FinSight: Towards Real-World Financial Deep Research

Jiajie Jin[1*], Yuyao Zhang[1*], Yimeng Xu[1], Yutao Zhu[1], Hongjin Qian[2], Zhicheng Dou[1†]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]BAAI
{jinjiajie, dou}@ruc.edu.cn
https://github.com/RUC-NLPIR/FinSight

Abstract

Professional financial reports are the cornerstone of investment decision-making, demanding deep analytical reasoning and multimodal synthesis. While recent deep research systems excel in open-domain search tasks, they struggle with financial reporting, specifically in processing structured financial data, ensuring analytical depth, and integrating professional visualizations. To address this gap, we introduce `FinSight` (**Fin**ancial In**Sight**), the first multi-agent framework for end-to-end automation of professional multimodal financial reports. At its core, we propose a Code Agent with Variable Memory architecture, unifying financial data, domain tools, and agent modules into a programmable variable space to enable flexible data manipulation and reasoning via executable code. To guarantee report quality, `FinSight` adopts a Two-Stage Writing Framework with Generative Retrieval. It first distills raw data into structured Chain-of-Analysis segments, then synthesizes them into a coherent, citation-aware, and multimodal narrative complying with financial reporting norms. Additionally, an Iterative Vision-Enhanced Mechanism uses visual feedback to refine code-generated charts to expert standards. Experiments on company and industry-level tasks demonstrate that `FinSight` significantly outperforms state-of-the-art deep research systems in factual accuracy, analytical depth, and presentation quality, validating the effectiveness of our framework. Our code is available at `https://github.com/RUC-NLPIR/FinSight`.

1 Introduction

High quality financial research reports, characterized by deep analytical narratives interleaved with rich visualizations, are the cornerstone of investment decisions worth billions of dollars (Tian et al., 2025). These multimodal documents go beyond simple summaries by synthesizing raw market data into strategic insights for asset managers and institutional investors. However, producing such reports remains a challenging task due to the overwhelming volume of financial data and the demand for rapid, high-quality analysis (Ren et al., 2021; Jimeno-Yepes et al., 2024; Jin et al., 2025a). Recent advances in artificial intelligence, particularly in deep research applications (OpenAI, 2025a; Gemini, 2025; Grok, 2025; Camara, 2025), present great potential in automating these labor-intensive tasks. Despite these technical advances, significant challenges persist in automating the generation of full financial research reports that meet the high standards for data accuracy, analytical depth, and multimodal content integration (Yang et al., 2025; Dong et al., 2025a).

However, significant challenges persist in adapting these general methods for professional financial reporting. Most existing systems are designed for open-domain search, often **lacking the capability to integrate real time, heterogeneous financial data** (Hu et al., 2025; Li et al., 2025c, 2026). Furthermore, they **typically produce plain text outputs** that miss critical multimodal visualizations, such as dynamic charts and tables (Yang et al., 2025). Finally, the reliance on **rigid, single-pass workflows** limits their ability to dynamically adjust research strategies, resulting in insufficient analytical depth (Trivedi et al., 2023; Li et al., 2025a; Jin et al., 2025b, 2026).

To address these challenges, we introduce **`FinSight`**, a multi-agent system designed for professional, multimodal financial reporting. At the core of `FinSight` is the **Code Agent with Variable Memory (CAVM)**, which unifies heterogeneous data and tools into a programmable variable space, orchestrating three specialized agents for Data Collection, Analysis, and Report Generation. Built upon CAVM, we orchestrate three specialized agents responsible for Data Collection,

[*] Equal Contribution, [†] Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5868–5894

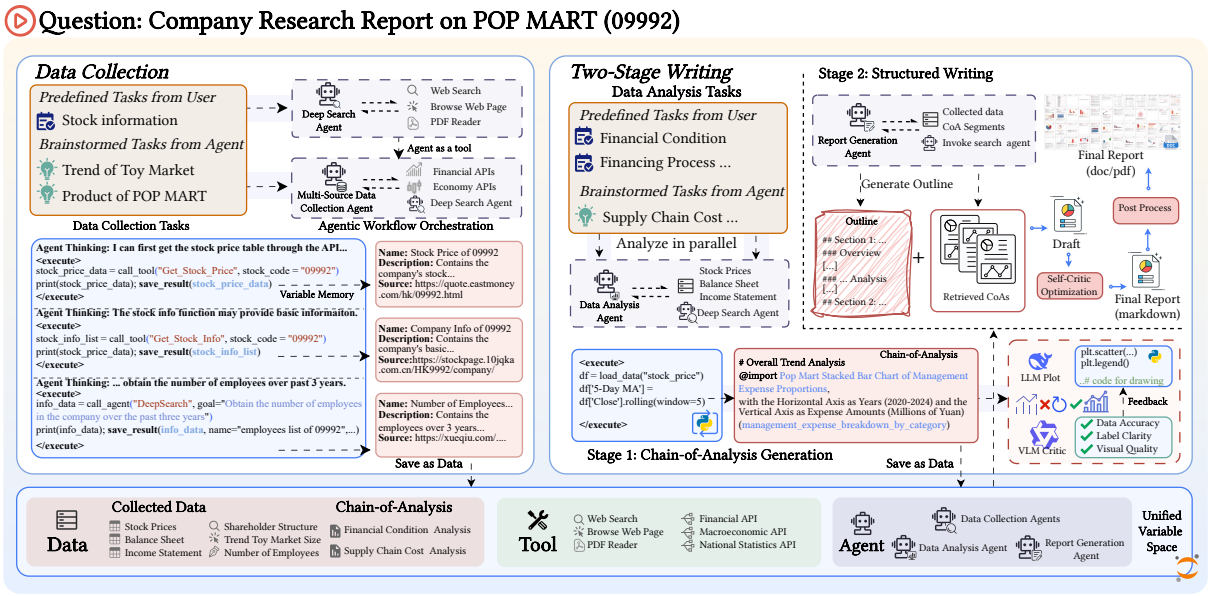

Figure 1: Overview of the FinSight Framework.

Data Analysis, and Report Generation. To guarantee professional quality, FinSight incorporates a **Two-Stage Writing Framework with Generative Retrieval**, which first distills findings into a structured Chain-of-Analysis (CoA) backbone before generating the full narrative with seamlessly interleaved citations and figures. Furthermore, an **Iterative Vision-Enhanced Mechanism** leverages VLM feedback to refine code-generated charts, ensuring they meet professional aesthetic standards.

We further construct a comprehensive benchmark covering company and industry-level tasks across multiple markets. We establish a rigorous evaluation protocol centered on Factual Accuracy, Analytical Depth, and Presentation Quality, employing both human and LLM-as-a-Judge assessments. Empirical results demonstrate that FinSight achieves state-of-the-art performance, surpassing existing methods in report accuracy, citation reliability, and visual integration.

Our core contributions are as follows:

1. To the best of our knowledge, we present the first exploration of the **Multimodal Deep Research Task** within the financial domain. We propose a system capable of generating professional-grade reports that combine deep textual analysis with rich visual elements, with empirical results demonstrating state-of-the-art performance on this task.

2. We propose the **Code Agent with Variable Memory (CAVM)** architecture, which unifies data, tools, and agents into a programmable variable space, enabling the flexible orchestration of complex, data-intensive report generation tasks.

3. We introduce two novel mechanisms to ensure high-quality output: an **Iterative Vision-Enhanced Mechanism** that refines code generated charts via visual feedback, and a **Two-Stage Writing Framework** that progresses from concise analytical chains to comprehensive, evidence-based narratives.

2 Method

2.1 Problem Formulation

We formalize the task of *Professional Financial Report Generation* a hierarchical and multimodal generation task. Given a research query q, the system aims to generate a structured report R. We model R as a hierarchical ordered sequence:

$$R = \{S_1, S_2, \ldots, S_N\},$$

where N is the number of sections derived from a dynamic outline $\mathcal{O}$. Each section S_i is further defined as an ordered sequence of multimodal elements $S_i = (e_{i,1}, e_{i,2}, \ldots, e_{i,m})$, where each element $e_{i,j} \in \{T, V, C\}$ represents text segments, visualization figures, or citations, respectively.

2.2 The Framework of FinSight

FinSight is a multi-agent framework designed to emulate the rigorous workflow of professional fi-

nancial analysts. As illustrated in Figure 1, the system operates through a hierarchical, autonomous process. Given a research target, `FinSight` first initiates a broad **Data Collection** phase to aggregate a comprehensive repository of structured and unstructured market intelligence. Subsequently, the framework executes a **Two-Stage Writing** process: it first distills raw data into a set of multi-dimensional insights and visualizations (termed the *Chain-of-Analysis*), and then synthesizes these segments into a coherent, professional-grade final report. Crucially, this workflow is not a rigid linear pipeline. Agents possess the autonomy to recursively invoke one another to resolve information gaps dynamically, effectively transitioning the system from handling raw, heterogeneous data to producing a structured multimodal report.

To support this complex interaction, our framework is built upon the **Code Agent with Variable Memory (CAVM)** architecture (detailed in Sec 2.3). This architecture unifies heterogeneous resources(e.g. data, tools, agents) into a programmable variable space, enabling agents to manipulate them as executable objects through code. Within this unified space, we orchestrate three specialized agent classes, as shown in Figure 1: (1) **Data Collection Agents** serve as the interface to the external world, transforming web knowledge and database inputs into executable Python objects (e.g., DataFrames) within the variable memory; (2) the **Data Analysis Agent** acts as the core reasoning engine, transforming raw data into analytical insights to produce multimodal Chain-of-Analysis (CoA) segments; and (3) the **Report Generation Agent** synthesizes these insights into the final report, ensuring overall narrative coherence and structural consistency.

2.3 Code Agent with Variable Memory

Motivation: From Reading Context to Manipulating Variables Traditional agents typically rely on unstructured text or vector embeddings as memory. While sufficient for general tasks, this paradigm struggles in professional financial scenarios, which require precise calculations and handling of massive heterogeneous data. To address this, we propose **Code Agent with Variable Memory (CAVM)**, a novel architecture that redefines agent memory as a *Programmable State Representation*. The core philosophy is to shift the agent's interaction mode from *reading context* to *manipulating variables*. This design empowers agents to maintain data as executable objects (e.g., DataFrames) rather than static text, enabling rigorous mathematical operations and significantly reducing hallucinations in numerical reasoning.

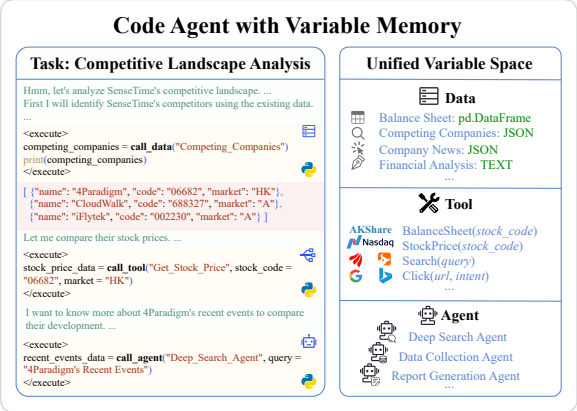

Figure 2: The design philosophy of CAVM.

Unified Variable Space We abstract the multi-agent collaboration environment into a unified variable space $\mathcal{V}$, encompassing three distinct types as shown in Figure 2: (1) **Data ($\mathcal{V}_{data}$):** Stores both structured (e.g., 'pandas.DataFrame' for financial tables) and unstructured data as executable Python objects; (2) **Tools ($\mathcal{V}_{tool}$):** Functional interfaces for external interaction; (3) **Agents ($\mathcal{V}_{agent}$):** encapsulated agent instances that can be invoked recursively.

$$\mathcal{V} = \mathcal{V}_{data} \cup \mathcal{V}_{tool} \cup \mathcal{V}_{agent}.$$

This unified scope allows heterogeneous elements to be accessed via a standard code interface. For instance, an agent can perform statistical analysis on $\mathcal{V}_{data}$ or invoke another expert agent from $\mathcal{V}_{agent}$ within the same code block, supporting hierarchical reasoning that static context windows cannot achieve.

Foundation Agent with Code Action Built upon this variable space, the agent operates in an iterative loop of reasoning and code execution. Unlike purely generative agents, our agent actively decides which variables to retrieve or modify via code, ensuring **contextual conciseness**. Formally, at step t, the agent generates a reasoning trace $\mathcal{R}_t$ and a code action $\mathcal{C}_t$:

$$\begin{aligned} &P_\theta(\mathcal{R}_t, \mathcal{C}_t \mid q, \mathcal{V}_{t-1}, \mathcal{H}_{t-1}) \\ &= \underbrace{P_\theta(\mathcal{R}_t \mid \Phi(\mathcal{V}_{t-1}), \cdot)}_{\text{Reasoning}} \cdot \underbrace{P_\theta(C_t \mid \mathcal{R}_t, \Phi(\mathcal{V}_{t-1}), \cdot)}_{\text{Code Action}}, \end{aligned}$$

Table 1: Comparison of agent roles in terms of their tool sets (including agent-as-a-tool) and data access.

Agent	Tool & Agent Set	Data
Deep Search	Search, Browse, PDF Reader	–
Data Collection	Fin. APIs, Econ. APIs, Deep Search Agent	–
Data Analysis	Deep Search Agent	Collected Data
Report Gen.	Deep Search Agent	Data, Analysis

Algorithm 1 CAVM Agent Execution Loop

Require: Task input q, Variable space $\mathcal{V}_0$, Max iterations T
Ensure: Final result and updated variable space

$\mathcal{H}_0 \leftarrow \text{PreparePrompt}(q, \Phi(\mathcal{V}_0))$
for $t = 1$ **to** T **do**
  $\mathcal{R}_t, \mathcal{C}_t \leftarrow \text{LLM}(\mathcal{H}_{t-1})$ *// Reason & generate code*
  if $\mathcal{C}_t$ is `FINAL_ANSWER` **then**
    return $\mathcal{R}_t, \mathcal{V}_{t-1}$
  end if
  $\mathcal{V}_t, \text{out}_t \leftarrow \text{Execute}(\mathcal{C}_t, \mathcal{V}_{t-1})$ *// Run code*
  $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \oplus (\mathcal{R}_t, \mathcal{C}_t, \text{out}_t)$
end for
return $\mathcal{H}_T, \mathcal{V}_T$

where Φ is a formatting function that summarizes the metadata of variables in $\mathcal{V}_{t-1}$. The code $\mathcal{C}_t$ is then executed by a Python interpreter to update the variable space:

$$\mathcal{V}_t, \text{output}_t = \text{Execute}(\mathcal{C}_t, \mathcal{V}_{t-1}), \quad (1)$$

$$\mathcal{H}_t = \mathcal{H}_{t-1} \oplus \text{output}_t. \quad (2)$$

This mechanism allows the agent to maintain a "working memory" of precise data states throughout long-horizon tasks.

CAVM Agent Loop Algorithm 1 presents the simplified execution loop of each CAVM-based agent. All agents share this unified loop; the key difference lies in their assigned tool sets, callable sub-agents, and available data, as summarized in Table 1.

2.4 Two-Stage Writing with Generative Retrieval

Motivation A complete report encompasses analyses from multiple perspectives, which can be regarded as an integration of several Chains-of-Analysis. To generate long-form financial research reports with both textual depth and multimodal coherence, we design a **two-stage writing framework** augmented with generative retrieval. It decomposes the report writing process into (1) Chain-of-Analysis Generation and (2) Structured Writing with Generative Retrieval.

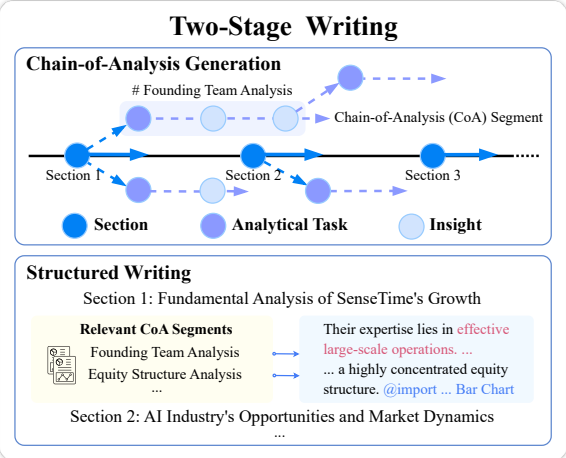

Figure 3: Chain-of-Analysis Illustration.

Stage 1: Chain-of-Analysis Generation Given the research question q, the Data Analysis Agent first generates a set of analytical perspectives $\mathcal{P} = \{p_1, p_2, ..., p_K\}$. The agent then performs parallel data analysis for each p_i, producing corresponding Chain-of-Analysis (CoA) that capture insights from distinct viewpoints.

Each CoA is generated based on the interaction history $\mathcal{H}_i$, accumulated during the data analysis process. To ensure coherence between textual content and referenced elements (e.g. figure, reference), this process is augmented with a **generative retrieval mechanism** that jointly produces textual contents along with element identifiers.

Specifically, for *citations*, the agent inserts natural language placeholders (e.g., `[Source: SenseTime Income Statement]`) directly into the text stream during generation, leveraging its memory of previously browsed data. These descriptions are then matched against data metadata via BM25 retrieval, with a similarity threshold to filter out unreliable matches. For *figures*, the agent generates import tokens (e.g., `@import "revenue_trend_chart"`), describing the required visual content. Since charts are created during the Vision-Enhanced stage, we pre-generate captions via a VLM and use embedding retrieval to match the agent's description with the most relevant chart. This unified natural language interface enables the model to plan the report structure and its multimodal references within a single, uninterrupted autoregressive generation, ensuring a smooth narrative flow. The process can be formal-

ized as:

$$P(\mathcal{A} \mid q, \mathcal{V}) = P(\mathcal{P} \mid q, \mathcal{V}) \cdot \prod_{i=1}^{|\mathcal{P}|} P(a_i \mid p_i, \mathcal{V}).$$

Stage 2: Structured Writing Building on CoAs, a Report Generation Agent first constructs a report outline $\mathcal{O} = \{o_1, o_2, ..., o_n\}$, and then writes each section sequentially. For each section s_i, the agent dynamically retrieves the most relevant data and CoA segments from the unified variable memory $\mathcal{V}$, formalized as:

$$P(R \mid \mathcal{A}, \mathcal{V}, q) = P(\mathcal{O} \mid \mathcal{A}, q) \cdot \prod_{i=1}^{n} P(A_{\text{sel}}^{(i)}, \mathcal{V}_{\text{sel}}^{(i)} \mid \mathcal{A}, \mathcal{V}, \cdot) \cdot P(s_i \mid s_{<i}, A_{\text{sel}}^{(i)}, \cdot).$$

To prevent hallucination of non-existent references and figures, agent is instructed to follow the identifiers established in $\mathcal{A}$. To ensure reference accuracy, the agent strictly follows the identifiers established during the stage 1.

2.5 Iterative Vision-Enhanced Mechanism for Visualization

Motivation Generating high-quality visualizations is a persistent challenge in automated report generation, particularly in data-intensive domains like finance that require nuanced analysis and presentation. Existing methods often rely on single-pass code execution or employ Vision-Language Models (VLMs) without incorporating visual feedback, which frequently leads to suboptimal outcomes. Drawing inspiration from Chain-of-Thought (Wei et al., 2022) and Actor-Critic (Schulman et al., 2017), we propose a framework where an agent learns to progressively improve visualizations. This is achieved by iteratively plotting a chart and refining it based on critical feedback, ensuring both stable generation and continuous quality enhancement.

Iterative Vision-Enhanced Mechanism Specifically, the final output of the Data Analysis Agent includes the target chart specifications along with the corresponding descriptions and data. As shown in Figure 4, the agent generates an initial visualization through executable plotting code, which is then evaluated by a VLM to give potential issues of visual cues (e.g., missing labels, inappropriate color schemes). These feedbacks are sent to the system, directing the iterative code generation until the output reaches professional quality.

$$P(\mathcal{C}_{vis} \mid \mathcal{V}) = \prod_{t=1}^{M} P_\theta(\mathcal{C}_t^{vis} \mid \mathcal{C}_{t-1}^{vis}, \mathcal{F}_{t-1}, \mathcal{V}),$$
$$\mathcal{F}_{t-1} = \text{VLM}(\text{Execute}(\mathcal{C}_{t-1}^{vis})),$$

where M is the maximum number of iterations. The iteration continues until convergence or a predefined quality threshold is satisfied.

3 Experiments

3.1 Benchmark Construction

Financial research report generation remains an under-explored problem lacking appropriate evaluation benchmarks and metrics. To address this gap, we construct a comprehensive benchmark specifically designed for multimodal financial report generation. This benchmark consists of a curated dataset of research targets, human written reports as ground truth, and a rigorous multidimensional evaluation protocol.

Dataset Composition. To simulate real-world investment research scenarios, our dataset encompasses research targets at both company and industry levels. For company-level analysis, we curated a diverse list of companies covering different markets, industry sectors, and market capitalizations. For industry-level analysis, we selected high-attention industries as research targets. To establish a rigorous ground truth, we collected in-depth analysis reports authored by professional brokerage institutions to serve as Golden References. To ensure a high quality bar, we applied stringent filtering criteria, selecting only reports exceeding 20 pages in length and containing more than 20 charts and visualizations. In total, we collected 20 samples: 10 company-level and 10 industry-level targets.

Evaluation Metrics. We design 9 automated evaluation metrics across three critical dimensions, each ranging from 0 to 10 points. Detailed description of each metric can be found in Appendix F.

(1) Factual Accuracy: Measures the reliability and correctness of generated content through Core Conclusion Consistency (alignment with reference conclusions), Textual Faithfulness (proper citation support), and Text-Image Coherence (consistency between textual and visual elements).

(2) Information Effectiveness: Evaluates the analytical value delivered to investors via Information Richness (distinct information points), Cover-

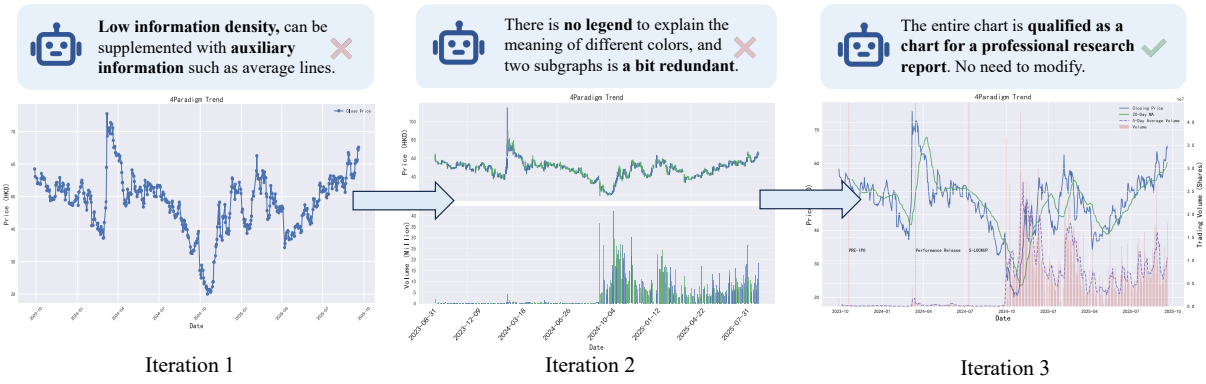

Figure 4: An example of our Iterative Vision-Enhanced Mechanism of Visualization. The chart is generated by matplotlib and seaborn package in Python.

age (proportion of key reference information captured), and Analytical Insight (critical analysis and forward-looking recommendations).

(3) Presentation Quality: Assesses professional standards through Structural Logic (organizational coherence), Language Professionalism (adherence to financial terminology), and Chart Expressiveness (effective visualization utilization and aesthetic quality).

Details of selected targets and evaluation metrics can be found in Appendix F.

3.2 Baselines

We compare `FinSight` against two categories of baselines: (1) **LLMs with Search Tools:** Leading large language models directly combined with search tools for report generation, including OpenAI GPT-5 (OpenAI, 2025b), DeepSeek-R1 (DeepSeek-AI et al., 2025), and Claude-4.1-Sonnet (anthropic, 2025); (2) **Deep Research Agents:** State-of-the-art commercial deep research products, including Gemini-2.5-Pro Deep Research (Gemini, 2025), Grok Deep Search (Grok, 2025), OpenAI Deep Research (OpenAI, 2025a), and Perplexity Deep Research[1]. Details of baseline implementations are shown in Appendix D.

3.3 Implementation Details

Our framework utilizes DeepSeek-V3 and DeepSeek-R1 as backbone models. For search, we employ the Google Search API to fetch the top-10 results. To ensure robust evaluation, we use Gemini-2.5-Pro as the primary judge model, with GPT-5 serving as an auxiliary evaluator to mitigate single-model bias. In main experiments, we report mean scores with 95% confidence intervals across three independent runs to verify statistical significance. To complement our automated metrics, we conducted a rigorous human evaluation with 6 experts, using the same evaluation metrics. Each annotator reviewed a random subset of 10 research targets, referencing professional brokerage reports as the golden standard. We report the inter-rater reliability using Krippendorff's α and human-LLM judgement correlation using Pearson's r. For additional details, please refer to Appendix C.5.

3.4 Main Results

Table 2 compares the performance of `FinSight` against all baselines. **Overall, `FinSight` achieves the highest average score (7.93)**, significantly outperforming closed-source commercial agents including Gemini DR (5.73) and OpenAI DR (6.44). This validates the effectiveness of our multi-agent framework in generating professional-grade reports.

Regarding *Factuality*, `FinSight` secures top scores in **citation faithfulness and text-image consistency**, demonstrating the efficacy of the identifier mechanism within our Chain-of-Analysis. Notably, while our consistency score (6.84) is marginally lower than OpenAI DR (6.87), case studies suggest this stems from our preference for **comprehensive data acquisition over simplified conclusions**, which occasionally introduces complex, data-driven variances.

Crucially, `FinSight` demonstrates a clear advantage in **Analytical Quality**, securing the highest scores across richness, coverage, and insightfulness. This superiority also extends to **Presentation Quality**, where our system maintains a comprehensive lead in logic, language, and particularly **visualization**. Notably, the visualization score (8.57)

[1] `https://www.perplexity.ai`

Table 2: Overall evaluation results on financial report generation benchmark (averaged over three runs). **Bold** denotes the highest score in each column, Underlined denotes the second highest. Full results are shown in Appendix C.1.

Model	Factual			Analytical			Presentation			Avg.
	Cons.	Faith.	T-I.	Rich.	Cover.	Ins.	Logic	Lang.	Vis.	
LLM with Search Tools										
GPT-5 w/ Search	5.95	6.35	4.77	5.43	4.52	5.09	6.53	5.87	3.90	5.38
Claude-4.1-Sonnet w/ Search	5.78	5.92	3.55	5.58	5.25	5.01	6.34	6.07	2.59	5.12
DeepSeek-R1 w/ Search	6.26	5.92	4.08	6.68	6.33	6.62	7.03	6.79	3.35	5.90
Deep Research Agent										
Grok Deep Search	4.71	5.72	4.21	4.90	4.03	4.35	5.87	5.61	3.76	4.79
Perplexity Deep Research	5.02	5.74	4.03	3.88	3.40	3.65	5.47	4.92	3.42	4.39
Gemini-2.5-Pro Deep Research	5.92	6.66	4.32	6.19	6.03	5.74	6.77	6.70	3.23	5.73
OpenAI Deep Research	**6.87**	6.78	4.58	6.79	6.83	7.33	7.56	7.58	3.66	6.44
FinSight (ours)	6.84	**7.59**	**7.84**	**8.49**	**8.44**	**7.78**	**7.82**	**7.98**	**8.57**	**7.93**

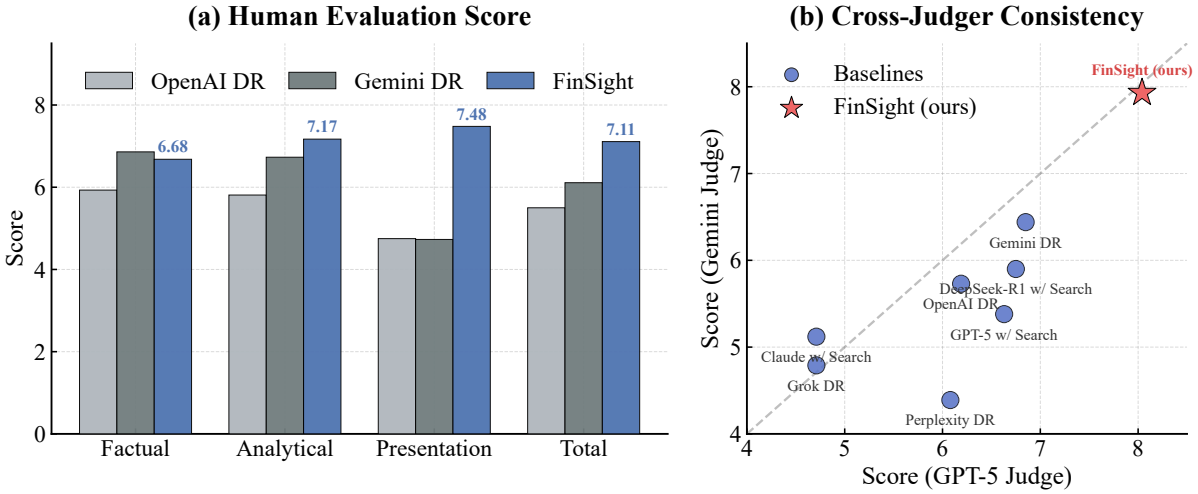

Figure 5: **Results across different evaluators. (a)** Comparison of human scores (0-10 scale). **(b)** Evaluation results using GPT-5 and Gemini-2.5-Pro as judges.

far surpasses all baselines, underscoring the system's advanced multimodal capabilities in generating professional-grade charts.

3.5 Ablation Studies

We conduct ablation studies (Table 3) to evaluate the contribution of our key components. Key findings are as follows: (1) Removing iterative VLM feedback causes a significant decline in Presentation (8.12 → 6.63) and Analytical Quality (8.23 → 7.23). This confirms that high-quality visuals are prerequisite for the model to generate insightful, chart-based analysis. (2) Merging analysis and writing into a single process leads to a sharp drop in Analytical Quality (8.23 → 5.83) and Factual Accuracy (7.42 → 6.20), validating the superiority of our analyze-then-write strategy over single-pass generation. (3) Eliminating dynamic search degrades performance across all dimensions (e.g., Factual Accuracy 7.42 → 5.97), highlighting the necessity of acquiring supplementary knowledge during the analysis and drafting phases.

4 Analysis

4.1 Human Evaluation

We recruited six financial experts to benchmark FinSight against two leading commercial baselines, using Golden Reports and strict rubrics for objectivity (details in Appendix G.2). As shown in Figure 5(a), FinSight achieves the highest total score (**7.11**), significantly outperforming Gemini DR (6.11) and OpenAI DR (5.50). While Gemini DR retains a slight edge in *Factual* accuracy (6.86 vs. 6.68), which is likely due to proprietary data access, FinSight dominates in *Analytical Depth* and

Table 3: Ablation studies of our key design.

Method	Fact.	Ana.	Pres.
FinSight	7.42	8.23	8.12
w/o Two-Stage Writing	6.20	5.83	6.20
w/o Iterative Feedback.	7.10	7.23	6.63
w/o Dynamic Search.	5.97	5.80	7.03

Table 4: Statistics of our generation process at both the CoA level and the final report level.

Chain of Analysis Level			
# Tokens	2,761	# Images	5.3
Final Report Level			
# Fin. API Calls	18.3	# Tokens	62,586
# Search Queries	983.2	# Images	51.2
# Browse Pages	469.8	# CoA Segments	17.6

Presentation Quality. Notably, FinSight scores **7.48** in presentation (baselines < 5.0), demonstrating the decisive value of our multimodal chart generation.

4.2 Cross-Model Verification

To mitigate potential self-preference bias in our primary judge (Gemini DR), we employed GPT-5 as an independent evaluator using identical protocols. Figure 5(b) confirms that FinSight consistently ranks first regardless of the evaluator. We observe high ranking consistency between judges (Kendall's $\tau = 0.764$, $p = 0.008$). Even under GPT-5 evaluation, Gemini DR remains second, confirming that FinSight's lead stems from objective improvements in analytical and visual depth rather than evaluator-specific bias.

4.3 Reliability and Factuality Analysis

Key Fact Accuracy. Directly measuring the factuality of long-form reports is challenging. We introduced a **Golden Facts Evaluation** by extracting 13 core financial indicators (e.g., Gross Margin, ROE) from professional reports as Ground Truth. We manually verified the accuracy of these data points. As shown in Figure 6(a), FinSight achieves an accuracy of **54.6%**, surpassing Gemini DR (38.5%) and OpenAI DR (30.0%) by a large margin. This demonstrates our method's superiority in uncovering deep, quantitative details that general-purpose agents often miss.

Citation Quality Analysis. We further evaluated the quality of the references cited in our report. (1) **Citation Faithfulness:** We manually verified the top-50 citations per report to check if the source explicitly supported the text. FinSight achieves a superior accuracy of **72.9%** (342/469 verified), outperforming Gemini DR's **69.8%**. This high faithfulness is attributed to our generative retrieval mechanism, which identifies references during the drafting process rather than via post-hoc appending. (2) **Source Authority:** We classified citations into High, Medium, and Low authority based on human-written rules. Figure 6(b) reveals that FinSight utilizes High Authority sources at a rate (36.5%) comparable to Gemini DR. While our reliance on open web search results in a slightly higher portion of Low Authority sources, the high faithfulness score ensures that the information extracted remains valid.

4.4 Generation Process Statistics

Table 4 provides a detailed summary of the generation process. At the Chain-of-Analysis level, each CoA is a self-contained multimodal block, averaging 2,761 tokens and 5.3 images. At the final report level, a typical report synthesizes approximately 17.6 CoA segments, resulting in 62,586 tokens and 51.2 images. The data collection phase involves an average of 983.2 search queries and 469.8 browsed web pages, demonstrating the depth of information gathering enabled by our framework.

5 Related Work

Deep Research Systems Deep research systems have evolved from simple retrieval to agentic knowledge synthesis, utilizing ReAct loops(Yao et al., 2022) or multi-agent collaboration to automate information gathering (OpenAI, 2025a; Li et al., 2025c; Hu et al., 2025; Tang et al., 2025; Li et al., 2025b). However, existing frameworks exhibit significant limitations in multimodal processing (Yang et al., 2025) and domain-specific applications (Jimeno-Yepes et al., 2024; Tian et al., 2025; Jin et al., 2024b). Due to the text-centric design of report generation workflows and the base models' lack of native image generation capabilities (Ren et al., 2021; Chen et al., 2024; Dong et al., 2025c), current systems produce reports deficient in visual elements such as charts and diagrams. Furthermore, these systems demonstrate inadequate adaptation to financial domains, particularly in their inability to support for professional-grade chart generation, limited real-time market data integration, creating

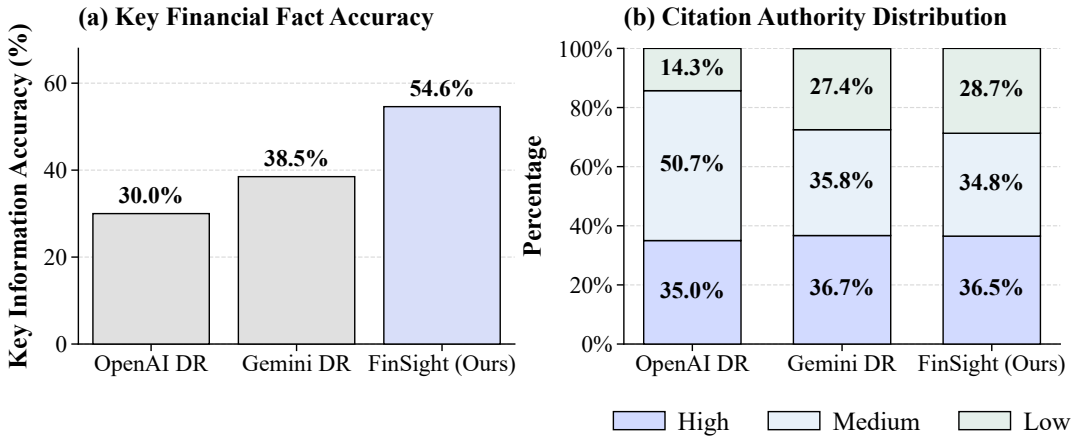

Figure 6: Factuality Evaluation. (a) **Key Financial Fact Recall**: The percentage of ground truth facts from golden reports covered by each model. (b) **Citation Authority**: The distribution of citations classified by source reliability.

substantial gaps between system outputs and professional requirements.

LLM Agents in Financial Domain Existing financial agents largely focus on specific predictive tasks (Zhang et al., 2025; Xiao et al., 2025; Dong et al., 2025b) or employ simplified, **single-pass workflows** for report generation (Wu et al., 2025; Yang et al., 2024; Dong et al., 2025d). These approaches often suffer from superficial analysis, lacking the iterative reasoning required for deep insight (Dong et al., 2024). While some platforms offer data tools (Zhang et al., 2025; Jin et al., 2024a), they fail to provide a unified framework that combines **deep analytical reasoning** with **multimodal integration**. Consequently, current systems cannot meet the high standards of professional financial research in terms of data breadth, analytical depth, and presentation quality.

Crucially, these systems differ from `FinSight` in two key aspects. First, regarding **comprehensiveness and depth**: existing methods integrate limited, fixed data sources and lack the "deep search" capability required to mine diverse data, resulting in short, rigid analytical texts (∼300 tokens) focused on a few fixed dimensions, far below professional financial research standards. Second, regarding **multimodal capabilities**: current methods often lack visual outputs entirely or are restricted to a few predefined, low-quality charts. In contrast, `FinSight` targets the seamless generation of high-quality, interleaved text and professional visualizations through our Iterative Vision-Enhanced Mechanism.

6 Conclusion

In this paper, we present `FinSight`, a multi-agent framework designed for multimodal deep research. This framework is capable of generating comprehensive, long-form financial reports with interleaved text and images. By integrating the Code Agent with Variable Memory and an Iterative Vision-Enhanced Mechanism, `FinSight` achieves dynamic analysis and professional visualization. Empirical evaluations demonstrate that our system significantly outperforms state-of-the-art commercial agents in factual accuracy and analytical depth, successfully bridging the gap between raw market data and actionable investment insights. While benchmarked on financial tasks, our code-centric approach provides a promising paradigm for future general-purpose automated research. However, deploying such systems in the real world necessitates vigilance regarding potential risks. Without proper constraints, such a system could be misused to mass-produce deceptive reports for market manipulation. Furthermore, over-reliance on automated research without rigorous human oversight could amplify systemic biases. Future research must advance agent capabilities in tandem with stricter auditing mechanisms and safety guardrails to ensure the responsible application of AI in finance. A thorough discussion of potential risks is provided in Appendix B.

Limitations

While `FinSight` achieves promising results, we identify four limitations hindering immediate real-world adoption. (1) **Efficiency:** Our multi-agent

framework relies on large-scale LLMs for tasks like data analysis and plotting, resulting in high computational costs. Generating a single report consumes approximately 5 million tokens and takes about 20 minutes, which may limit its application in high-velocity environments. Future works could replace these with specialized, smaller models to improve inference efficiency. (2) **Factuality:** Despite using generative retrieval to produce text-image interleaved content, long-form generation is still prone to hallucinations, a common hurdle in deep research. Implementing a strict verification mechanism to ensure claim traceability is a necessary next step. Furthermore, the system currently processes a significant amount of noise and information from low-authority sources. Future iterations may require "tool middleware" to verify and denoise source data before processing. (3) **Generalizability:** Currently tailored for financial reports, the system lacks a flexible interface for other domains. Future work will explore human-in-the-loop designs to allow users to customize workflows for diverse report types. (4) **Data Latency:** `FinSight`'s workflow relies on data snapshots captured during the initial analysis phase. Because generating long-form reports is time-intensive, the system currently lacks a mechanism to dynamically capture and integrate high-frequency market shifts mid-writing. This may lead to slight data latency in extremely volatile market conditions. Consequently, `FinSight` is currently best suited for quarterly or weekly corporate research rather than real-time day trading.

Acknowledgements

This work was supported by National Natural Science Foundation of China No. 62272467. The work was partially done at the Beijing Key Laboratory of Research on Large Models and Intelligent Governance.

References

anthropic. 2025. Meet claude. https://www.anthropic.com/claude.

Nicholas Camara. 2025. open-deep-research. https://github.com/nickscamara/open-deep-research.

Yuemin Chen, Feifan Wu, Jingwei Wang, Hao Qian, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, and Meng Wang. 2024. Knowledge-augmented financial market analysis and report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1207–1217, Miami, Florida, US. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Yutao Zhu, and Zhicheng Dou. 2025a. Agentic entropy-balanced policy optimization. *CoRR*, abs/2510.14545.

Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. 2025b. Tool-Star: Empowering LLM-brained multi-tool reasoner via reinforcement learning. *CoRR*, abs/2505.16410.

Guanting Dong, Jiajie Jin, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2025c. RAG-Critic: Leveraging automated critic-guided agentic workflow for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025*.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2025d. Agentic reinforced policy optimization. *CoRR*, abs/2507.19849.

Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Zhicheng Dou, and Ji-Rong Wen. 2024. Understand what LLM needs: Dual preference alignment for retrieval-augmented generation. *CoRR*, abs/2406.18676.

Gemini. 2025. Gemini deep research. https://gemini.google/overview/deep-research.

Grok. 2025. Grok 3 beta — the age of reasoning agents. https://x.ai/news/grok-3.

David Hasler and Sabine Süsstrunk. 2003. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII, Santa Clara, CA, USA, January 20, 2003*, volume 5007 of *SPIE Proceedings*, pages 87–95. SPIE.

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, Zeyu Zhang, Yifeng Wang, Qianshuo Ye, Bernard Ghanem, Ping Luo, and Guohao Li. 2025. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation. *Preprint*, arXiv:2505.23885.

Antonio Jimeno-Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation. *CoRR*, abs/2402.05131.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2025a. Hierarchical document refinement for long-context retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025*.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Zhao Yang, Hongjin Qian, and Zhicheng Dou. 2025b. Decoupled planning and execution: A hierarchical reasoning framework for deep search. *CoRR*, abs/2507.02652.

Jiajie Jin, Yanzhao Zhang, Mingxin Li, Dingkun Long, Pengjun Xie, Yutao Zhu, and Zhicheng Dou. 2026. LaSER: Internalizing explicit reasoning into latent space for dense retrieval. *CoRR*, abs/2603.01425.

Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. 2024a. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *CoRR*, abs/2405.13576.

Jiajie Jin, Yutao Zhu, Yujia Zhou, and Zhicheng Dou. 2024b. BIDER: bridging knowledge inconsistency for efficient retrieval-augmented llms via key supporting evidence. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 750–761. Association for Computational Linguistics.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models. *CoRR*, abs/2501.05366.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Guanting Dong, Jiajie Jin, Yinuo Wang, Hao Wang, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2025b. DeepAgent: A general reasoning agent with scalable toolsets. *CoRR*, abs/2510.21618.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Shijian Wang, Guanting Dong, Jiajie Jin, Hao Wang, Yinuo Wang, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2026. OmniGAIA: Towards native omni-modal AI agents. *CoRR*, abs/2602.22897.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025c. Webthinker: Empowering large reasoning models with deep research capability. *CoRR*, abs/2504.21776.

OpenAI. 2025a. Introducing deep research. https://openai.com/index/introducing-deep-research.

OpenAI. 2025b. Openai gpt-5. https://openai.com/gpt-5/.

Yunpeng Ren, Wenxin Hu, Ziao Wang, Xiaofeng Zhang, Yiyuan Wang, and Xuan Wang. 2021. A hybrid deep generative neural model for financial report generation. *Know.-Based Syst.*, 227(C).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Jiabin Tang, Tianyu Fan, and Chao Huang. 2025. AutoAgent: A Fully-Automated and Zero-Code Framework for LLM Agents. *Preprint*, arXiv:202502.05957.

Yong-En Tian, Yu-Chien Tang, Kuang-Da Wang, An-Zi Yen, and Wen-Chih Peng. 2025. Template-based financial report generation in agentic and decomposed information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 2706–2710. ACM.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10014–10037. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Yingqian Wu, Qiushi Wang, Zefei Long, Rong Ye, Zhongtian Lu, Xianyin Zhang, Bingxuan Li, Wei Chen, Liwen Zhang, and Zhongyu Wei. 2025. Finteam: A multi-agent collaborative intelligence system for comprehensive financial scenarios. *arXiv preprint arXiv:2507.10448*.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2025. Tradingagents: Multi-agents llm financial trading framework. *Preprint*, arXiv:2412.20138.

Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, et al. 2024. Finrobot: An open-source ai agent platform for financial applications using large language models. *arXiv preprint arXiv:2405.14767*.

Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework. *Preprint*, arXiv:2506.02454.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Wentao Zhang, Yilei Zhao, Chuqiao Zong, Xinrun Wang, and Bo An. 2025. Finworld: An all-in-one open-source platform for end-to-end financial ai research and deployment. *arXiv preprint arXiv:2508.02292*.

Appendix

A Statement on the Use of LLMs

During the preparation of this manuscript, we use LLMs as a general-purpose assistance tool. The primary role of the LLM is to aid in improving the clarity and readability of the text, as well as to accelerate the implementation of our research ideas. Specific applications include: (1) Language and Grammar Correction: Polishing sentence structure, correcting grammatical erros, and refining word choices to enhance the overall quality of the writing. (2) Paraphrasing and Style Refinement: Rephrasing sentences and paragraphs to ensure consistency in tone and style throughout the paper. (3) Code Implementation Assistance: Generating code snippets and providing debugging support to help implement the proposed algorithms and experimental setups.

It should be noted that all core research concepts, experimental design, data analysis, and conclusions are developed exclusively by the human authors. Any content or suggestions generated by the LLM, including code, are critically checked, and substantially edited by the authors to ensure accuracy. The authors take full responsibility for the final content of this paper.

B Potential Risks

While FinSight demonstrates significant advancements in automated financial deep research, the deployment of such a system in real-world scenarios introduces several potential risks that must be carefully managed.

Over-Reliance and Financial Liability. Although FinSight outperforms existing baselines in factual accuracy (achieving 54.6% on key financial facts), it does not reach 100% precision. The professional presentation of the reports—characterized by high-quality visualizations and authoritative tone may inadvertently induce *automation bias*, where users trust the system's output without sufficient verification. In the financial domain, where decisions involve substantial capital, minor hallucinations in numerical data or misinterpretations of market sentiment can lead to significant economic loss. Therefore, FinSight should be positioned as an auxiliary tool for human analysts rather than a fully autonomous financial advisor.

Security Risks in Code Execution. The core of our framework, the Code Agent with Variable Memory (CAVM), relies on the autonomous generation and execution of Python code to process data and render charts. While this enables flexible data manipulation, it introduces security vulnerabilities common to code-interpreting agents. Without a strictly sandboxed environment, there is a risk that the agent could generate or execute malicious code, particularly if prompted by adversarial inputs or if it ingests compromised data from external web sources. Future iterations must implement rigorous sandboxing and code-vetting protocols.

Data Provenance and Copyright. FinSight aggregates information from the open web to generate commercial-grade reports. This raises concerns regarding data provenance and intellectual property. Analyzing and restructuring proprietary financial data or news content without explicit licensing may violate copyright regulations or terms of service of

specific data providers. Furthermore, the system's reliance on web search means the quality of the report is bound by the quality of available public data, which may contain biases or outdated information.

Potential for Market Manipulation. The efficiency of FinSight in generating high-quality, persuasive financial narratives could be misused by malicious actors. The system could potentially be engineered to rapidly produce large volumes of biased reports intended to manipulate market sentiment (e.g., "pump and dump" schemes) or damage the reputation of targeted companies. Safeguards, such as watermarking generated content and restricting API access, are necessary to mitigate the risk of generating synthetic financial misinformation at scale.

C Further Analysis of `FinSight`

C.1 Stability Analysis

To measure the sensitivity of our evaluation to stochastic variations, we repeated the evaluation process three times using `Gemini-2.5-Pro` as the judge model. For each run, we use the same evaluation prompt and rubric but with different random seeds. The 95% Confidence Intervals (CI) are calculated as $\mu \pm 1.96 \times \sigma/\sqrt{n}$ where $n = 3$.

As shown in Table 5, the 95% confidence intervals are consistently within ± 1.0 point across all metrics, demonstrating the stability of our evaluation protocol. We attribute this to the robust evaluation design that anchors scoring against a provided Golden Report and utilizes a detailed, listwise grading rubric.

C.2 Dimension-Specific Variance Analysis

We further analyzed the standard deviation across specific evaluation dimensions to understand which aspects of report quality are more reliably assessed by LLM judges.

Table 6 reveals important patterns: (1) **Structural and linguistic metrics** (e.g., *Structural Logic*, *Analytical Insight*) exhibit high stability with Std < 1.0, indicating that LLM judges reliably assess writing quality. (2) **Factual metrics** (e.g., *Textual Faithfulness*) show moderate variance, reflecting inherent difficulty in verifying factual claims. (3) **Visual metrics** (e.g., *Text-Image Coherence*, *Chart Expressiveness*) display slightly higher variance, as multimodal assessment involves more subjective judgment. Despite these variations, the overall ranking of methods remains consistent across runs.

C.3 Cross-Model Verification

To investigate whether `Gemini-2.5-Pro` exhibited "self-preference bias" (favoring its own outputs), we employed `GPT-5` as an independent evaluator using the identical prompt and rubric.

As shown in Table 7, the ranking order remains highly consistent between the two judges (Kendall's $\tau = 0.764$, $p = 0.008$). Notably, even when evaluated by GPT-5, Gemini-2.5-Pro Deep Research retains the second-place position, and `FinSight` consistently achieves the top rank. This confirms that `FinSight`'s superior performance is attributable to objective report quality rather than evaluator bias.

C.4 Quantitative Visual Analysis

To rigorously quantify the **presentation quality** and validate the effectiveness of our **Iterative Vision-Enhanced Mechanism**, we implement three reference-free Image Quality Assessment (IQA) metrics (Hasler and Süsstrunk, 2003):

- **Colorfulness**: Measures the chromatic distinction between visual elements, computed as $\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3 \times \sqrt{\mu_{rg}^2 + \mu_{yb}^2}$, where $rg = R - G$ and $yb = 0.5(R + G) - B$.
- **RMS Contrast**: Measures luminance contrast using the root-mean-square of pixel intensities, correlating with the legibility of labels and grid lines.
- **Edge Density**: Measures information density versus visual clutter using Canny edge detection, computed as the ratio of edge pixels to total pixels.

We argue that pixel-wise metrics (e.g., MSE, SSIM) are ill-suited for chart evaluation, as different rendering engines produce large pixel discrepancies even when plotting identical data. The IQA metrics above provide a more meaningful assessment of visual quality. As shown in Table 8, our Iterative Vision-Enhanced Mechanism approximately doubles the scores across all three dimensions, objectively validating that the VLM critic loop significantly improves the aesthetic quality and information density of the generated charts.

C.5 Human Evaluation

As illustrated in Figure 7, our key observations are: **(1) Validation of Automated Metrics:** Figure 5(b)

Table 5: Overall evaluation results with 95% confidence intervals (subscript). **Bold** denotes the highest score, Underlined denotes the second highest.

Model	Factual			Analytical			Presentation			Avg.
	Cons.	Faith.	T-I.	Rich.	Cover.	Ins.	Logic	Lang.	Vis.	
LLM with Search Tools										
GPT-5 w/ Search	$5.95_{\pm 0.48}$	$6.35_{\pm 0.39}$	$4.77_{\pm 0.72}$	$5.43_{\pm 0.58}$	$4.52_{\pm 0.51}$	$5.09_{\pm 0.49}$	$6.53_{\pm 0.27}$	$5.87_{\pm 0.35}$	$3.90_{\pm 0.79}$	$5.38_{\pm 0.51}$
Claude-4.1-Sonnet w/Search	$5.78_{\pm 0.52}$	$5.92_{\pm 0.55}$	$3.55_{\pm 0.65}$	$5.58_{\pm 0.49}$	$5.25_{\pm 0.51}$	$5.01_{\pm 0.46}$	$6.34_{\pm 0.26}$	$6.07_{\pm 0.40}$	$2.59_{\pm 0.51}$	$5.12_{\pm 0.48}$
DeepSeek-R1 w/ Search	$6.26_{\pm 0.36}$	$5.92_{\pm 0.44}$	$4.08_{\pm 0.37}$	$6.68_{\pm 0.26}$	$6.33_{\pm 0.34}$	$6.62_{\pm 0.41}$	$7.03_{\pm 0.20}$	$6.79_{\pm 0.29}$	$3.35_{\pm 0.45}$	$5.90_{\pm 0.35}$
Deep Research Agent										
Grok Deep Search	$4.71_{\pm 0.81}$	$5.72_{\pm 0.59}$	$4.21_{\pm 0.57}$	$4.90_{\pm 0.55}$	$4.03_{\pm 0.58}$	$4.35_{\pm 0.55}$	$5.87_{\pm 0.42}$	$5.61_{\pm 0.48}$	$3.76_{\pm 0.53}$	$4.79_{\pm 0.56}$
Perplexity Deep Research	$5.02_{\pm 0.61}$	$5.74_{\pm 0.63}$	$4.03_{\pm 0.97}$	$3.88_{\pm 0.60}$	$3.40_{\pm 0.53}$	$3.65_{\pm 0.59}$	$5.47_{\pm 0.44}$	$4.92_{\pm 0.47}$	$3.42_{\pm 0.97}$	$4.39_{\pm 0.65}$
Gemini-2.5-Pro Deep Research	$5.92_{\pm 0.61}$	$6.66_{\pm 0.56}$	$4.32_{\pm 0.91}$	$6.19_{\pm 0.65}$	$6.03_{\pm 0.61}$	$5.74_{\pm 0.62}$	$6.77_{\pm 0.40}$	$6.70_{\pm 0.43}$	$3.23_{\pm 0.95}$	$5.73_{\pm 0.64}$
OpenAI Deep Research	$\mathbf{6.87}_{\pm 0.52}$	$6.78_{\pm 0.34}$	$4.58_{\pm 0.51}$	$6.79_{\pm 0.35}$	$6.83_{\pm 0.30}$	$7.33_{\pm 0.40}$	$7.56_{\pm 0.25}$	$7.58_{\pm 0.31}$	$3.66_{\pm 0.50}$	$6.44_{\pm 0.39}$
`FinSight` (ours)	$6.84_{\pm 0.59}$	$\mathbf{7.59}_{\pm 0.52}$	$\mathbf{7.84}_{\pm 0.52}$	$\mathbf{8.49}_{\pm 0.47}$	$\mathbf{8.44}_{\pm 0.50}$	$\mathbf{7.78}_{\pm 0.48}$	$\mathbf{7.82}_{\pm 0.38}$	$\mathbf{7.98}_{\pm 0.33}$	$\mathbf{8.57}_{\pm 0.57}$	$\mathbf{7.93}_{\pm 0.49}$

Table 6: Average standard deviation across all models for each evaluation dimension.

Dimension	Avg. Std.
Structural Logic (Logic)	0.675
Professional Language (Lang.)	0.848
Analytical Insight (Ins.)	0.932
Information Coverage (Cover.)	0.966
Core Conclusion Consistency (Cons.)	1.010
Information Richness (Rich.)	1.058
Textual Faithfulness (Faith.)	1.213
Chart Expressiveness (Vis.)	1.340
Text-Image Coherence (T-I.)	1.470

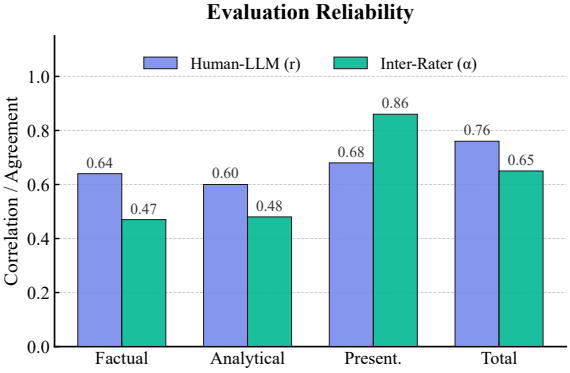

Figure 7: **Reliability analysis metrics.** High Human-LLM correlation (r) validates our automated judges, and robust Inter-Rater reliability (α) confirms the consensus among human experts.

details the reliability metrics. We observe a strong positive correlation between human and LLM scoring (Total Score $r = 0.76$), which **validates the reliability of the automated evaluation framework** used in our main experiments. **(2) Robust Inter-Rater Agreement:** The consistently high α values in Figure 5(b), particularly in the Presentation dimension ($\alpha = 0.86$), underscore that the

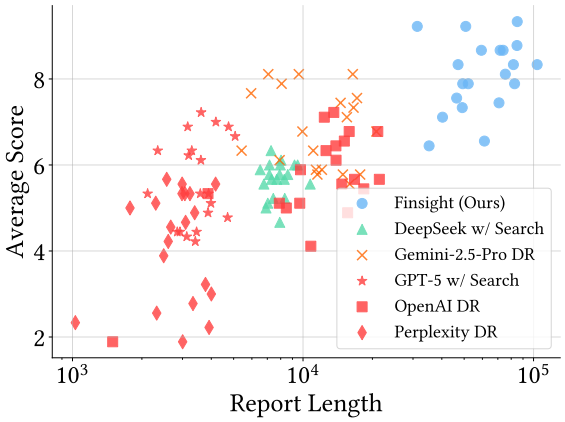

Figure 8: Correlation between report length and quality score across different methods.

visual and structural advantages of `FinSight` are objectively recognizable and consensus-based.

C.6 Report Length and Quality

To further investigate the characteristics of the generated reports, we analyze the relationship between report length and overall quality score, as illustrated in Figure 8. The plot shows that the outputs from our method are concentrated in the top-right quadrant, which indicates that our generated reports are not only comprehensive and of substantial length (typically over 20,000 words) but also of superior quality. We attribute this strong and consistent performance to our proposed two-stage writing framework. By first generating a concise Chain-of-Analysis, the model can then compose the final report based on richer, well-structured information, ensuring both analytical depth and coherence.

In contrast, baseline methods exhibit significant limitations. Simpler approaches like LLM with search tool, which often rely on single-pass gener-

Table 7: Cross-model evaluation comparison between GPT-5 and Gemini-2.5-Pro as judges.

Method	Score (GPT-5)	Score (Gemini)	Rank (GPT-5)	Rank (Gemini)
GPT-5 w/ Search	6.63	5.38	4	5
Claude-4.1-Sonnet w/ Search	4.71	5.12	7	6
DeepSeek-R1 w/ Search	6.75	5.90	3	3
Grok Deep Search	4.71	4.79	8	7
Perplexity Deep Research	6.08	4.39	6	8
OpenAI Deep Research	6.19	5.73	5	4
Gemini-2.5-Pro Deep Research	6.85	6.44	2	2
FinSight (ours)	**8.04**	**7.93**	**1**	**1**

Table 8: Quantitative visual quality comparison. Higher Colorfulness and Contrast indicate better aesthetic quality; Edge Density reflects information density.

Metric	FinSight (Full)	*w/o* Iter. Vision
Colorfulness	**32.35**	15.81
Contrast	**31.71**	15.47
Edge Density	**0.0056**	0.0027

ation, are typically constrained to shorter reports. Meanwhile, other deep research agents such as OpenAI DR and Perplexity DR display a wide scatter of data points across the plot, which signifies a critical lack of consistency. For these methods, a greater length does not reliably translate into higher quality, highlighting the effectiveness of our structured, two-stage approach.

Length-Controlled Evaluation. To address potential length bias in LLM-based evaluations, we conducted an additional experiment with strict length constraints. We applied a truncation strategy to limit FinSight's output to approximately 10,000 words, aligning it with baseline models.

As shown in Table 9, even with forced truncation (which naturally penalizes coherence and completeness), FinSight (10k) achieves an overall score of 6.93, surpassing Gemini-2.5-Pro (6.82) and significantly outperforming OpenAI Deep Research (6.11). This demonstrates that FinSight's performance gain derives from high-quality content synthesis rather than mere verbosity.

D Implementation Details of Baselines

We mainly compare our method with the following two types of baselines:

(1) LLMs with Search Tools.

- **OpenAI GPT-5 w/Search**: The latest OpenAI's GPT model with web search API for research question.
- **Claude-4.1-Sonnet w/Search** The latest Anthropic's reasoning LLM with web search API for research question.
- **DeepSeek-R1 w/ Search**: The DeepSeek's LLM integrated with web search API for research question.

(2) Deep Research Agents.

- **Grok Deep Search**: The xAI's Deep Search applications, powered by the latest Grok model.
- **Perplexity Deep Research**: A commercial AI research assistant integrating multi-step search and analysis, optimized for rapid information aggregation.
- **OpenAI Deep Research**: A multi-step web research agent built on ChatGPT that searches, analyzes, and synthesizes information from multiple sources to produce research-grade reports with citations.
- **Gemini-2.5-Pro Deep Research**: Google's advanced research agent featuring multi-turn planning, deep web navigation, and multi-source evidence integration.

We evaluate these baselines directly on their official API and web applications, the evaluation date is September 2025. For consistency across different systems, we use the following unified prompt template to get the report. As commercial "black-box" systems, we have no control over their internal search routing or region settings. We utilized English prompts for task instructions across all models, while keeping entity names (e.g., company names) in their original Chinese characters to ensure correct query interpretation.

We argue this setting is fair and potentially disadvantageous to FinSight for two reasons: (1) **Relevance over Bias**: Retrieval quality is driven primarily by query language and specificity rather

Table 9: Length-controlled evaluation results. FinSight (10k) represents reports truncated to 10,000 words.

Method	Cons.	Faith.	T-I.	Rich.	Cover.	Ins.	Logic	Lang.	Vis.	Avg.
OpenAI DR	5.60	7.45	4.90	6.35	6.40	5.90	6.90	6.85	4.65	6.11
Gemini-2.5-Pro DR	7.10	6.80	4.65	7.45	7.75	7.85	7.65	7.85	4.25	6.82
FinSight (Full)	6.85	7.50	7.85	8.70	8.30	8.45	8.05	8.10	9.00	**8.09**
FinSight (10k)	6.20	6.70	7.00	6.60	6.75	7.05	6.60	7.50	8.05	6.93

than region settings. The China region setting was necessary to retrieve specific local filings. (2) **Commercial Advantage**: Commercial deep research systems often have access to high-quality search resources and curated financial databases with sophisticated internal query rewriting. In contrast, FinSight relies solely on the open Google Search API.

PROMPT

Please help me write a detailed research report on the corporate finance of {topic}, which should be rich in both text and charts. Give me the standardized citations at the end of the report (including serial numbers and corresponding references).

E Implementation Details of FinSight

Backbone For Multi-source Data Collection Agent, Deep Search Agent and Data Analysis Agent, we use the DeepSeek-V3 as the backbone model. For Report Generation Agent, we use DeepSeek-R1 as the backbone model. The maximum input length is 81,920 tokens, and the maximum output length is 16,384 tokens.

Data Collection We implement the financial api tool based on akshare [2] package in Python. For web search, we use the Google Search API and the number of retrieved results fixed at the top 10. For web content acquisition, we employ Playwright [3] to simulate a browser for webpage content extraction.

Retrieval We use Qwen3-Embedding-0.6B to generate embeddings for data and CoA segments. Then we use the cosine similarity to select the relevant data and CoA segments for each section.

Iterative Vision-Enhanced Mechanism We use the Qwen2.5-VL-72B as the critic vision-language model in the chart generation stage. To balance effectiveness and cost, we perform three iterations of the critic process.

[2]https://github.com/akfamily/akshare

[3]https://playwright.dev/

Ablation Study We conduct ablation study on 5 company questions, which includes: Cambricon Technologies, Li Auto-W, Pop Mart, 3SBio, China Mobile. Some variants are as follows:

- **w/o Iteration Vision-Enhanced Mechanism** We remove the iterative refinement process and plot charts in a single pass.
- **w/o Two-Stage Writing Framework** We only concatenate the CoA segments to output the final report.
- **w/o Dynamic Search-Enhanced Strategy** We remove the Dynamic Search-Enhanced Strategy from the Data Collection and Report Generation process.

F Construction of the Financial Report Generation Benchmark

F.1 Questions

We select the most popular five A-share companies, five Hong Kong-stock companies, and ten representative industries from https://www.djyanbao.com as the benchmark research questions. These companies and industries cover a diverse set of market sectors and provide a comprehensive foundation for evaluating the effectiveness of deep research systems.

F.2 Golden Referenced Reports

To establish human expert-level benchmark, we collect the latest equity and industry research reports from well-known Chinese securities firms, as shown in Table 10. These golden references cover both company-level and industry-level analyses across A-shares, Hong Kong stocks, and major industries. We will released all golden reports after camera-ready version.

F.3 Evaluation Metrics

We further illustrate the metrics we used for evaluation:

Table 10: Golden Referenced Reports from Chinese Securities Firms

Market	Company / Industry	Securities Firm
A-shares	SMIC (688981)	Soochow Securities
	Cambricon Technologies (688256)	Donghai Securities
	China Mobile (600941)	Zhongtai Securities
	Skshu Paint (603737)	Huatai Securities
	Yiwu China Commodities City (600415)	Guolian Minsheng Securities
Hong Kong Stocks	Pop Mart (09992)	Zhongtai Securities
	SenseTime (00020)	Zhongtai Securities
	Li Auto-W (02015)	Huayuan Securities
	3SBio (01530)	Huatai Securities
	UBTECH Robotics (09880)	Guohai Securities
Industries	Semiconductor Industry	Kaiyuan Securities
	Food & Beverage Industry	Huachuang Securities
	Basic Chemical Industry	Zhongtai Securities
	Steel Industry	Orient Securities
	Construction & Decoration Industry	Guosheng Securities
	Environmental Protection & Public Utilities (Controlled Nuclear Fusion)	Huachuang Securities
	Light Manufacturing (Durable Consumer Goods)	Guotai Haitong Securities
	K12 Education Industry	Guosheng Securities
	Media Industry (Short Drama Overseas Expansion)	Soochow Securities
	Transportation (Cross-border E-commerce Logistics)	Maigao Securities

(1) Factual Metrics Measure the textual quality and factual accuracy of the final report.

- **Core Conclusion Consistency**: Whether the core conclusions in the generated report are consistent with those in the reference report.
- **Textual Faithfulness**: Whether the arguments in the report are properly supported by citations from the reference.
- **Text-Image Coherence**: Whether the report integrates images into the discussion, and whether the textual and visual descriptions align.

(2) Analysis Effectiveness Measure whether the financial report provides sufficient information and insights for investors.

- **Information Richness**: The number of distinct information points included in the report.
- **Coverage**: The extent to which key information from the golden reference report is covered.
- **Analytical Insight**: Whether the report provides critical analysis, original insights, and forward-looking recommendations.

(3) Presentation Quality Measure the presentation quality of the final report.

- **Structural Logic**: The logical organization of each section and the overall structural soundness of the report.
- **Language Professionalism**: Whether the language conforms to financial terminology, using the golden report as a reference.
- **Chart Expressiveness**: The effectiveness of charts in supporting the narrative, including their informativeness and aesthetic quality.

G Evaluation Details

G.1 LLM Evaluation Process

We adopt `Gemini-2.5-Pro` as the backbone evaluation model. To ensure fair comparison across reports, we employ a list-wise evaluation strategy, where the model is provided with all candidate reports along with the golden reference report and assigns scores accordingly. The nine metrics mentioned above can be divided into two parts, one is unrelated to the golden report and the other is related to the golden report. For these two types, we have designed two types of prompts, which are listed below.

Evaluation Instruction for Golden Report Irrelevant Metrics

[TASK]
Your task is to act as an expert financial analyst and editor. You will perform a rigorous, **comparative evaluation** of a list of financial research reports. Your goal is to produce a structured critique for each report based on how effectively it addresses the central **Research Question**, using the provided **Golden Standard Report** as a quality benchmark.
[INPUTS]
* **Research Question:** Research Question * **Golden Standard Report:** Given in file format, the one starting with 'golden' is the 'golden standard report' * **Reports to Evaluate:** Reports
[EVALUATION METHODOLOGY]
To ensure fairness and accuracy, you must follow this three-step process for **each report** in the 'Reports to Evaluate' list:
1. **Step 1: Establish the Benchmark (Internal Thought Process)**
* For each of the six evaluation dimensions, first thoroughly analyze the **Golden Standard Report**. Identify its key characteristics, depth, and quality to create a mental benchmark for what constitutes a high-quality, professional report (which corresponds to a score of 7).
2. **Step 2: Comparative Analysis (Internal Thought Process)**
* Now, analyze the report currently being evaluated. For each dimension, find concrete evidence (e.g., specific quotes, data points, chart quality, structural features). * **Directly compare** this evidence against the benchmark established in Step 1. Note where the report meets, exceeds, or falls short of the Golden Standard.
3. **Step 3: Score and Justify (Final Output Generation)**
* Based on the comparison in Step 2, assign a score from 1 to 10 for the dimension, following the 'Benchmark-Based Scoring' rules below. * Write a **concise, one-sentence rationale** that justifies your score by referencing your comparative findings.
[SCORING GUIDELINES]
Adhere strictly to these principles to maintain objectivity:
* **Benchmark-Based Scoring:**
* **The Golden Standard Report is the benchmark for a score of 7.** * A report demonstrating a **similar level of quality**, depth, and execution as the Golden Standard on a specific dimension should receive a score of **7**. * Scores of **8-10** are reserved for reports that **demonstrably exceed** the Golden Standard in that dimension (e.g., providing deeper insights, more comprehensive data, or superior visualizations). * Scores of **1-6** indicate that the report **falls short** of the Golden Standard's quality in that dimension, with the score reflecting the degree of the gap.
* **Justification for Extremes:** Scores of **9-10** (exceptional) or **1-2** (critically flawed) require a particularly strong and specific justification in the rationale.
[EVALUATION FRAMEWORK and CRITERIA]
Dimension 1: Information Richness (Score 1-10)
* **Definition:** Measures the concentration of substantive, verifiable facts and data points relevant to the research question, while minimizing filler content.
Dimension 2: Textual Faithfulness (Score 1-10)
* **Definition:** Measures whether significant claims, data, and forecasts are verifiably supported by provided "References / Data Sources".
Dimension 3: Text-Image Coherence (Score 1-10)
* **Definition:** Assesses if charts and tables are consistent with the text and if the text provides meaningful interpretation that supports the core analysis.
Dimension 4: Analytical Insight (Score 1-10)
* **Definition:** Evaluates the quality of the analysis, focusing on critical thinking, original insights, and actionable, forward-looking conclusions that directly address the research question.
Dimension 5: Structural Logic (Score 1-10)
* **Definition:** Measures the structural integrity and logical flow of the argument, assessing if the report builds a clear and compelling case from evidence to conclusion.
Dimension 6: Chart & Table Expressiveness (Score 1-10)
* **Definition:** Focuses on the quality of data visualizations themselves—their clarity, ability to reveal patterns, and effectiveness in communicating key information.
[OUTPUT FORMAT]
Provide your evaluation in the following strict JSON format. **For each score, you must provide a brief, one-sentence rationale.** Do not add any conversational text outside of this structure. Use the file name of each report as its report id.
Now start your evaluation of the given reports. Carefully read each report and give a score.

Evaluation Instruction for Golden Report Relevant Metrics

[ROLE] You are an expert financial analyst and editor, specializing in the comparative analysis of research reports. **[TASK]** Your task is to rigorously evaluate a list of **Generated Reports** by comparing each one against a **Benchmark Report** (a professionally written 'gold standard'). You will assess each Generated Report's quality across three key dimensions on a scale of 1 to 10, producing a structured JSON output with scores and justifications.
[INPUTS]
1. **'Benchmark Report'**: A high-quality, professional research report that serves as the "gold standard" for this evaluation. All comparisons should be made against this document. The file name of the benchmark report begins with "golden_". 2. **'Generated Reports'**: A list of one or more reports to be evaluated against the Benchmark Report. 3. **'Report ID'**: An identifier for each Generated Report. Use the file name as the report ID.
[EVALUATION METHODOLOGY]
To ensure fairness and accuracy, you must follow this three-step process for **each Generated Report**:
1. **Step 1: Establish the Benchmark (Internal Thought Process)**

* For each of the three evaluation dimensions, first thoroughly analyze the **Benchmark Report**. Identify its key characteristics, depth, and quality to create a mental benchmark for what constitutes a score of **7**.
2. **Step 2: Comparative Analysis (Internal Thought Process)**
* Now, analyze the Generated Report. For each dimension, find concrete evidence (e.g., specific conclusions, data points included/omitted, linguistic style). * **Directly compare** this evidence against the benchmark established in Step 1. Note where the report meets, exceeds, or falls short of the Benchmark Report.
3. **Step 3: Score and Justify (Final Output Generation)**
* Based on the comparison in Step 2, assign a score from 1 to 10 for the dimension, following the 'SCORING GUIDELINES' below. * Write a **concise, one-sentence rationale** that justifies your score by referencing your comparative findings.
[SCORING GUIDELINES]
Adhere strictly to these principles to maintain objectivity:
* **Benchmark-Based Scoring:** * **The Benchmark Report is the standard for a score of 7.** * A report demonstrating a **similar level of quality**, depth, and execution as the Benchmark Report on a specific dimension should receive a score of **7**. * Scores of **8-10** are reserved for reports that **demonstrably exceed** the Benchmark Report in that dimension (e.g., providing a more nuanced conclusion, broader data coverage, or more sophisticated language). * Scores of **1-6** indicate that the report **falls short** of the Benchmark Report's quality in that dimension, with the score reflecting the degree of the gap. * **Justification for Extremes:** Scores of **9-10** (exceptional) or **1-2** (critically flawed) require a particularly strong and specific justification in the rationale.
[EVALUATION FRAMEWORK & CRITERIA]
Dimension 1: Core Conclusion & Data Consistency (Score 1-10)
* **Definition:** Measures the alignment of the Generated Report's core thesis, key arguments, and supporting data points with those presented in the Benchmark Report.
Dimension 2: Information Coverage (Score 1-10)
* **Definition:** Assesses the extent to which the Generated Report includes the key information points, topics, and analytical angles present in the Benchmark Report.
Dimension 3: Professional Language & Tone (Score 1-10)
* **Definition:** Evaluates the linguistic quality of the Generated Report, using the Benchmark Report's writing style, tone, and vocabulary as the standard for professional financial analysis.
[OUTPUT FORMAT] Provide your evaluation in the following strict JSON format. For each score, you must provide a brief, one-sentence rationale that explains the score relative to the benchmark. Do not add any conversational text outside of this structure.
Now start your evaluation of the given reports. Carefully read each report and give a score.

G.2 Human Evaluation Process

To validate our automated evaluation and substantiate claims about report quality, we conducted a comprehensive human evaluation study.

We recruited **6 graduate students with financial backgrounds** (majoring in Finance, Economics, or related fields) to serve as expert annotators. To save costs, we selected the two strongest baselines, **Gemini-2.5-Pro Deep Research** and **OpenAI Deep Research**, to compare against `FinSight`. Each annotator reviewed a random subset of 10 research topics, evaluating all three systems' outputs for each topic. In reviewing process, raters were provided with "Golden Reports" (professional analyst reports from top-tier securities firms) as ground truth references to anchor their judgments.

Scoring Protocol. To manage cognitive load when evaluating long-form reports, raters scored on a 0–5 scale with 0.5 increments across three consolidated dimensions: *Factual* (combining Consistency, Faithfulness, Text-Image Coherence), *Analytical* (combining Richness, Coverage, Insight), and *Presentation* (combining Logic, Language, Visualization). Scores were scaled ($\times 2$) to align with our 0–10 automated metrics.

Table 11: Human evaluation scores (scaled to 0-10).

Model	Fact.	Anal.	Pres.	Total
OpenAI DR	5.93	5.81	4.75	5.50
Gemini-2.5-Pro DR	**6.86**	6.73	4.73	6.11
`FinSight`	6.68	**7.17**	**7.48**	**7.11**

Table 12: Human-LLM alignment and inter-rater reliability metrics.

Dimension	Pearson r	Krippendorff's α
Factual	0.6360	0.4667
Analytical	0.6003	0.4752
Presentation	0.6757	0.8570
Total Score	**0.7587**	**0.6474**

We calculated Inter-Rater Reliability using Krippendorff's Alpha (α) and Human-LLM Alignment using Pearson correlation coefficient (r). Our key findings are as follows: (1) `FinSight` achieves the highest total score (7.11), significantly outperforming both commercial baselines. (2) The strong positive correlation between human and LLM scoring (Pearson $r > 0.75$ for Total Score) validates the reliability of our automated evaluation framework. (3) The overall inter-rater reliability ($\alpha = 0.64$) indicates solid consensus among experts, with exceptionally high agreement in the Presentation dimension ($\alpha = 0.86$), confirming that `FinSight`'s multimodal capabilities provide objectively recognizable advantages.

The instruction for human raters is as follows.

Evaluation Instruction for Human Raters

General Instructions

Thank you for participating in this evaluation. Please assess each report independently based on the three core dimensions defined below. Each dimension is scored on a **1 to 5 point scale**, allowing for half-points (e.g., 3.5).

- **5 points (Excellent):** Significantly exceeds expectations; outstanding performance in all aspects.
- **4 points (Good):** Solid and reliable; comprehensively meets all requirements for a professional report.
- **3 points (Passable):** Fundamentally adequate, but with clear deficiencies in some areas.
- **2 points (Poor):** Contains serious flaws; fails to deliver core value.
- **1 point (Very Poor):** Contains almost no usable information; logically incoherent or factually incorrect.

Dimension 1: Factual - Accuracy & Comprehensiveness

Definition: Assesses the **truthfulness, completeness,** and **objective evidence** of the information provided in the report. This dimension concerns the solidity of the report's foundation.

Score	Evaluation Criteria
5 (Excellent)	Information is extremely dense, facts are cross-verified and accurate, all key topics are covered, and crucial data is clearly supported by sources.
4 (Good)	Information is solid, facts are generally accurate, most key topics are covered, and major data points are supported by sources.
3 (Passable)	Contains basic facts, but coverage is insufficient (e.g., missing key information points), or there are minor factual errors / missing sources.
2 (Poor)	Contains numerous factual errors or severe gaps in information; most claims are not supported by data or sources.
1 (Very Poor)	Filled with unverified information, obvious factual errors, or large-scale content omissions.

Dimension 2: Analytical - Depth & Logic

Definition: Assesses the **quality of analysis, insightfulness,** and **argumentative structure** of the report. This dimension concerns whether the report provides added value beyond a simple recitation of facts.

Score	Evaluation Criteria
5 (Excellent)	Insights are profound, drawing unique and forward-looking conclusions from the data. The logical chain is complete, rigorous, and highly persuasive.
4 (Good)	Analysis is reasonable and capable of effective deduction based on facts. The logic is clear, the structure is complete, and the conclusion is consistent with the argumentation.
3 (Passable)	Contains basic analysis, but lacks depth (often just restating facts). The logic is generally coherent but not sufficiently rigorous.
2 (Poor)	Analysis is superficial or contains logical leaps. There is a weak connection between arguments and evidence; the structure is chaotic.
1 (Very Poor)	Almost no analysis, or filled with logical contradictions. Fails to form a coherent argument.

Dimension 3: Presentation - Quality & Professionalism

Definition: Assesses the **readability, effectiveness of charts,** and **professionalism of the language**. This dimension concerns whether the report can be understood efficiently and clearly.

Score	Evaluation Criteria
5 (Excellent)	Language is precise, professional, and authoritative. Charts are exceptionally well-designed, perfectly complementing the text and greatly enhancing the argument.
4 (Good)	Language is professional and fluent. Charts are clear, easy to understand, and effectively support the text's points; figure-text consistency is good.
3 (Passable)	Language is generally professional but occasionally verbose or inappropriate. Chart quality is average (e.g., unclear, low information), or the connection to the text is weak.
2 (Poor)	Language is unprofessional or contains many errors. Chart quality is poor (e.g., misleading, unreadable), or there is a serious disconnect between figures and text.
1 (Very Poor)	Language is confusing and difficult to read. No charts are used, or the charts provided are completely ineffective.

G.3 Citation Accuracy Evaluation

To rigorously evaluate the faithfulness of generated citations, we conducted a comprehensive manual verification study.

Methodology. Human experts checked the top 50 citations in each generated report to verify whether the cited source actually supported the

Table 13: Citation verification results across all company-level and industry-level tasks.

Model	Total Checked	Accuracy
FinSight	469	**72.92%** (342/469)
Gemini-2.5-Pro DR	414	69.81% (289/414)

generated claim. For each citation, annotators classified it as *Accurate* (the source directly supports the claim), *Partially Accurate* (the source is related but does not fully support the claim), or *Inaccurate* (the source is irrelevant or contradicts the claim). We report the overall accuracy as the proportion of Accurate citations.

As shown in Table 13, even while generating a higher volume of citations, FinSight maintains higher accuracy. We attribute this to our **Two-Stage Writing Framework** and generative retrieval mechanism, which identifies references during the drafting process rather than via post-hoc appending.

Citation Authority Analysis. We further analyzed source quality by classifying citations into three authority levels:

- **High Authority**: Government/Regulatory bodies (SEC, IMF), Official Company Filings, Top Academic/Research Institutions.
- **Medium Authority**: Mainstream Financial Media (Bloomberg, Reuters), Known Market Research Firms.
- **Low Authority**: Social Media, Personal Blogs, Content Farms, or unverified aggregators.

FinSight utilizes High Authority sources at a rate comparable to commercial baselines ($\sim$36.5%). The slightly higher usage of Low Authority sources compared to OpenAI DR reflects our reliance on open web search versus proprietary filtered databases, indicating a direction for future refinement.

G.4 Key Fact Recall Evaluation

Directly measuring the factuality of long reports is challenging due to the absence of strict ground truth. To quantify factual accuracy, we introduced a **Golden Facts Evaluation** methodology.

Methodology. We extracted **13 core financial indicators** from the professional Golden Reports as ground truth, covering: (1) Profitability (Gross Margin, Net Margin), (2) Growth (Revenue Growth, Profit Growth), (3) Financial Health (Cash Flow, Debt Ratio), (4) Valuation (PE Ratio, PB Ratio), and (5) Efficiency (ROE, ROA). Human annotators then manually verified how many of these specific data points were accurately retrieved and reported by each model across company-level tasks.

FinSight achieves significantly higher recall rate, demonstrating superior coverage of critical financial data compared to commercial deep research systems.

H A Case of Company Research Question

To demonstrate the practical application of our system, this section shows the case of **SenseTime Technology (0020.HK)**, a leading artificial intelligence company in China.

We present the collecting tasks of the Data Collection process in Table 16, and an analytical tasks of the Data Analysis process in Table 17.

I Report Gallery

We have presented an overview of the report generated by ours here, and the complete report can be obtained from https://anonymous.4open.science/r/FinSight-5841.

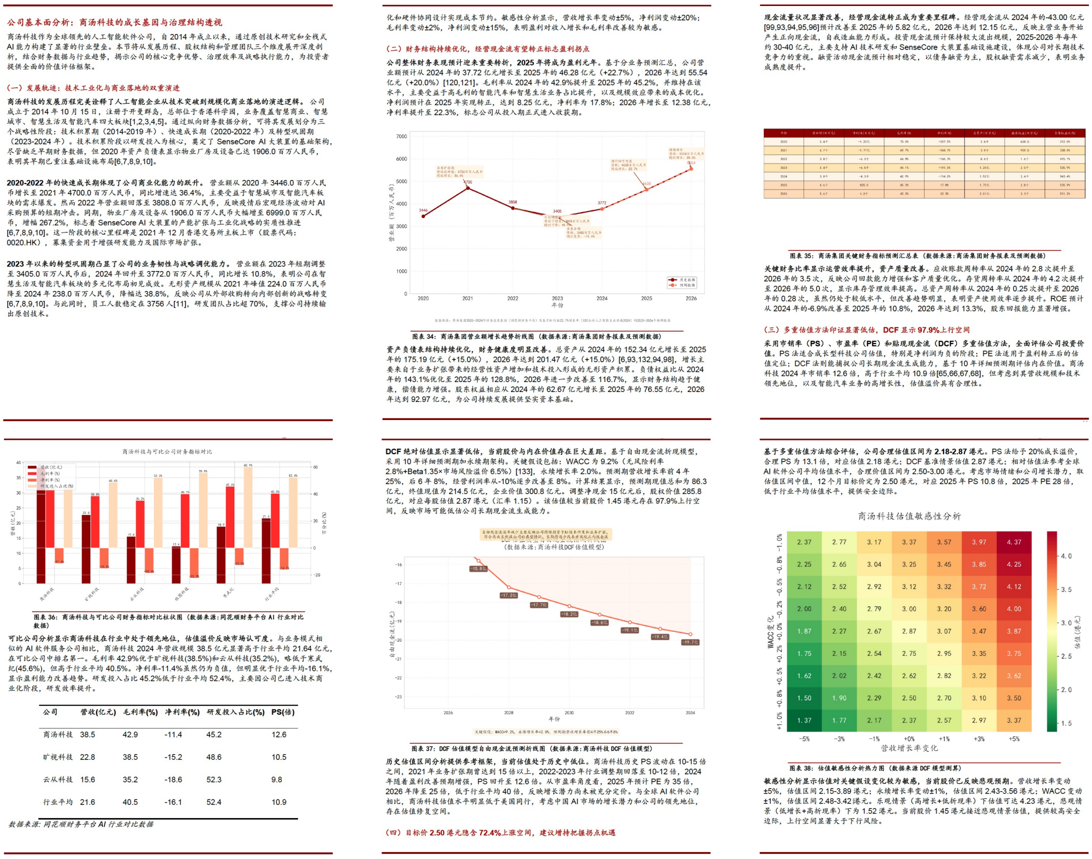

Figure 9: The final report of The SenseTime (part).

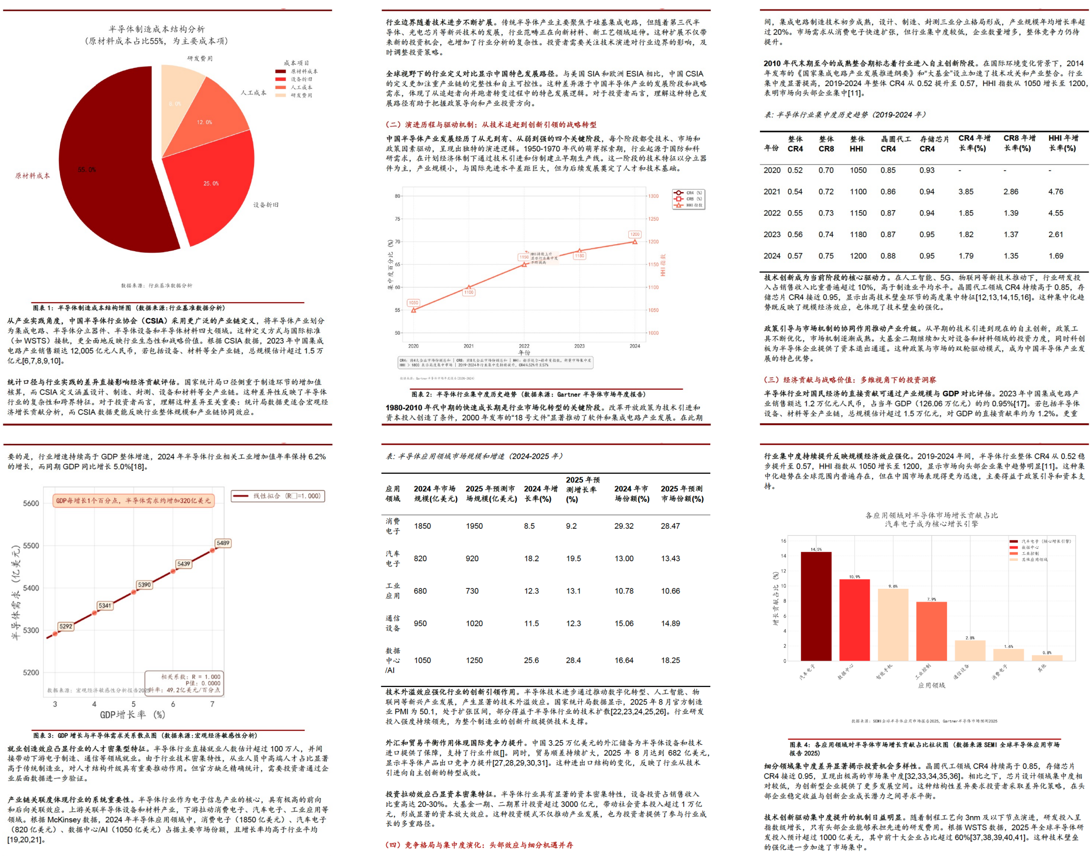

Figure 10: The final report of semiconductor industry (part).

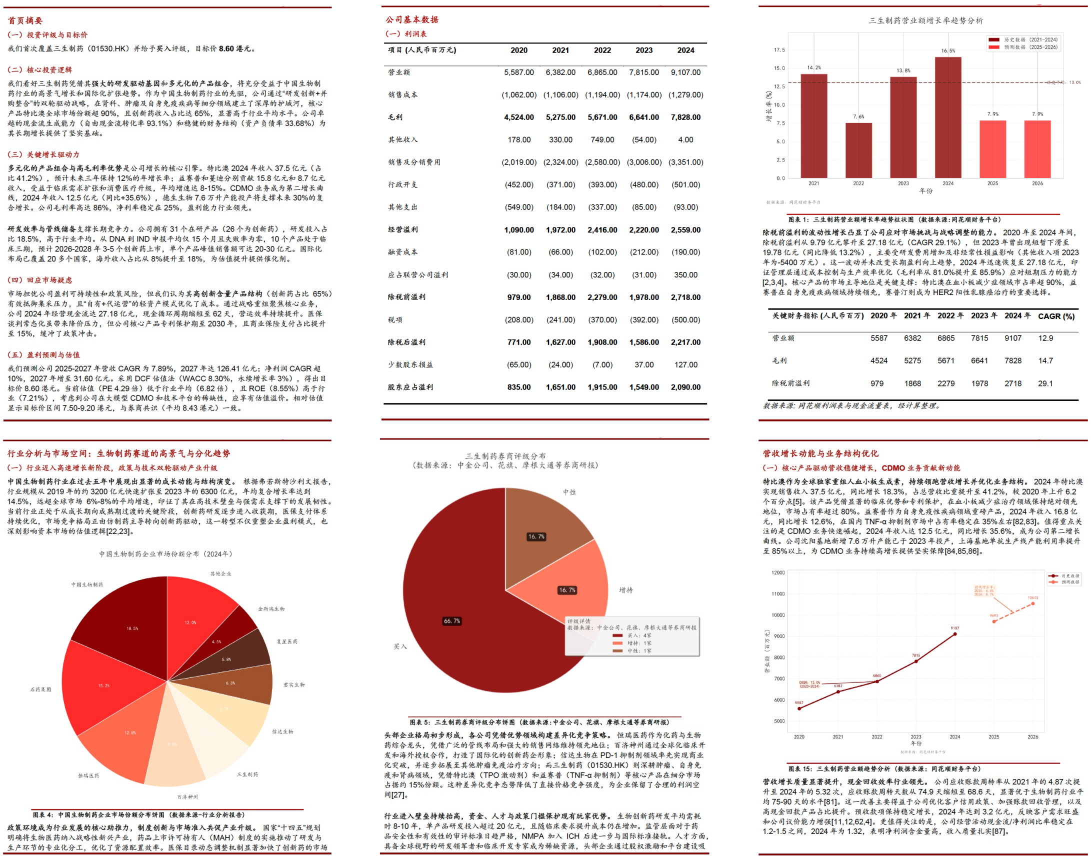

Figure 11: The final report of The 3SBio Inc. (part).

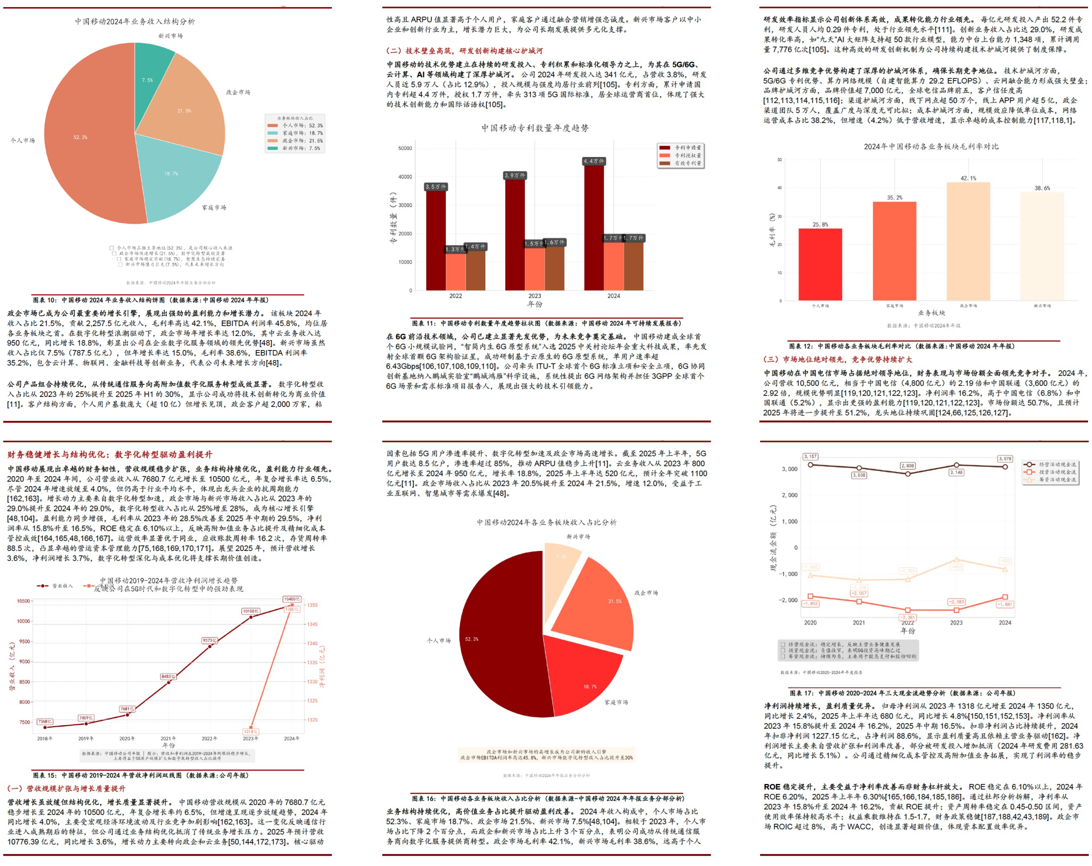

Figure 12: The final report of The China Mobile (part).

Table 14: Key financial information recall across methods.

Method	Avg. Hits (out of 13)	Recall Rate	Relative
FinSight	**7.1**	**54.6%**	–
Gemini-2.5-Pro DR	5.0	38.5%	-29.6%
OpenAI DR	3.9	30.0%	-45.1%

Table 15: Distribution of citation authority across different methods.

Model	High	Medium	Low	Total
Gemini-2.5-Pro DR	312 (36.7%)	304 (35.8%)	233 (27.4%)	849
OpenAI DR	348 (35.0%)	504 (50.7%)	142 (14.3%)	994
FinSight	**334 (36.5%)**	319 (34.8%)	263 (28.7%)	916

Table 16: The predefined and brainstormed data collection tasks.

Data Collection

```
Predefined Tasks:
"company": [
    {"name": "Balance Sheet"},
    {"name": "Income Statement"},
    {"name": "Cash Flow Statement"},
    {"name": "Basic Stock Information"},
    {"name": "Shareholder Structure"},
    {"name": "Stock Price"},
    {"name": "Stock-related Financial Data"},
    {"name": "CSI 300 Daily Index Data"},
    {"name": "Hang Seng Daily Index Data"},
    {"name": "NASDAQ Daily Index Data"},
    {"name": "Investment Rating",
    "description": "Collect analyst investment ratings and target prices from
    major securities firms (e.g., CITIC Securities, CICC). Select the most
    frequent rating across institutions."
    },
    {"name": "Price-to-Sales Ratio (PS)",
    "description": "Collect the company's PS data (PS = Stock Price / Sales per
    Share)."
    },
    {"name": "Return on Equity (ROE, %)",
    "description": "Collect the company's ROE data, defined as (Net Income
    Attributable to Shareholders / Equity) or equivalently (Net Profit Margin x
    Asset Turnover x Equity Multiplier)."
    },
    { "name": "Competitor Data",
    "description": "Collect comparable data from 5-8 major peer companies,
    including financial and operational indicators."
    },
]

Brainstormed Tasks: [
    {"name": "Growth Data for AI Subfields in 2024"},
    {"name": "The Development History of The SenseTime and Main Business"}
]
```

Table 17: The illustration of data analysis process and CoA output of a task.

Data Analysis (We show a predefine task of data analysis, and its CoA output.)

Predefined Task 1: Company Development History and Main Business Analysis
Required Data: Years and corresponding company achievements.
Description: Organize the company's key milestones by year (e.g., founding, expansions, product launches, strategic shifts, and M&A) and present them as a timeline with annotated events. Also summarize its current core business scope, including main products, business segments, and market positioning.

CoA Output:

SenseTime: Evolution from Computer Vision Pioneer to AI Platform Enterprise
Executive Summary
As a global leading artificial intelligence software company, SenseTime has grown into an AI giant valued at over USD 10 billion since its founding in 2014. With the mission of "persisting in originality and enabling AI to advance human progress", the company has built a comprehensive AI ecosystem covering four major business segments: Smart Business, Smart City, Smart Life, and Smart Auto, through 11 years of continuous innovation and business expansion [Source: SenseTime Group Company Profile]. By 2024, SenseTime employed 3,756 people, with operations spanning over 20 cities worldwide and R&D centers in Hong Kong, Shanghai, Beijing, and Shenzhen [Source: SenseTime Basic Information].
The company's growth trajectory reflects a typical path of technology firms: from technical breakthroughs, to productization, then to platformization and ecosystemization. Its IPO on the Hong Kong Stock Exchange in 2021 (Ticker: 0020.HK) marked the beginning of a new capital-intensive phase, while the launch of the "Riyadh New" large model system in 2023 signified its strategic pivot toward generative AI [Source: Development History Search Data]. Currently, SenseTime reports annual revenue of RMB 3.772 billion, maintaining technological leadership in computer vision and artificial intelligence [Source: Stock 0020 Income Statement Data].

Development History Analysis
Early Stage of Founding and Technology Accumulation (2014–2016)
Founded on October 15, 2014, in the Cayman Islands by the CUHK Multimedia Lab team, SenseTime initially focused on core R&D in computer vision and deep learning [Source: SenseTime Basic Information]. In 2015, it launched its first facial recognition algorithm and completed its first funding round, laying the foundation for subsequent commercialization. By 2016, it had become the world's most valuable AI unicorn, reflecting strong capital market recognition.

@import "SenseTime Development Timeline (2014-2025), horizontal axis = year, vertical axis = milestone importance, annotated with key events (milestone_df)"

Productization and Market Expansion (2017–2019)
In 2017, the launch of the SenseAR augmented reality platform marked the transition from pure research to productization. By 2018, it had completed Series C financing with a valuation exceeding USD 4.5 billion. In 2019, it rolled out Smart City solutions, partnering with key domestic cities and expanding its scope from technology output to integrated solutions.
Platformization and Capitalization (2020–2022)
In 2020, the release of the SenseCore AI Infrastructure introduced the concept of industrializing AI, aiming to reduce the cost of AI production factors. On December 30, 2021, SenseTime was listed on the Hong Kong Stock Exchange Main Board, raising about HKD 6 billion [Source: SenseTime IPO Search Data]. In 2022, it launched the "Riyadh New" foundation model system, officially entering the generative AI space and advancing toward multimodal development.
Ecosystemization and Deepening Applications (2023–2025)
In 2023, the company released the "SenseChat" large language model, further completing its model matrix. In 2024, its Smart Auto business grew rapidly, with partnerships across major automakers. By 2025, SenseTime continued to optimize its large model technologies and achieved deep application deployments in healthcare, finance, and manufacturing.

Core Business Structure Analysis

Smart Business (35%)
Core Products: SenseFoundry Enterprise Ark, AI Content Generation Platform Market Positioning: Provides AI solutions for enterprise digital transformation, focusing on finance, retail, and manufacturing. This is the main revenue driver, contributing about 35% of revenue in 2024 [Source: Business Segment Data].
Smart City (30%)
Core Products: City Ark, traffic management, and surveillance systems
Market Positioning: Provides smart governance and public service solutions. Established partnerships with major cities, gaining competitive advantages in smart transportation and public safety.
Smart Life (20%)
Core Products: SenseMARS Metaverse Platform, mobile AI applications
Market Positioning: Consumer-facing AI applications and metaverse ecosystem. Focuses on AR/VR and mobile internet scenarios, with strong growth potential.
Smart Auto (15%)
Core Products: SenseAuto Platform, autonomous driving solutions
Market Positioning: Provides intelligent driving and V2X systems for automakers. This segment grew rapidly in 2024, becoming a new growth engine [Source: Smart Auto Data].

@import "SenseTime 2024 Revenue Breakdown Pie Chart, showing four business segments (business_df)"

Technical Strength and Competitive Advantages
SenseTime possesses deep academic roots and originality in AI technologies. The company has built full-stack AI capabilities across perception intelligence, decision intelligence, content generation, and content enhancement [Source: Company Profile]. By 2024, it had over 8,000 AI-related patents, maintaining leadership in computer vision, large model training, and AI chips [Source: Patent Data].
Core Competitive Advantages:
1. Technological Originality: Commitment to independent fundamental research and development
2. Platformization Capability: SenseCore enables AI industrial-scale production
3. Ecosystem Completeness: Full product portfolio covering four major domains
4. Talent Density: A world-class AI R&D team

Market Position and Prospects
Industry data show SenseTime as a global leader in computer vision, ranking among top players in market share. In 2024, it was positioned in the Leaders Quadrant of the AI maturity curve, highly rated in both innovation and execution [Source: AI Market Growth Data 2024].
Future Directions:
- Large Model Enhancement: Continuous optimization of the ”Riyadh New" large model system
- Vertical Integration: Deepened applications in finance, healthcare, and manufacturing
- International Expansion: Strengthening presence in Southeast Asia and the Middle East
- Tech Convergence: Driving integration of AI with IoT and blockchain

@import "SenseTime Revenue Growth Trend (2020-2024), showing revenue growth from RMB 3.446B to RMB 3.772B (income_statement)"

Conclusion
Through 11 years of development, SenseTime has evolved from a single computer vision technology provider into a comprehensive AI platform enterprise. Its trajectory reflects a clear progression: technology-driven, product iteration, platform building, and ecosystem expansion. With balanced business segments and strong technological capabilities, SenseTime holds a favorable competitive position in the AI industrialization wave.

Looking forward, with continuous AI evolution and expanding applications, SenseTime is poised to play a more significant role in the global AI market. To remain competitive, the company must sustain R&D investment, deepen vertical applications, and enhance international collaborations in response to intensifying market and technological challenges.