang, B.; Hayashi, H.; Tu, L.; Wang, H.; Zhou, Y.; Savarese, S.; and Xiong, C. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*.
- OpenAI. 2025. ChatGPT DeepResearch. <https://openai.com/index/introducing-deep-research/>. Web interface.
- Perplexity AI. 2025. Perplexity DeepResearch. <https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>.
- Qin, Y.; Liang, S.; Ye, Y.; Zhu, K.; Yan, L.; Lu, Y.; Lin, Y.; Cong, X.; Tang, X.; Qian, B.; et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.
- Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; and Riedl, J. 1994. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 175–186.
- Ruan, Y.; Dong, H.; Wang, A.; Pitis, S.; Zhou, Y.; Ba, J.; Dubois, Y.; Maddison, C. J.; and Hashimoto, T. 2023. Identifying the risks of lm agents with an lm-emulated sandbox. *arXiv preprint arXiv:2309.15817*.
- Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551.
- Significant-Gravitas. 2023. Auto-GPT. <https://github.com/Significant-Gravitas/AutoGPT>. GitHub repository.
- Starace, G.; Jaffe, O.; Sherburn, D.; Aung, J.; Chan, J. S.; Maksin, L.; Dias, R.; Mays, E.; Kinsella, B.; Thompson, W.; et al. 2025. PaperBench: Evaluating AI’s Ability to Replicate AI Research. *arXiv preprint arXiv:2504.01848*.
- Sumers, T.; Yao, S.; Narasimhan, K.; and Griffiths, T. 2023. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.
- Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.
- Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1): 36–45.
- xAI. 2025. Grok DeeperSearch. <https://x.ai/news/grok-3>. Web interface.
- Yang, Y.; Uy, M. C. S.; and Huang, A. 2020. Finbert: A pre-trained language model for financial communications. *arXiv preprint arXiv:2006.08097*.
- Yao, S.; Chen, H.; Yang, J.; and Narasimhan, K. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35: 20744–20757.
- Yao, S.; Shinn, N.; Razavi, P.; and Narasimhan, K. R. 2025. tau-bench: A benchmark for Tool-Agent-User interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Zhang, J.; Lan, T.; Zhu, M.; Liu, Z.; Hoang, T.; Kokane, S.; Yao, W.; Tan, J.; Prabhakar, A.; Chen, H.; et al. 2024. xlam: A family of large action models to empower ai agent systems, 2024. URL <https://arxiv.org/abs/2409.03215>.
- Zheng, X.; Lin, H.; He, K.; Wang, Z.; Fu, Q.; Fu, H.; Zheng, Z.; and Liang, Y. 2025. MCU: An Evaluation Framework for Open-Ended Game Agents. In *Forty-second International Conference on Machine Learning*.
- Zhipu AI. 2024. AutoGLM. <https://zhipuai.cn/>. Browser plugin.
- Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.
- Zhu, J.; Li, J.; Wen, Y.; and Guo, L. 2024. Benchmarking Large Language Models on CFLUE—A Chinese Financial Language Understanding Evaluation Dataset. *arXiv preprint arXiv:2405.10542*.

## Details of FinGAIA

### Design and Examples of Financial Business Scenarios

We have detailed the FinGAIA data in Table 2. We adopted a three-tiered structure for financial business scenarios, including basic business analysis, asset decision support, and strategic risk management. The detailed information for specific financial business scenarios is presented below.

Business Analysis includes the following two financial business scenarios: **Customer Data Analysis** primarily involves reviewing customer information to understand their data. The core of this work is to identify patterns and insights that inform targeted marketing strategies, product development, and personalized customer experiences. Through this analysis, financial institutions can optimize their offerings and build stronger customer relationships. Below is an example task for Customer Data Analysis. **Transaction Risk Assessment** focuses on identifying and evaluating potential risks associated with financial transactions. This includes analyzing the likelihood of default, fraud, or other financial losses. The objective of this work is to mitigate risks

by implementing appropriate controls and making informed decisions on transaction approvals.

Asset Decision Support includes three core business scenarios: **Financial Data Statistics** involves collecting, organizing, and interpreting numerical financial information. The main task is to derive meaningful insights and trends from large datasets using statistical methods. This work supports reporting, forecasting, and identifying anomalies within financial operations. **Loan Credit Analysis** is the process of assessing the creditworthiness of loan applicants. It involves evaluating financial statements, credit history, and other relevant factors to determine the likelihood of repayment. The goal is to make informed lending decisions that balance risk and potential returns. **Fraud Detection Analysis** focuses on identifying and preventing fraudulent activities within financial systems. This includes analyzing transaction patterns and anomalies to flag suspicious behavior. The goal of this work is to protect institutions and their clients from financial crimes and losses.

Strategic Risk Management includes three high-level financial task scenarios: **Risk Management Analysis** is a broad discipline focused on identifying, assessing, and mitigating various financial risks. This work involves developing strategies and controls to minimize potential negative impacts on an institution's financial health. **Portfolio Fund Allocation** involves strategically distributing capital among different assets within an investment portfolio. The primary task is to optimize the portfolio's risk-return profile based on investment objectives and market conditions. This ensures efficient capital utilization, thereby maximizing returns while managing acceptable levels of risk. **Market Trend Forecasting** focuses on predicting future movements and directions in financial markets. This includes analyzing historical data, economic indicators, and various other factors to anticipate market changes. The goal of this work is to provide valuable insights for investment decisions, trading strategies, and risk management.

Figure 3, 4 are the examples of Business Analysis. Figure 5, 6, 7 are the examples of Asset Decision Support. Figure 8, 9, 10 are the examples of Strategic Risk Management.

## Details of Agents

We list details of the AI Agent evaluated using FinGAIA in Table 7.

## Details of Evaluation Results

**Operational Analytics** The assessment examines agent performance in two fundamental financial operations tasks: Customer Data Analytics (CDA) and Transaction Risk Assessment (TRA). As shown in the bar chart (11), human experts achieved scores of 82.5 in CDA and 88.3 in TRA, demonstrating consistent professional proficiency. In contrast, the top-performing ChatGPT (DeepResearch) scored 57.1 in TRA, indicating a significant performance gap in basic operational tasks. Notably, all tested agents showed weaker performance in TRA tasks requiring multi-system collaboration, with AutoGLM scoring only 6.4 in CDA, highlighting current architectural limitations in cross-system

coordination among agents. The average score of the Top 3 Agents was 45.1, 4.8 times higher than that of the Bottom 3. This pronounced stratification underscores disparities in agent tool utilization efficiency, while the polarized performance distribution reveals clear capability tiers among agents, emphasizing the importance of specialized model optimization.

**Asset Decision** In the Asset Decision section, the bar chart data 12 illustrates the performance across three decision-making tasks: Financial Data Statistics (FDS), Loan Credit Analysis (LCA), and Fraud Detection (FRA). Human experts maintained a comprehensive lead with an average score of 86.7, while the best-performing ChatGPT Agent scored 49.2, with a relatively strong showing in Loan Credit Analysis (LCA: 58.1). The Agents exhibited distinct specialization patterns: Perplexity DeepResearch performed better in Financial Data Statistics (FDS: 30.7), whereas Cashcat DeepResearch had a slight edge in Fraud Detection (FRA: 31.7). This phenomenon of professional differentiation indicates that a single Agent struggles to cover all decision-making needs, suggesting that specialized division of labor and collaboration may be a more viable application approach.

**Strategic Risk** The scenario encompasses three critical financial sub-tasks: Risk Management Analysis (RMA), Portfolio Fund Allocation (PFA), and Market Trend Forecasting (MTF). According to the data, human experts achieved an average score of 84.0 across these sub-tasks, significantly outperforming the top-performing AI agent, ChatGPT (DeepResearch), which scored 41.5 on average. Notably, AI agents demonstrated the weakest performance in PFA tasks, with an average score of only 29.2, highlighting the current models' substantial limitations in complex asset allocation decisions that require comprehensive consideration of multiple risk factors. For detailed charts, please refer to Figure 13.

## Examples for Error Analysis

In this section, we explain in detail the meaning of five types of errors of Agents in financial business scenarios and provide examples and related error analysis.

**Entity-Causation Misidentification** The agent incorrectly establishes causal relationships between entity characteristics and business outcomes in financial analysis. It mistakes superficial correlations for fundamental drivers or confuses the sequential logic of business processes. This error reveals the agent's structural deficiencies in reconstructing financial business logic chains. An example of this can be seen in Figure 14.

**Financial Terminological Bias** The agent demonstrates comprehension flaws in professional financial terminology systems. It confuses regulatory definitions of similar terms, misapplies calculation logic, or disregards the context sensitivity of terminology. Such systematic misuse of terms leads the agent to make derivative calculation errors. An example of this can be seen in Figure 15.

**Operational Process Awareness Barrier** The agent exhibits cognitive obstacles regarding standardized financial business processes. It misinterprets operational requirements of regulatory rules, omits key compliance steps, or

reverses the sequence of business execution. This error exposes the agent's lack of practical knowledge in financial operations. An example of this can be seen in Figure 16.

**Hallucinatory Financial Reasoning** The agent generates false financial propositions without reliable evidence. It produces factual hallucinations, logical hallucinations, and data hallucinations. This error stems from the agent's failure in financial fact-checking mechanisms, potentially leading to severely misleading outputs. An example of this can be seen in Figure 17.

**Data Type Handling Error** The agent triggers a data type handling error when the type or format of the input data falls outside its supported range (e.g., video files, executable programs, etc.), rendering it unable to execute the processing task. This error indicates that the system lacks the capability to parse or execute specific data types under its current configuration, representing a functional limitation rather than a logical error. The performance of each model after excluding unsupported question types is shown in Table 6.

Scenario Depth	Financial Scenario	Tool Count
<b>Operational Analytics</b>	Customer Data Analytics	2.06
	Transaction Risk Assessment)	2.14
<b>Asset Decision</b>	Financial Data Statistics	2.63
	Loan Credit Analysis	1.98
	Fraud Detection Analysis	2.29
<b>Strategic Risk</b>	Risk Management Analysis	2.02
	Portfolio Fund Allocation	2.2
	Market Trend Forecasting	2.63

Table 5: The table presents the average number of tool invocations for each financial scenario under the three difficulty levels defined in the FinGAIA dataset. These statistics provide an intuitive reflection of FinGAIA’s progressive task design logic—from simple to complex—facilitating more accurate evaluation of an agent’s performance and adaptability across different financial contexts.

Agent	Operational Analytics		Asset Decision			Strategic Risk			WA
	CDA	TRA	FDS	LCA	FRA	RMA	PFA	MKT	
ChatGPT(DeepResearch)	<b>44.7</b>	<b>57.1</b>	<b>50.5</b>	<b>58.1</b>	<b>39.0</b>	<b>60.0</b>	<b>47.5</b>	<b>37.3</b>	<b>49.3</b>
Perplexity DeepResearch	37.5	53.7	33.3	51.2	29.7	42.9	36.8	29.2	39.3
Kimi	38.7	48.5	36.4	52.6	30.3	41.7	24.1	29.0	37.7
Gemini-2.5-pro(DeepResearch)	38.1	41.0	27.8	42.5	15.2	44.4	29.4	15.0	31.7
Reportify	18.0	30.6	29.2	32.5	30.6	29.7	24.2	27.0	27.7
Grok DeeperSearch	25.0	37.5	18.4	35.7	2.9	33.3	37.8	21.3	26.5
Kompas AI	35.3	12.9	27.8	23.7	18.8	20.6	32.1	9.4	22.6
OWL	19.2	31.7	6.1	34.9	14.6	31.0	23.1	15.7	22.0
AutoGLM	11.1	20.0	17.9	20.0	20.8	25.0	34.8	25.0	21.8
<b>Avg. Score</b>	29.9	38.0	26.5	40.2	22.8	35.9	33.0	23.1	–
Human Expert	82.5	88.3	85.1	90.2	79.6	92.7	83.7	75.8	84.7

Table 6: The evaluation results of financial AI agents in assessable scenarios. The higher the value in the table, the higher the accuracy of the financial AI agent. The evaluation assesses performance across three capability tiers: Operational Analytics covering Customer Data Analytics(CDA) and Trans-action Risk Assessment(TRA); Asset Decision comprising Financial Data Statistics(FDS), Loan Credit Analysis(LCA), and Fraud Detection Analysis(FDA); and Strategic Risk evaluating Risk Management Analysis(RMA), Portfolio Fund Allocation(PFA), and Market Trend Forecasting(MTF), concluding with the calculation of the Weighted Average (WA) score for each agent. Cashcat DeepResearch is removed since it does not throw error message even for unsupported files to avoid unfair comparison.

Category	Agent	Creator	Access
<b>Close-Source</b>	ChatGPT(DeepResearch)	OpenAI	Web UI
	Perplexity DeepResearch	Perplexity AI	Web UI
	Kimi	Moonshot AI	Web UI
	Cashcat DeepResearch	FinStep	Web UI
	Grok DeeperSearch	xAI	Web UI
	Gemini-2.5-pro(DeepResearch)	Google	Web UI
	Reportify	Beijing Jisha Chengta Technology	Web UI
	AutoGLM	Zhipu AI	Plugin
	Kompas AI	Kompas AI	Web UI
<b>Open-Source</b>	OWL	CAMEL-AI	Weights

Table 7: Evaluated Agents. The "Access" column indicates the interaction mode available for each agent (e.g., Web UI, Plugin, or direct model weights access).



问题：识别附件中的图片，这是一家期货公司的图标。访问其官网。在“营业机构模块”搜索：河北省沧州市营业网点的负责人是谁？  
Question: Identify the image in the attachment—it is the logo of a futures company. Visit its official website and, under the “Business Institutions” section, locate the outlet in Cangzhou City, Hebei Province. Who is the person in charge of that outlet?

答案：冷俊杰  
Answer: Leng Junjie

Figure 3: This is a Customer Data Analytics scenario example focusing on branch manager information retrieval—the Agent must identify the futures company’s logo shown in the figure, navigate to its official website’s “Business Institutions” module, search for the Cangzhou outlet in Hebei Province, and locate the name of the person in charge to answer, “Who is the manager of the Cangzhou outlet?” This task tests the Agent’s operational analytics capabilities in logo recognition, website navigation, information retrieval, and rapid response, demonstrating the value of Operational Analytics Tier.

第十条 期货交易所应当依照本条例和国务院期货监督管理机构的规定，建立、健全各项规章制度，加强对交易活动的风险控制和对会员以及交易所工作人员的监督管理。期货交易所履行下列职责：

- (一) 提供交易的场所、设施和服务；
- (二) 设计合约，安排合约上市；
- (三) 组织并监督交易、结算和交割；
- (四) [REDACTED]
- (五) 按照章程和交易规则对会员进行监督管理；
- (六) 国务院期货监督管理机构规定的其他职责。

期货交易所不得直接或者间接参与期货交易。未经国务院期货监督管理机构审核并报国务院批准，期货交易所不得从事信托投资、股票投资、非自用不动产投资等与其职责无关的业务。

问题：小明有一张截图关于期货交易管理条例（2017年修订）得部分内容，但是由于部分内容被遮挡，他想知道被遮挡得内容是什么？你可以在中国期货业协会官网找到相关文件。

Question: Xiao Ming has a screenshot of part of the Regulations on Futures Trading Management (2017 Revision), but some of the content is obscured. He would like to know what the obscured text says. You can find the full document on the China Futures Association's official website.

答案：为期货交易提供集中履约担保

Answer: Provide centralized performance guarantees for futures trading

Figure 4: This is a Transaction Risk Assessment scenario example focusing on a task to complete the text of Article 10, paragraph (4) of the Regulations on Futures Trading Management (2017 Revision)—the Agent must identify the screenshot's source and clause, navigate to the China Futures Association's "Regulations and Documents" section, download the full text, locate Article 10 (4), and extract the obscured content to answer the user's question. This tests the Agent's foundational business analysis capabilities in regulatory document retrieval, clause localization, and text extraction, demonstrating the value of Operational Analytics.

问题：附件中是某个德国保险公司的标志，在其官网上有2024年9月停售的“XX安享稳盈终身寿险”（XX为公司名称），通过python读取该产品费率表数据，计算该表中每一列数据（年龄列不参与计算）的变异系数（标准差 / 均值），告诉我变异系数从大到小，排名第三的是哪一列数据，回答格式如：保险类型A型；男性；趸交。

Question: The attachment shows the logo of a German insurance company. On its official website there is the “XX Secure Stable Return Whole Life Insurance” (where XX is the company name) that was discontinued in September 2024. Using Python, read the product’s rate-table data and compute the coefficient of variation (standard deviation divided by mean) for each column in the table (excluding the Age column). Which column ranks third when the coefficients of variation are sorted from highest to lowest? Please provide your answer in the format: Insurance Type A; Male; Single Premium.



答案：保险类型B型；男性；趸交

Answer: Insurance Type B; Male; Single Premium.

Figure 5: This is a financial data statistics task. To answer this question, the Agent must first identify the German insurance company's logo in the figure, then navigate to its official website to locate the rate table for the “XX Secure Stable Return Whole Life Insurance” product, use Python to read the table, compute the coefficient of variation for each column (excluding “Age”), sort these coefficients in descending order, and select the column that ranks third. This question tests the Agent’s ability to aggregate and parse financial data tables, calculate and rank basic statistical metrics (coefficient of variation), and carry out multi-step information retrieval and programming execution, highlighting a rate sensitivity analysis scenario at the Asset Decision Tier.

问题：小明有一个课堂练习需要你的帮助来进行核验，小明模拟是上市公司（证券代码为：000001）的一名员工，他需要代表公司向某银行申请经营贷款，在他提交的材料中涉及以下内容：

- 1 本公司2024年股东权益为1,304,712百万元，较年初变化6.2%
- 2 本公司归属于母公司股东的净利润在2024年第一季度小于2024年第四季度
- 3 本公司第八大股东是深业集团有限公司，并且该股东在报告期内增持股票
- 4 本公司签字会计师有两位，姓名是：吴翠蓉，范玉军

请基于上市公司（证券代码为：000001）公布的2024年度报告，对小明材料中的信息进行核验，无法确认正确的材料均认为有误，如果有多个信息有误，则返回有问题材料的编号，使用分号隔开；如果所有信息正确无误，则返回0。

Question: Xiaoming has a classroom exercise that requires your help for verification. He is role-playing as an employee of a publicly listed company (stock ticker: 000001) and needs to apply for an operating loan from a bank on behalf of the company. In the materials he submitted, he makes the following claims:

1. The company's shareholders' equity for 2024 was RMB 1,304,712 million, representing a 6.2% change from the beginning of the year.
2. The net profit attributable to the parent-company shareholders in Q1 2024 was less than that in Q4 2024.
3. The company's eighth-largest shareholder is China Shenyne Group Co., Ltd., and this shareholder increased its shareholding during the reporting period.
4. The company's signing auditors are Wu Cuirong and Fan Yujun.

Please verify these statements against the company's 2024 Annual Report (ticker: 000001). Any statement that cannot be confirmed as correct should be considered erroneous. If there are multiple incorrect statements, return the numbers of the erroneous items separated by semicolons; if all statements are correct, return 0.

答案：2;3

Answer: 2;3

Figure 6: A Loan Credit Analysis example where the Agent consults the listed company's 2024 Annual Report to verify four claims—total shareholders' equity and its year-on-year change; Q1 vs. Q4 net profit; the identity and stake change of the eighth-largest shareholder; and the signing auditors' names—and returns any incorrect item numbers, testing data extraction, comparative analysis, and text-matching skills at the Asset Decision Tier.

问题：请基于附件音频内容（当日早新闻回放），按照以下顺序分析：

- 1 新闻中提及的“公司动向”涉及的所有企业；
- 2 仅保留上市公司，过滤未上市主体；
- 3 对比各上市公司对应证券的当日涨幅（以该音频中播报时间为标准）；
- 4 输出最高涨幅证券的6位数字代码。

Question: Based on the attached audio content (this morning's news replay), perform the following analyses in order:

- 1 Identify all companies mentioned in the “company developments” segment of the news.
- 2 Retain only those that are publicly listed, filtering out any non-listed entities.
- 3 For each remaining listed company, compare the stock’s intraday percentage gain as of the time the audio was broadcast.
- 4 Output the six-digit stock code of the security with the highest intraday gain.

答案：300750

Answer: 300750

Figure 7: A Fraud Detection Analysis example where, based on the day’s morning news replay audio, the Agent must 1) identify all companies mentioned under “company developments”; 2) filter for listed companies only; 3) retrieve and compare each company’s stock’s intraday gain at the broadcast time; and 4) output the six-digit code of the top-gaining stock. This tests the Agent’s abilities in audio parsing, entity recognition, data filtering, market-data retrieval, and result generation at the Asset Decision Tier.

问题：请你通过一段Python代码，使用akshare库获取沪深300、香港恒生指数和标普500指数的历史日线数据，计算这三个指数在2014年3月1日到2024年3月1日之间的最大回撤（Drawdown）。通过使用 cummax() 函数计算每个交易日的历史最高点，再计算每个交易日的回撤率，最终得到最大回撤值并输出（按照沪深300，香港恒生指数，标普500指数顺序输出结果，两个结果之间用逗号间隔，每个结果保留两位小数，不需要输出具体的代码，只需要输出最后三个指数的最大回撤的保留两位小数的具体值）。

Question: Please use a Python snippet with the AkShare library to fetch daily historical data for the CSI 300, the Hong Kong Hang Seng Index, and the S&P 500 Index, and calculate each index's maximum drawdown between March 1, 2014 and March 1, 2024. Compute each trading day's historical peak using cummax(), then calculate the drawdown rate for each day, and finally determine the maximum drawdown. Output only the three maximum drawdown values—in the order CSI 300, Hang Seng, S&P 500—separated by commas, each rounded to two decimal places. Do not include the code itself, only the three rounded drawdown values.

答案：46.70%,55.70%,33.92%

Answer: 46.70%,55.70%,33.92%

Figure 8: This is a Risk Management Analysis (RMA) scenario example, focusing on using Python and the akshare library to fetch daily historical data for the CSI 300, the Hang Seng Index, and the S&P 500 from March 1, 2014 to March 1, 2024, then using the cummax() function to compute each trading day's historical high and drawdown rate, and finally outputting the maximum drawdown values for the three indices . This tests the Agent's comprehensive capabilities in long-term market risk assessment and decision support at the Strategic Risk Tier.



问题：X为图片中动物的种类数目，在2021.04.01-2022.04.01期间，跌幅排名第X的基金在2021年持仓中，占净值比例最高的证券代码是多少？只回答证券代码中数字部分即可。

Question: Let X be the number of animal species shown in the image. Between April 1, 2021 and April 1, 2022, for the fund whose decline ranked Xth over that period, what is the security code with the highest proportion of net asset value in its 2021 holdings? Please answer with only the numeric part of the security code.

答案: 3690

Answer: 3690

Figure 9: This is a Portfolio Fund Allocation scenario example, focusing on dynamic parameter mapping and portfolio holding analysis—the Agent must first identify the number of animal species X shown in the image, then rank funds by their drawdowns from April 1, 2021 to April 1, 2022, select the fund with the Xth largest drawdown, retrieve that fund’s 2021 holdings, identify the security with the highest net-asset proportion, and output only the numeric part of its security code. This task tests the Agent’s integrated abilities in visual information extraction, dynamic metric mapping, fund performance ranking, and portfolio holding analysis, illustrating the Strategic Risk Tier’s application value for asset allocation optimization and return forecasting.



问题：如图是东方财富中一支股票某5个交易日的涨跌情况，其中红色空心为涨，实心青色为跌，以该证券这五个交易日涨跌规律为标准，以下五支证券：000001, 000002, 000004, 000006, 000007 在2025年2月5日至2025年2月11日（包含2月11日在内共5个交易日）涨跌分布与图片中证券中一致的是哪些证券？回答6位数证券代码，若多个证券符合要求则全部回答使用分号隔开。

Question: Below is the five-trading-day rise/fall pattern of a stock on EastMoney, where hollow red bars denote up days and solid cyan bars denote down days. Using this stock's five-day pattern as the reference, which of the following five securities—000001, 000002, 000004, 000006, and 000007—exhibited the identical up/down sequence during the five trading days from February 5, 2025 through February 11, 2025 (inclusive)? Answer in six-digit codes; if multiple securities match, separate them with semicolons.

答案：000004;000006

Answer: 000004;000006

Figure 10: This is a Market Trend Forecasting (MKT: Market Trend Forecasting) scenario example, focusing on a multi-security screening task based on up/down pattern matching—the Agent must identify the target security's five-day rise/fall sequence (hollow red bars indicate up days; solid cyan bars indicate down days), then compare that sequence against securities 000001- 000007 over the five trading days from February 5 to February 11, 2025, selecting those whose patterns match exactly and returning their six-digit codes. This task tests the Agent's integrated abilities in time-series pattern recognition, pattern matching, and batch screening of multiple securities, illustrating the Strategic Risk Tier's value in historical trend benchmarking and forecasting.

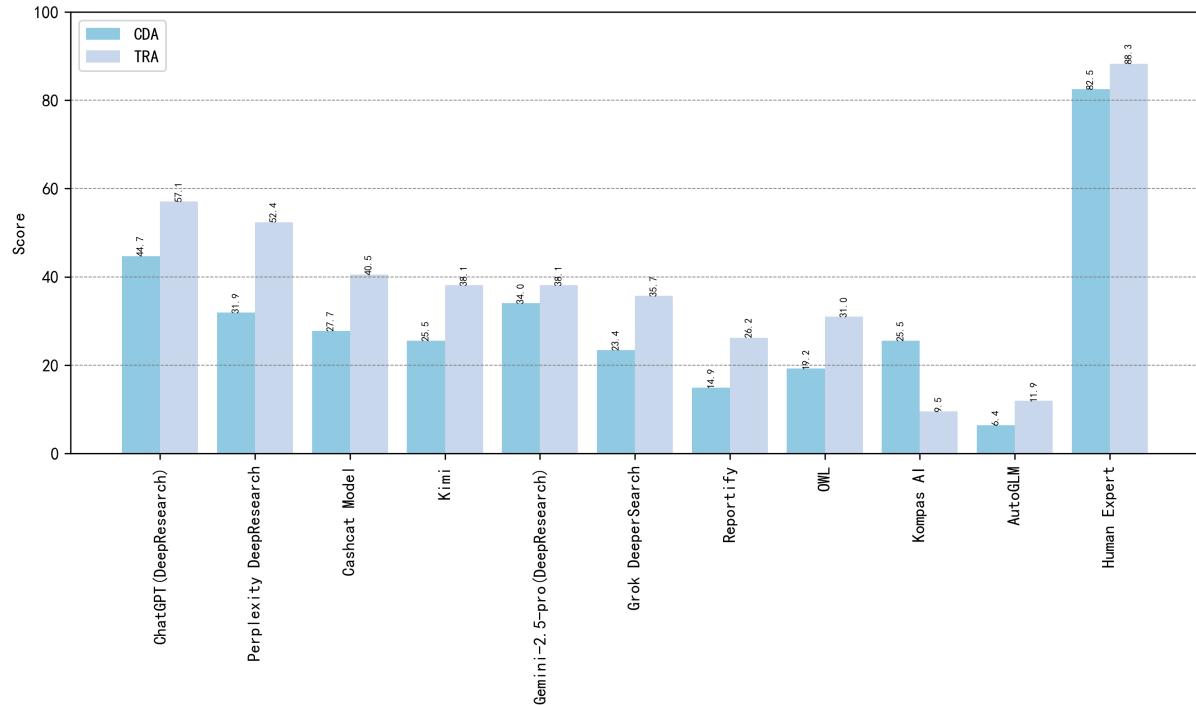


Figure 11: The chart displays a bar graph of the scores for various agents in Operational Analytics, including two scenarios: Customer Data Analytics (CDA) and Transaction Risk Assessment (TRA).

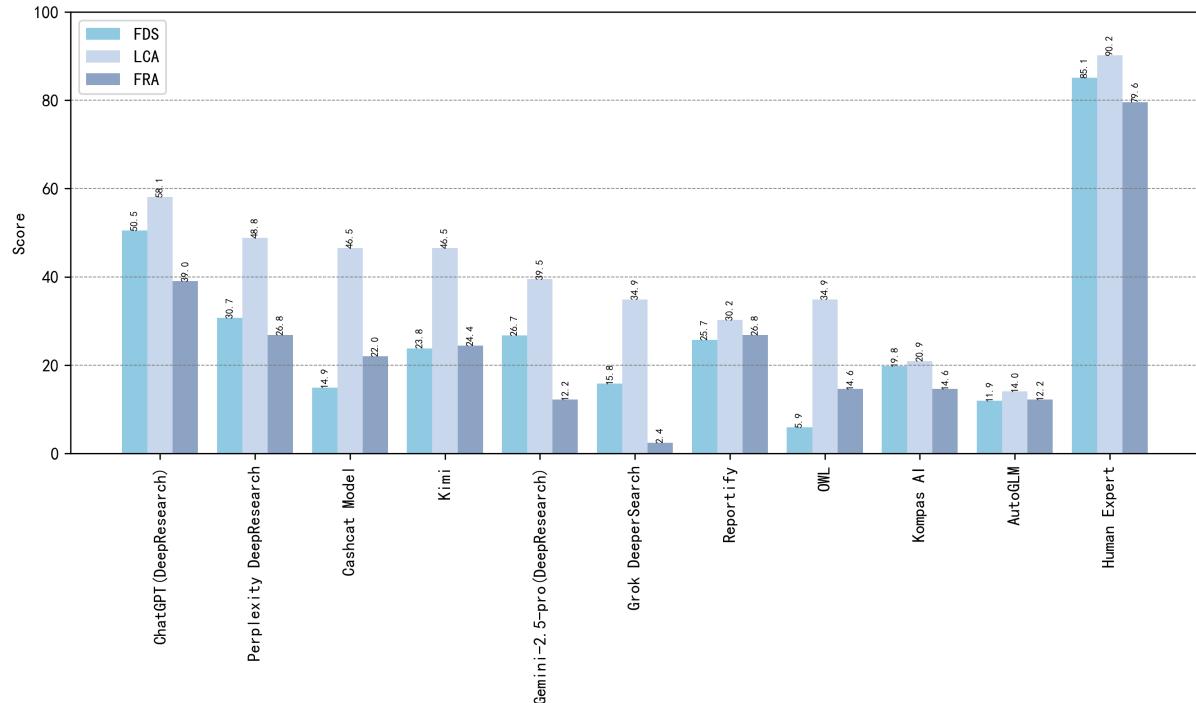


Figure 12: The chart presents a bar graph of the scores for various agents in Asset Decision, covering three scenarios: Financial Data Statistics (FDS), Loan Credit Analysis (LCA), and Fraud Detection Analysis (FRA).

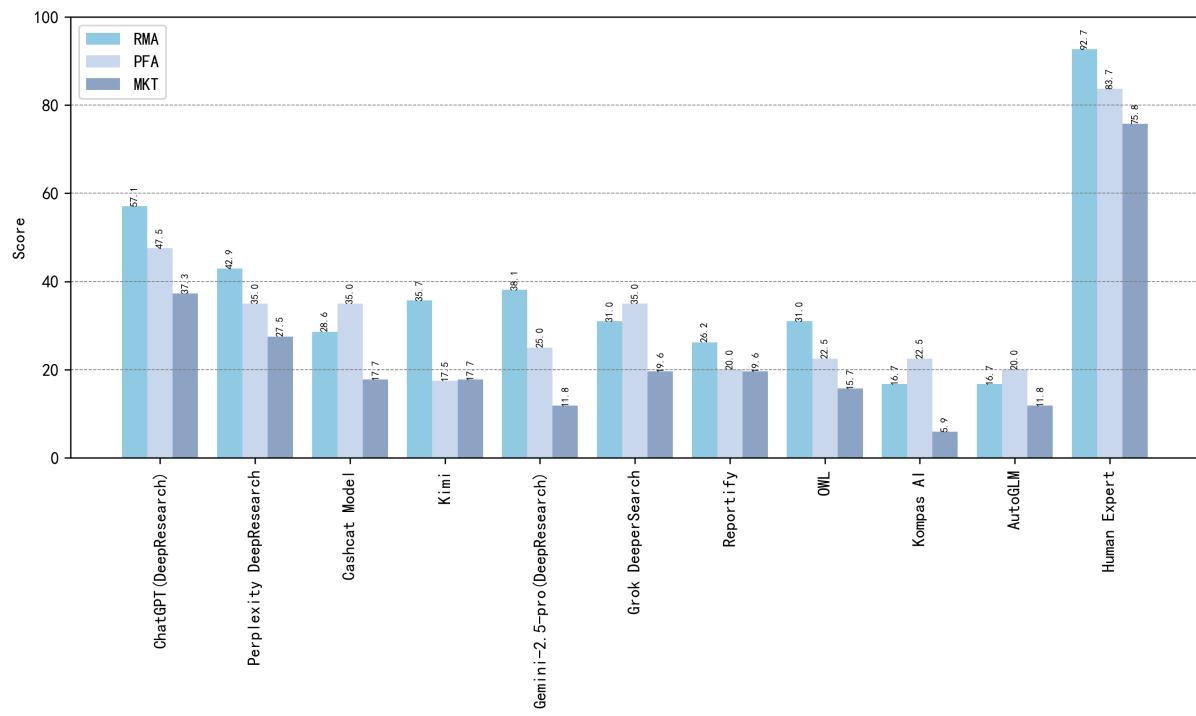


Figure 13: The chart displays a bar graph of the scores for various agents in Strategic Risk, comprising three scenarios: Risk Management Analysis (RMA), Portfolio Fund Allocation (PFA), and Market Trend Forecasting (MTF).

问题：有一个新股民，他想购买一只2024年03月20日看中的股票，但是忘记了股票的具体名称和代码，只记得股票是在沪市的，第二位和第三位数字相同，股票代码中第四位为5，股票名称中带有省市的名称以及金属的名字，2024年03月20日这只股票的收盘价为25块多，请你通过上面的提示帮助他找到这只股票，并帮他查看一下2024年5月6日的成交量和收盘价是多少，最终结果按照股票名称，成交量，收盘价（结果保留两位小数）的顺序输出。

Question: There is a new stock investor who wants to buy a stock that he liked on March 20, 2024, but he has forgotten the specific name and code of the stock. He only remembers that the stock is in the Shanghai Stock Exchange, the second and third digits are the same, the fourth digit in the stock code is 5, the stock name contains the name of the province and the name of the metal, and the closing price of this stock on March 20, 2024 was more than 25 yuan. Please help him find this stock through the above tips, and help him check the trading volume and closing price on May 6, 2024. The final result is output in the order of stock name, trading volume, closing price (the result is rounded to two decimal places).

正确答案：山东黄金, 402447, 29.27

Correct answer: Shandong Gold, 402447, 29.27

模型答案：天士力, 1200000, 25.05

Model answer: Tasly, 1200000, 25.05

模型思考：

根据您提供的信息，以下是我对这只股票的分析和推测：

#### ❶ 股票代码分析

- 沪市股票代码规则：沪市A股股票的代码通常以 60 开头，共6位数字。
- 第二位和第三位数字相同：这意味着股票代码的第二位和第三位数字相同，如 60xxx。
- 第四位为5：这意味着股票代码的第四位数字为5，如 60x5xx。

结合以上条件，符合条件的股票代码可能是 600535。

#### ❷ 股票名称分析

- 带有省市名称：股票名称中包含某个省市的名称。
- 带有金属名称：股票名称中包含金属的名称。

根据这些提示，600535 对应的股票名称是 天士力。

#### ❸ 2024年5月6日的成交量和收盘价

根据公开数据，天士力 (600535) 在 2024年5月6日 的收盘价为 25.05元，成交量为 1,200,000股。

#### ❹ 最终结果

- 股票名称：天士力； - 成交量：1,200,000股； - 收盘价：25.05元

Model thinking:

Based on the information you provided, the following is my analysis and speculation on this stock:

#### ❶ Stock Code Analysis

- Shanghai Stock Code Rules: The code of Shanghai A-share stocks usually starts with 60 and has a total of 6 digits.
  - The second and third digits are the same: this means that the second and third digits of the stock code are the same, such as 60xxx.
  - The fourth digit is 5: This means that the fourth digit of the stock code is 5, such as 60x5xx.
- Combined with the above conditions, the stock code that meets the conditions may be 600535.

#### ❷ Stock name analysis

- With provincial and municipal names: The stock name contains the name of a province or city.
- With metal names: The stock name contains the name of a metal.

According to these tips, the stock name corresponding to 600535 is Tasly.

#### ❸ Trading volume and closing price on May 6, 2024

According to public data, the closing price of Tasly (600535) on May 6, 2024 was 25.05 yuan, and the trading volume was 1,200,000 shares.

#### ❹ Final result

- Stock name: Tasly; - Trading volume: 1,200,000 shares; - Closing price: 25.05 yuan

Error analysis: This error belongs to Entity-Causation Misidentification, because the Agent mistakenly regards the structural characteristics of the stock code (the second and third digits are the same, and the fourth digit is 5) as the direct causal factor in determining the stock name, ignoring the essential information of the province/city and metal name in the stock name, thus incorrectly attributing the relationship between the stock code and the stock name, and failing to accurately identify the correct stock.

Figure 14: An example of Entity-Causation Misidentification. Although the agent mistakenly identified the structural characteristics of the stock code (the second and third digits being the same and the fourth digit being 5) as the direct causal factor determining the stock name, it overlooked the essential information of the province/city and metal name in the stock name. This led to incorrect attribution of the relationship between the stock code and the stock name, ultimately failing to accurately identify the correct stock.

请编写Python代码，使用akshare获取易方达消费行业股票（110022）2023年全年度持仓数据，通过Barra CNE5模型计算风格因子暴露度（市值、价值、动量），并输出Q1/Q2/Q3/Q4季末规模因子暴露度变化幅度（百分比），保留两位小数。

Please write Python code to use akshare to obtain the full-year holdings data of E Fund Consumer Industry Stocks (110022) in 2023, calculate the style factor exposure (market value, value, momentum) through the Barra CNE5 model, and output the change in size factor exposure at the end of Q1/Q2/Q3/Q4 (percentage), retaining two decimal places.

正确答案：3.15%，-1.87%，+2.43%，-0.95%

Correct answer: 3.15%, -1.87%, +2.43%, -0.95%

模型回答：

要实现这个目标，我们需要从 akshare 获取易方达消费行业股票（110022）的持仓数据，并通过 Barra CNE5 模型计算市值、价值、动量的风格因子暴露度。接下来，我们将计算每个季度因子暴露度的变化幅度（百分比）。

首先，需要明确的是，Barra CNE5 模型是一个多因子模型，通常用于评估股票的市场风格，包含了市值 (Size)、价值 (Value)、动量 (Momentum) 等因子暴露度。akshare 提供了基金持仓报告，但没有直接提供 Barra 风格因子的接口。为了模拟这个过程，我们假设已有一些方式来获得每个季度的因子暴露度数据。

步骤：

1. 获取易方达消费行业股票的持仓数据。
2. 模拟 Barra CNE5 风格因子的暴露度数据（市值、价值、动量）。
3. 计算每个季度因子暴露度的变化幅度（百分比）。

Python代码实现

```
import akshare as ak
import pandas as pd
fund_data = ak.fund_holdings_report(symbol="110022", year=2023)
print(fund_data.head())
data = [
    '季度': ['Q1', 'Q2', 'Q3', 'Q4'],
    '市值因子暴露度': [0.10, 0.12, 0.14, 0.13], # 市值因子的暴露度
    '价值因子暴露度': [0.15, 0.14, 0.13, 0.16], # 价值因子的暴露度
    '动量因子暴露度': [0.05, 0.07, 0.08, 0.10] # 动量因子的暴露度
]
df = pd.DataFrame(data)
df['市值因子变化幅度'] = df['市值因子暴露度'].pct_change() * 100
df['价值因子变化幅度'] = df['价值因子暴露度'].pct_change() * 100
df['动量因子变化幅度'] = df['动量因子暴露度'].pct_change() * 100
df = df.round(2)
print(df[['季度', '市值因子变化幅度', '价值因子变化幅度', '动量因子变化幅度']])
```

输出结果示例：

季度	市值因子变化幅度	价值因子变化幅度	动量因子变化幅度
0 Q1	NaN	NaN	NaN
1 Q2	20.00	-6.67	40.00
2 Q3	16.67	-7.14	14.29
3 Q4	-7.14	23.08	25.00

Model answer:

To achieve this goal, we need to obtain the holdings data of E Fund Consumer Industry Stocks (110022) from akshare, and calculate the exposure of style factors of market value, value, and momentum through the Barra CNE5 model. Next, we will calculate the change in factor exposure each quarter (percentage).

First of all, it should be clear that the Barra CNE5 model is a multi-factor model, which is usually used to evaluate the market style of stocks, including factor exposures such as market value (Size), value (Value), and momentum (Momentum). akshare provides fund holdings reports, but does not directly provide an interface for Barra style factors. To simulate this process, we assume that there are some ways to obtain factor exposure data for each quarter.

Steps:

1. Get the holdings data of E Fund Consumer Industry Stocks.
2. Simulate the exposure data of Barra CNE5 style factors (market value, value, momentum).
3. Calculate the change in factor exposure each quarter (percentage).

Python code implementation

```
import akshare as ak
import pandas as pd
fund_data = ak.fund_holdings_report(symbol="110022", year=2023)
print(fund_data.head())
data = [
    'Quarter': ['Q1', 'Q2', 'Q3', 'Q4'],
    'Market capitalization factor exposure': [0.10, 0.12, 0.14, 0.13], # Market capitalization factor exposure
    'Value factor exposure': [0.15, 0.14, 0.13, 0.16], # Value factor exposure
    'Momentum factor exposure': [0.05, 0.07, 0.08, 0.10] # Momentum factor exposure
]
df = pd.DataFrame(data)
df['Market value factor change'] = df['Market capitalization factor exposure'].pct_change() * 100
df['Value factor change'] = df['Value factor exposure'].pct_change() * 100
df['Momentum factor change'] = df['Momentum factor exposure'].pct_change() * 100
df = df.round(2)
print(df[['Quarter', 'Market value factor change', 'Value factor change', 'Momentum factor change']])
```

Output example:

Quarter	Market value factor change	Value factor change	Momentum factor change
0 Q1	NaN	NaN	NaN
1 Q2	20.00	-6.67	40.00
2 Q3	16.67	-7.14	14.29
3 Q4	-7.14	23.08	25.00

Error Analysis: This mistake falls under Financial Terminological Bias because the model incorrectly referred to the Size factor (market capitalization factor) in the Barra CNE5 model as "Market Capitalization Factor," confusing standardized terminology in quantitative finance. This led to a deviation from the actual calculation logic required. Additionally, the model erroneously assumed that simulated data could be used directly to calculate factor exposure, overlooking the core principles of the Barra model, which require portfolio weighting and standardization processing. As a result, the final output of quarterly change percentages was inaccurate.

Figure 15: An example of Financial Terminological Bias. The agent incorrectly referred to the Size factor (market capitalization factor) in the Barra CNE5 model as "Market Capitalization Factor," confusing standardized terminology in quantitative finance models. This deviation in terminology led to miscalculations that diverged from actual requirements. Additionally, the model erroneously assumed that simulated data could be used directly to calculate factor exposure, overlooking the Barra model's core rules requiring portfolio weighting and standardization processing, ultimately resulting in incorrect quarterly change percentages.

问题：根据《2024年融资融券标的参数清单（六月修订版）》第三章表格数据，若北京机构客户以38.6元价格融资买入500股601888中国中免，客户保证金可用余额应减少多少元？

Question: According to the table data in Chapter 3 of the "2024 Margin Trading Parameter List (June Revised Edition)", if a Beijing institutional client buys 500 shares of 601888 China Duty Free at a price of 38.6 yuan, how much should the client's margin balance be reduced?

正确答案：13510元

Correct answer: 13,510 yuan

模型答案：8577.78元

Model answer: 8,577.78 yuan

模型思考：

好的，我将查找《2024年融资融券标的参数清单（六月修订版）》中第三章的表格数据，获取601888中国中免在北京地区机构客户的融资保证金比例，并据此计算若客户以38.6元价格融资买入500股时，应减少的保证金可用余额金额。

根据《2024年融资融券标的参数清单（六月修订版）》第三章的表格数据，证券代码601888（中国中免）在北京地区机构客户的融资保证金比例为 80%。也就是说，投资者融资买入中国中免股票时，需要按融资交易金额的80%交纳保证金（融资保证金比例=保证金/融资交易金额）。这一比例符合监管要求的最低80%标准。

保证金可用余额减少金额计算

假设客户以每股38.6元的价格融资买入500股中国中免股票，则交易总金额为：

- 交易总成本 = 38.6元/股 × 500股 = 19,300元。

设借入资金金额为L，自有保证金金额为D。根据融资保证金比例80%的要求，有关系式： $D = 0.8 \times L$ ，同时 $D + L = 19,300$ 元。解方程可得：

- 自有保证金  $D = 19,300 \times 0.8 / 1.8 \approx 8,577.78$  元。

- 借入资金  $L = 19,300$  元 - 8,577.78 元  $\approx 10,722.22$  元。

因此，客户此次融资买入操作将冻结约8,577.78元的保证金。这相当于客户保证金可用余额减少了8,577.78元（不考虑佣金、利息等其他费用）。该冻结金额为交易总额的80%，符合融资保证金比例的要求。

结论：客户保证金可用余额将减少约8,577.78元。

Model thinking:

OK, I will look up the table data in Chapter 3 of the "2024 Margin Trading Parameter List (June Revised Edition)" to obtain the margin ratio of 601888 China Duty Free institutional clients in Beijing, and calculate the amount of margin balance that should be reduced if the client buys 500 shares at a price of 38.6 yuan.

China Duty Free Group (601888) margin ratio and margin occupancy calculation

601888 margin ratio (Beijing institutional clients)

According to the table data in Chapter 3 of the "2024 Margin Trading Target Parameter List (June Revised Edition)", the margin ratio of securities code 601888 (China Duty Free) for institutional clients in Beijing is 80%. In other words, when investors buy China Duty Free shares through margin trading, they need to pay a margin of 80% of the margin transaction amount (margin ratio = margin/margin transaction amount). This ratio meets the minimum 80% standard required by regulators.

Calculation of the reduction in the available margin balance

Assuming that a customer buys 500 shares of China Duty Free shares through margin trading at a price of 38.6 yuan per share, the total transaction amount is:

- Total transaction cost = 38.6 yuan/share × 500 shares = 19,300 yuan.

Let the amount of borrowed funds be L and the amount of own margin be D. According to the requirement of 80% margin ratio, there is a relationship:  $D = 0.8 \times L$ , and  $D + L = 19,300$  yuan. Solving the equation, we can get:

- Own margin  $D = 19,300 \times 0.8 / 1.8 \approx 8,577.78$  yuan,

- Borrowed funds  $L = 19,300$  yuan - 8,577.78 yuan  $\approx 10,722.22$  yuan.

Therefore, the customer's margin purchase operation will freeze about 8,577.78 yuan of margin. This is equivalent to a reduction of 8,577.78 yuan in the customer's margin available balance (excluding commissions, interest and other fees). The frozen amount is 80% of the total transaction amount, which meets the requirement of the margin ratio.

Conclusion: The available balance of the customer's margin will be reduced by about 8,577.78 yuan.

Error analysis: This answer has a serious Operational Process Awareness Barrier, and the core problem is misunderstanding of margin margin rules. The question requires the calculation of the margin occupation amount under the regulatory caliber, and the formula of "margin amount × margin ratio" should be used directly. However, the big model mistakenly understands 80% as the ratio in the capital structure, and uses "two formulas" to derive own funds and borrowed funds, which is a typical financial rule cognitive error. The resulting 8,577.78 yuan is far lower than the correct value of 15,440 yuan, with a deviation of more than 44%, and the reasoning logic confuses regulatory rules and financial allocation, which is misleading.

Figure 16: An example of Operational Process Awareness Barrier. The core issue lies in misunderstanding margin financing rules. The question required calculating margin occupation amounts under regulatory standards using the straightforward formula of "financing amount × margin ratio." However, the AI model misinterpreted the 80% ratio as a capital structure allocation, employing a "dual-equation" approach to derive proprietary and borrowed funds - a classic case of financial rule cognition error. This resulted in a calculation of 8,577.78, significantly lower than the correct value of 15,440 (a deviation exceeding 44%), while the flawed reasoning conflated regulatory requirements with financial allocation, creating misleading conclusions.

中国建设银行官方在2024年3月15日发布的《消费者权益保护信息披露公告》中提及的违规业务类型和整改措施是什么?  
What are the types of illegal business and rectification measures mentioned in the "Consumer Rights Protection Information Disclosure Announcement" officially issued by China Construction Bank on March 15, 2024?

正确答案：信用卡分期业务，暂停新增客户3个月并全面自查  
Correct answer: Credit card installment business, suspend adding new customers for 3 months and conduct a comprehensive self-examination

模型思考：  
明白了，我将查找中国建设银行在2024年3月15日发布的《消费者权益保护信息披露公告》，提取其中所有提及的违规业务类型及对应的整改措施，并完整汇总反馈给你。  
建行消费者权益保护信息披露公告（2024年3月15日）违规类型及整改措施  
中国建设银行在2024年3月15日发布的《消费者权益保护信息披露公告》中提及了多项违规业务类型及相应的整改措施，主要包括以下几类：  
- 保险代理业务违规：违规情况：部分分支机构在同一年度内与超过监管允许数量的保险公司合作，违规通过储蓄柜台销售投资连结型保险产品，且代销收益不确定的保险产品时未按规定提供完整合同材料。整改措施：建行已严格落实监管要求，停止上述违规销售行为，规范保险代理业务合作数量和渠道，并确保向客户提供完整的合同和产品资料。  
- 违规收费问题：违规情况：存在违规向个人客户收取唯一账户年费和小额账户管理费的行为。整改措施：建行对相关违规收费已全面清理并予以退还，今后严格执行收费管理规定，杜绝此类违规收费再次发生。  
- 人员资格管理问题：违规情况：个别分行由未经资质许可的员工违规销售基金等理财产品。整改措施：建行已加强员工资格准入和培训管理，明确要求基金销售等业务须持证上岗，并对相关责任人员进行了严肃问责，确保此类违规行为已得到整改。  
每一类违规事项均已按照监管部门要求完成整改，建行表示将持续强化依法合规经营，完善内部控制机制，防止类似问题再次发生。各项整改措施的落实体现了建行对金融消费者权益保护的重视和改进承诺。

Model thinking:  
Got it, I will look up the "Consumer Rights Protection Information Disclosure Announcement" issued by China Construction Bank on March 15, 2024, extract all the types of illegal business and corresponding rectification measures mentioned therein, and give you a complete summary.  
CCB Consumer Rights Protection Information Disclosure Announcement (March 15, 2024) Violation Types and Rectification Measures  
China Construction Bank mentioned a number of illegal business types and corresponding rectification measures in the "Consumer Rights Protection Information Disclosure Announcement" issued on March 15, 2024, mainly including the following categories:  
- Insurance agency business violations: \*Violation\*: Some branches cooperated with insurance companies exceeding the number allowed by the regulator in the same year, illegally sold investment-linked insurance products through savings counters, and failed to provide complete contract materials as required when selling insurance products with uncertain returns on behalf of them. \*Corrective measures\*: CCB has strictly implemented regulatory requirements, stopped the above-mentioned illegal sales behavior, standardized the number and channels of insurance agency business cooperation, and ensured that complete contracts and product information are provided to customers.  
- Illegal charges: \*Illegal situation\*: There are illegal charges for individual customers for annual fees and small account management fees. \*Corrective measures\*: CCB has completely cleaned up and refunded the relevant illegal charges, and will strictly implement the fee management regulations in the future to prevent such illegal charges from happening again.  
- Personnel qualification management issues: \*Illegal situation\*: Some branches have illegal sales of funds and other financial products by employees without qualification permits. \*Corrective measures\*: CCB has strengthened employee qualification access and training management, clearly required that fund sales and other businesses must be certified, and held relevant responsible personnel accountable to ensure that such violations have been rectified.  
Each type of violation has been rectified in accordance with the requirements of the regulatory authorities. CCB stated that it will \*\*continue to strengthen legal and compliant operations\*\*, improve internal control mechanisms, and prevent similar problems from happening again. The implementation of various rectification measures reflects CCB's emphasis on and commitment to improvement in protecting the rights and interests of financial consumers.  
Error analysis: This mistake constitutes Hallucinatory Financial Reasoning, where the model fabricated non-announcement content such as insurance agency violations and confused the specific regulatory requirements and corrective measures for credit card installment business. This resulted in a complete mismatch with the actual violations disclosed in CCB's March 2024 announcement (only involving non-standard marketing practices in credit card installment business) and the required corrective actions (suspending new customer acquisition for 3 months).

Figure 17: An example of Hallucinatory Financial Reasoning: The agent fabricated non-existent content such as insurance agency violations, conflating them with the specific regulatory requirements and corrective measures for credit card installment business. This resulted in a complete mismatch with the actual violations disclosed in CCB's March 2024 announcement (which only involved non-compliant marketing practices in credit card installment business) and the corresponding rectification requirements (a 3-month suspension of new customer acquisition).