

# FINDEEPRSEARCH: Evaluating Deep Research Agents in Rigorous Financial Analysis

FENGBIN ZHU<sup>\*</sup>, XIANG YAO NG<sup>\*</sup>, ZIYANG LIU<sup>\*</sup>, CHANG LIU<sup>◊</sup>, XIANWEI ZENG<sup>\*</sup>, CHAO WANG<sup>\*</sup>, TIANHUI TAN<sup>◊</sup>, XUAN YAO<sup>◊</sup>, PENGYANG SHAO<sup>\*</sup>, MIN XU<sup>\*</sup>, ZIXUAN WANG<sup>\*</sup>, JING WANG<sup>\*</sup>, XIN LIN<sup>\*</sup>, JUNFENG LI<sup>\*</sup>, JINGXIAN ZHU<sup>◊</sup>, YANG ZHANG<sup>\*</sup>, WENJIE WANG<sup>\*</sup>, FULI FENG<sup>\*</sup>, RICHANG HONG<sup>◊</sup>, HUANBO LUAN<sup>\*</sup>, KE-WEI HUANG<sup>◊</sup>, TAT-SENG CHUA<sup>\*</sup>,

<sup>\*</sup>National University of Singapore, Singapore

<sup>\*</sup>6Estates Pte Ltd, Singapore

<sup>◊</sup>Asian Institute of Digital Finance, Singapore

<sup>◊</sup>Hefei University of Technology, China

<sup>\*</sup>University of Science and Technology of China, China

Deep Research (DR) agents, driven by Large Language Models (LLMs), have recently garnered increasing attention for their capability in conducting complex research tasks. However, existing literature lacks a rigorous and systematic evaluation of DR agent's ability in critical analysis tasks. To fill this gap, we first propose *HisRubric*, a novel evaluation framework with an expert-designed hierarchical analytical structure and a fine-grained grading rubric for rigorously assessing DR agents in corporate financial analysis. This framework mirrors the professional analyst's workflow, progressing from data recognition to metric calculation, and finally to strategic summarization and interpretation. Built on this framework, we construct a FINDEEPRSEARCH benchmark that comprises 64 listed companies from 8 financial markets across 4 languages, encompassing a total of 15,808 grading items. We further conduct extensive experiments on the FINDEEPRSEARCH with 16 representative methods, including 6 DR agents, 5 LLMs equipped with both deep reasoning and search capabilities, and 5 LLMs with deep reasoning capabilities only. The results reveal the strengths and limitations of these methods across diverse capabilities, financial markets, and languages, offering valuable insights for future advancements. The benchmark and leaderboard will be made publicly available soon.

## ACM Reference Format:

Fengbin Zhu<sup>\*</sup>, Xiang Yao Ng<sup>\*</sup>, Ziyang Liu<sup>\*</sup>, Chang Liu<sup>◊</sup>, Xianwei Zeng<sup>\*</sup>, Chao Wang<sup>\*</sup>, Tianhui Tan<sup>◊</sup>, Xuan Yao<sup>◊</sup>, Pengyang Shao<sup>\*</sup>, Min Xu<sup>\*</sup>, Zixuan Wang<sup>\*</sup>, Jing Wang<sup>\*</sup>, Xin Lin<sup>\*</sup>, Junfeng Li<sup>\*</sup>, Jingxian Zhu<sup>◊</sup>, Yang Zhang<sup>\*</sup>, Wenjie Wang<sup>\*</sup>, Fuli Feng<sup>\*</sup>, Richang Hong<sup>◊</sup>, Huanbo Luan<sup>\*</sup>, Ke-Wei Huang<sup>◊</sup>, Tat-Seng Chua<sup>\*</sup>. 2025. FINDEEPRSEARCH: Evaluating Deep Research Agents in Rigorous Financial Analysis. 1, 1 (October 2025), 25 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

---

\*Corresponding author: Fengbin Zhu, fengbin@nus.edu.sg.

Author's Contact Information: Fengbin Zhu<sup>\*</sup>, Xiang Yao Ng<sup>\*</sup>, Ziyang Liu<sup>\*</sup>, Chang Liu<sup>◊</sup>, Xianwei Zeng<sup>\*</sup>, Chao Wang<sup>\*</sup>, Tianhui Tan<sup>◊</sup>, Xuan Yao<sup>◊</sup>, Pengyang Shao<sup>\*</sup>, Min Xu<sup>\*</sup>, Zixuan Wang<sup>\*</sup>, Jing Wang<sup>\*</sup>, Xin Lin<sup>\*</sup>, Junfeng Li<sup>\*</sup>, Jingxian Zhu<sup>◊</sup>, Yang Zhang<sup>\*</sup>, Wenjie Wang<sup>\*</sup>, Fuli Feng<sup>\*</sup>, Richang Hong<sup>◊</sup>, Huanbo Luan<sup>\*</sup>, Ke-Wei Huang<sup>◊</sup>, Tat-Seng Chua<sup>\*</sup>,

<sup>\*</sup>National University of Singapore, Singapore

<sup>\*</sup>6Estates Pte Ltd, Singapore

<sup>◊</sup>Asian Institute of Digital Finance, Singapore

<sup>◊</sup>Hefei University of Technology, China

<sup>\*</sup>University of Science and Technology of China, China.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM XXXX-XXXX/2025/10-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

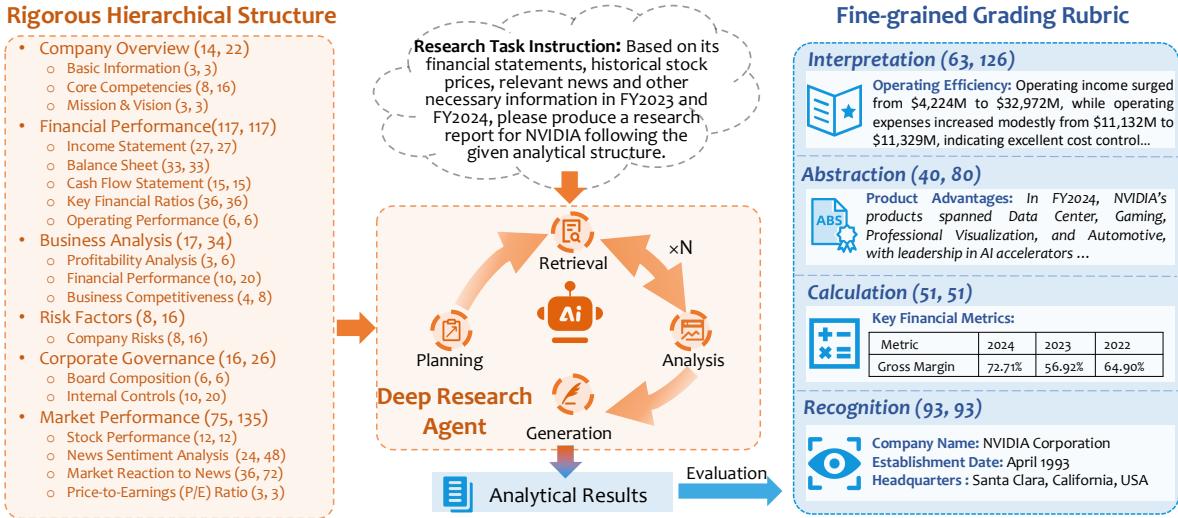


Fig. 1. An overview of the *HisRubric* evaluation framework. The numbers in brackets indicate the number of grading items (left) and the corresponding full marks (right).

## 1 Introduction

The advent of Deep Research (DR) agents, powered by the advancements in Large Language Models (LLMs), marks a pivotal shift in the ways of complex research tasks being tackled [2]. They are capable of automatically navigating the Web, aggregating and synthesizing relevant information, and producing comprehensive reports in response to complex research tasks, such as scientific discovery [9, 15, 36] and financial analysis [6, 14]. Due to such amazing capabilities, the DR agents have rapidly achieved widespread adoption [33], such as Gemini DR [1], OpenAI DR [12], and Grok DR [29]. Yet, rigorous and systematic approaches to evaluating their capabilities remain scarce in current literature, hindering a comprehensive understanding of their strengths and limitations.

Existing evaluation methods generally fall into two groups. On one hand, a line of work focuses on answer-centric verification in a Question Answering (QA) setting, reducing evaluation to a single correctness check while ignoring the substantive analysis outcomes [6, 26, 31, 36]. On the other hand, some research pursues holistic quality assessment through high-level, subjective metrics like "helpfulness" [2, 14], or based on indeterminate report structures [9, 13], which yields scores that are often superficial or irreproducible. Thus, a fundamental tension emerges, as focusing on verifiable facts often overlooks the evaluation of analytical coherence, while an emphasis on holistic quality frequently lacks sufficient grounding in verifiable detail. Critically, a high-quality analysis depends on both a systematic analytical structure (rigor) and specific, accurate claims (precision) simultaneously.

To overcome this, we explore a unified framework that integrates the principal goals of both research streams, which we define as two measurable criteria: *Structural Rigor*, which examines whether the agent's findings and reasoning are organized into a coherent, verifiable analytical structure; and *Information Precision*, which inspects whether its claims are specific, accurate, and traceable to their sources. By combining these two, the framework offers a more complete and faithful measure of an agent's ability to ensure rigorous analytical quality, thereby enhancing the applicability of DR agents in critical real-world scenarios. In this work, we ground the initial development of this framework in financial analysis, specifically focusing on corporate financial analysis. This task serves as an ideal testbed due to its exceptionally clear and strict professional standards. First, it requires a concrete analytical flow for *Structural Rigor*, as analyses must follow standardized structures that cover

company fundamentals, financial tables, and stock price trends, etc. Furthermore, it provides stringent validation of *Information Precision*, demanding error-free reporting of granular details, such as year-over-year revenue growth and specific stock prices.

In this work, we introduce ***HisRubric***, a novel framework built on two key mechanisms: an expert-designed **Hierarchical structure** to guide DR agents to conduct rigorous financial analysis and a fine-grained grading **Rubric** for a comprehensive assessment. Developed with senior financial experts, our hierarchical structure defines a practical analytical structure for corporate financial analysis, comprising 6 major sections and 18 subsections. The Rubric is composed of 247 fine-grained grading items designed to assess 4 progressive capabilities of DR agents, *i.e.*, *Recognition*, *Calculation*, *Abstraction*, and *Interpretation*. These dimensions align closely with established evaluative frameworks for financial analysis and the quality of analyst reports from an academic perspective [3, 5], and are also consistent with best practices recognized in global financial markets from an industry perspective. In practice, leading institutions such as *Institutional Investor's All-America Research Team Awards* and *Refinitiv StarMine Analyst Awards* systematically evaluate research quality based on cognitive accuracy, reasoning depth, and interpretive insight, while the *CFA Institute's Graham & Dodd Awards* highlight excellence in applied financial analysis and communication clarity.<sup>1</sup> Together, these industry standards reinforce the relevance of the four dimensions as key indicators of analytical rigor and professional competence.

With the *HisRubric* framework, we construct a **FINDEEPRSEARCH** benchmark, encompassing companies from 8 financial markets (*i.e.*, United States, United Kingdom, China, Hong Kong, Australia, Singapore, Malaysia, and Indonesia) across 4 languages (*i.e.*, English, Simplified Chinese, Traditional Chinese, Bahasa Indonesia). From each financial market, we select 8 companies, resulting in a total of 64 listed companies with 15,808 grading items. These companies are distributed across 10 industries, defined by the Bloomberg Industry Classification Standard (BICS), including Communications, Energy, Health Care, Materials, and Technology, etc. For each company, DR agents are required to generate a research report that follows the hierarchical analytical structure and is grounded in a diverse set of data, such as financial statements, stock prices, financial news, market indices, etc.

On the constructed **FINDEEPRSEARCH** benchmark, we conduct extensive experiments with 16 representative methods, including 6 DR agents, 5 LLMs with thinking and search capabilities, and 5 LLMs with thinking capability only. The experimental results reveal that: 1) Most methods generally conform to the expert-designed analytical structure, but they consistently fall short in generating precise information. 2) DR agents consistently exhibit superior performance compared to the methods in the other two categories, with their advantage being particularly pronounced in the Recognition and Calculation capabilities. 3) All evaluated methods face significant challenges in mastering the Interpretation capabilities and in performing corporate financial analysis of non-English markets.

In summary, the major contributions of this work are threefold:

- We introduce a novel *HisRubric* evaluation framework built upon a practical hierarchical analytical structure and a fine-grained grading rubric for assessing Deep Research agents in critical and rigorous financial analysis.
- We construct a **FINDEEPRSEARCH** benchmark comprising companies from 8 financial markets across 4 languages, resulting in a total of 64 listed companies with 15,808 grading items.
- We conduct extensive experiments on **FINDEEPRSEARCH** with 16 models, including advanced DR agents and representative LLMs equipped with web search and/or deep reasoning capabilities. The results indicate that while most methods successfully adhere to the prescribed analytical structure, they consistently struggle with producing precise information.

---

<sup>1</sup>See Institutional Investor Research Awards: <https://www.institutionalinvestor.com/research>; Refinitiv StarMine Awards: <https://www.refinitiv.com/en/star-mine>; CFA Institute Graham & Dodd Awards: <https://rpc.cfainstitute.org/research/financial-analysts-journal/graham-and-dodd-awards-of-excellence>.

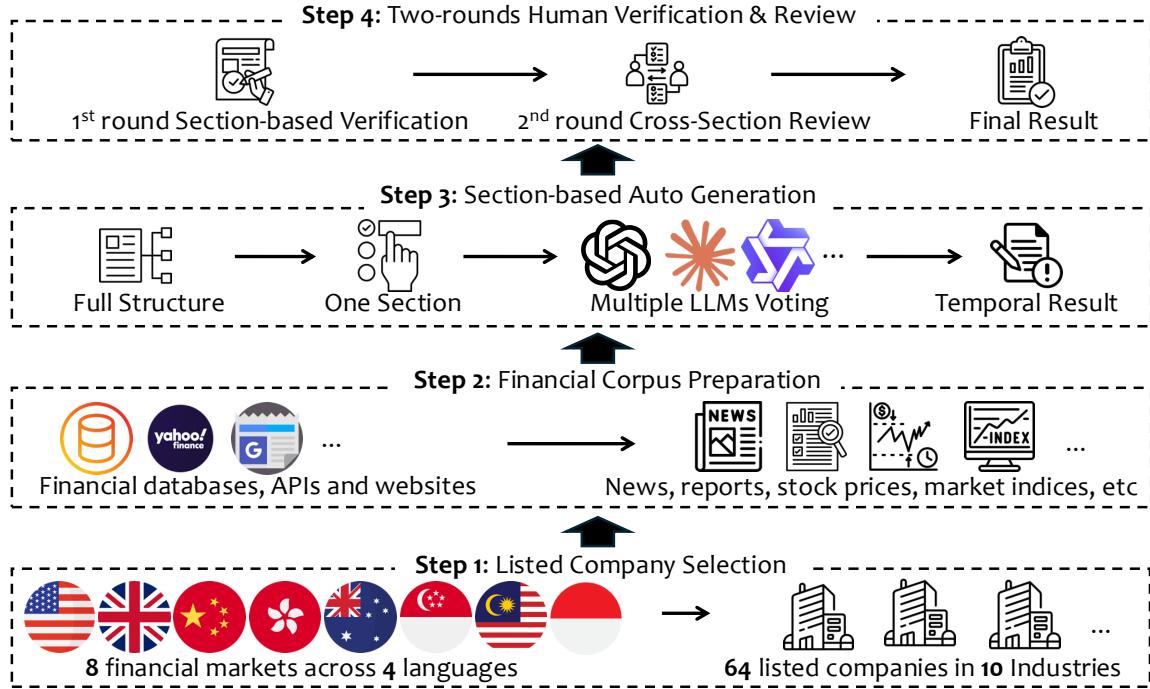


Fig. 2. An overview for constructing FINDEEPRSEARCH.

## 2 HisRubric Framework

In Figure 1, we present the proposed *HisRubric* evaluation framework, which integrates an expert-defined hierarchical analytical structure with a fine-grained grading rubric to systematically assess recognition, calculation, abstraction, and interpretation capabilities of deep research methods.

### 2.1 Task Definition

To ensure the high standard of the research outcomes, we devise a comprehensive hierarchical analytical structure to guide the analysis. Formally, given a research task instruction  $i$  with a desired analytical structure  $S$ , a method  $\mathcal{M}$  is required to produce a research report  $\mathcal{R}$  strictly following the analytical structure  $S$ .

$$\mathcal{R} = \mathcal{M}(i, S) \quad (1)$$

In this work, the instruction  $i$  is provided in natural language. Both the analytical structure  $S$  and the generated research report  $R$  are formatted in Markdown to facilitate easy evaluation.

### 2.2 Rigorous Hierarchical Structure

As shown in Figure 1, to achieve a comprehensive and rigorous evaluation, we employ proficient financial experts to devise a practical hierarchical analytical structure for corporate finance analysis with 6 major sections and 18 subsections, covering the key perspectives in real-world corporate analysis as follows:

- **Section 1: Company Overview.** This section provides a concise overview of the company, including its basic information, industry background, key strengths, and strategic direction. It is divided into 3 subsections: Basic Information, Core Competencies, and Mission & Vision.

- **Section 2: Financial Performance.** This section presents a detailed analysis of the company’s financial health, including primary financial statements and key performance metrics. It comprises 5 subsections: Income Statement, Balance Sheet, Cash Flow Statement, Key Financial Ratios, and Operating Performance.
- **Section 3: Business Analysis.** Through a deep analysis of the obtained data, this section identifies key insights regarding the company’s business, financial performance, and profitability. This section includes 3 subsections: Profitability Analysis, Financial Performance Summary, and Business Competitiveness.
- **Section 4: Risk Factors.** This section identifies and discusses the principal risks the company faces, including market, financial, operational, and regulatory risks, along with the strategies in place to manage them.
- **Section 5: Corporate Governance.** This section outlines the company’s governance framework, including the board of directors, executive leadership, governance policies, and practices, ensuring transparency and accountability. This section contains 2 subsections: Board Composition and Internal Controls.
- **Section 6: Market Performance.** This section provides a comprehensive analysis of the company’s stock performance, the news events that shape its public narrative, and its current market valuation. It is structured into 4 subsections: Stock Performance, News Sentiment Analysis, Market Reaction to News, and Price-to-Earnings Ratio.

### 2.3 Fine-grained Grading Rubric

To facilitate a comprehensive evaluation of the generated financial research report, a fine-grained grading rubric is applied. From each section in the structure, we select specific data items for scoring, termed “*grading items*”, which are designed to ensure full coverage of all key analytical perspectives. Each of these grading items is then mapped to one of four critical capabilities of DR agents:

- **Recognition.** The capability to accurately identify and extract specific factual data from vast and complex data sources, serving as a fundamental skill.
- **Calculation.** The ability to precisely compute and verify numerical values, which is essential for rigorous quantitative analysis.
- **Abstraction.** One critical competency to synthesize complex relationships and summarize valuable patterns, enabling the distillation of essential perspectives from messy data.
- **Interpretation.** The capacity to conduct deep analysis on the existing data to deliver insightful findings and implications, reflecting the highest level of reasoning.

In total, we obtain 247 distinct grading items, and the distribution across the four capabilities is presented on the right of Figure 1. According to financial expert assessment, the competencies assessed under Abstraction and Interpretation are more complex than those under the other two categories. Consequently, items in Abstraction and Interpretation are weighted at 2 marks each, while items in the remaining categories are weighted at 1 mark each. This weighting scheme yields a total possible score of 350 marks.

### 2.4 Evaluation Protocol

To assess the *Information Precision*, for each grading item, we first obtain the predicted answer from the generated result and then compare it with the corresponding ground truth. Three distinct evaluation protocols are applied to different types of grading items.

- *Accuracy:* We employ an advanced LLM to evaluate the correctness by comparing the predicted answer to the ground truth. It gives a score of 1 for a match, otherwise 0. This method is applied to all grading items in *Recognition* and *Calculation* and to a subset of items with concrete answers in *Interpretation*.
- *Claim-based Score:* We first adopt an advanced LLM to identify three to five critical reference claims from the ground truth, depending on the length of the ground truth. For each claim, we apply the LLM to determine whether it is adequately covered in the predicted answer[7]. The proportion of covered claims constitutes this

Table 1. Statistics of FINDEEPRSEARCH.

Statistic	Number
<b>Basic Information</b>	
Number of Languages	4
Number of Financial Markets	8
Number of Industries	10
Number of Selected Companies	64
<b>Analytical Structure</b>	
Number of Major Sections	6
Number of Subsections	18
<b>Grading Items</b>	
Number of Grading Items per Report	247
Full Marks for each Report	350
Total Number of Grading Items	15,808

claim-based score, ranging from 0 to 1. This method is applied to all grading items in *Abstraction* and to a subset of items formed in a summary format in *Interpretation*.

- **Criterion-based Score:** For items requiring nuanced reasoning and qualitative analysis, a simple binary or claim-based evaluation is insufficient. We therefore introduce a criterion-based scoring approach[35]. This process begins by prompting an advanced LLM to act as the role of a financial expert (*e.g.*, a financial professor) to generate a detailed 10-point scoring criterion based on the ground truth. This criterion deconstructs the ideal answer into its core analytical components. Subsequently, the LLM is used to grade the predicted answer against the criterion. The final score is the sum of the awarded points, normalized to a scale of 0 to 1. This method is applied to a subset of the *Interpretation* items where the quality of argumentation and the depth of analysis are key assessment factors.

After summing the scores from all grading items, the total score is normalized by the maximum possible value (*i.e.*, 350) to yield a final score ranging from 0 to 1, termed “accuracy score”.

In addition, we also assess the *Structural Rigor* of the generated markdown result with a rule-based validation method. Our method evaluates structural compliance by scoring the 6 main sections, 18 subsections, and 18 markdown tables. The scoring awards 1 point for each correct element and deducts 1 point for errors, yielding a format score out of a maximum of 42 points. The raw score is then normalized by the maximum (*i.e.*, 42) to produce a final score between 0 and 1, termed “structure score”, which provides a quantitative measure of structural fidelity.

### 3 FINDEEPRSEARCH Benchmark

This section introduces the construction and quality control of FINDEEPRSEARCH, presents a statistical analysis of its properties, and compares it with existing deep research benchmarks.

#### 3.1 Construction of FINDEEPRSEARCH

We illustrate the overall pipeline for constructing our FINDEEPRSEARCH in Figure 2, including four key steps:

- **Step 1: Listed Company Selection** To ensure a comprehensive and diverse evaluation, we select companies from eight financial markets: the United States (US), the United Kingdom (UK), China (CN), Hong Kong (HK), Australia (AU), Singapore (SG), Malaysia (MY), and Indonesia (ID). This selection covers four languages: English (EN), Simplified Chinese (zh-CN), Traditional Chinese (zh-HK), and Bahasa Indonesia (BI). Finally, we obtain 64 listed companies from 10 industries (*i.e.*, Property & Real Estate, Healthcare & Communications, Consumer

Table 2. Analysis of FINDEEPR ESEARCH across 8 financial markets.

Metric	US	UK	CN	HK	AU	SG	MY	ID
#Selected Companies	8	8	8	8	8	8	8	8
Min #Chars per Company	31,985	21,605	15,040	11,940	20,358	16,034	18,683	27,535
Avg #Chars per Company	46,032	31,029	22,928	26,680	25,218	26,420	24,320	30,528
Med #Chars per Company	40,017	30,544	20,835	23,625	23,123	23,560	21,784	29,455
Max #Chars per Company	82,257	40,041	42,713	46,967	36,169	45,036	37,740	36,181
Min #Chars per Grading Item	3	3	3	2	2	3	3	2
Avg #Chars per Grading Item	173	112	78	93	88	94	85	109
Med #Chars per Grading Item	17	16	20	18	16	18	19	18
Max #Chars per Grading Item	3,803	3,526	2,590	3,159	2,332	3,940	1,898	2,235

Discretionary, Consumer Staples, Energy, Health Care, Industrials, Materials, Real Estate, Technology, Utilities) based on the Bloomberg Industry Classification Standard (BICS).

- **Step 2: Financial Corpus Preparation** After the companies are selected, we obtain the associated financial data from a variety of data providers, including established financial databases (e.g., Bloomberg), API services (e.g., Alpha Vantage), Google News), and public financial websites (e.g., Yahoo Finance <sup>2</sup>). The collected data includes fundamental data, annual reports, historical stock prices and market indices, and relevant news, etc.
- **Step 3: Section-based Auto Generation** Next, we generate a reference report for each selected company. For every section in the expert-designed structure, multiple Large Language Models (LLMs) are leveraged to take the relevant corpus as input and generate candidate results separately. The predominant result for each grading item in the section is then selected among the multiple candidates as the definitive value. Upon completion of all sections, these results are synthesized into a provisional full report for subsequent verification.
- **Step 4: Two-rounds Human Verification & Review** Finally, our financial experts conduct two rounds of data verification. To enhance consistency and efficiency of the human review process, we conduct a section-based verification technique in the first round. First, we divide the financial experts into 6 groups, with each group responsible for verifying a specific section. This round is concluded only after all sections have been verified. In the second round, a panel of senior financial experts is assigned to review the entire report, performing a cross-verification of all sections to ensure both consistency and accuracy.

### 3.2 Quality Control

We maintain the high quality of FINDEEPR ESEARCH by implementing a rigorous quality-control process throughout its construction, including,

- **Proficient Financial Experts.** The cohort of financial experts comprises over 30 professional practitioners, academic researchers, and graduate students in economics and related disciplines from leading institutions and universities. These experts are deeply involved in the entire benchmark construction process, from structure design to section verification and report review. To ensure structural rigor and practical applicability, a dedicated senior team comprising industry experts with over ten years of experience, finance professors, and postdoctoral

<sup>2</sup>Bloomberg:<https://www.bloomberg.com/>; Alpha Vantage:<https://www.alphavantage.co/>; Google News: <https://gnews.io/>; Yahoo Finance:<https://finance.yahoo.com/>

Table 3. Comparison between our FINDEEPRSEARCH and other deep research benchmarks.

Name	Domain	Structured	Languages	#Answers/Items	Retrieval Corpus
GAIA [10]	General	✗	EN	466	Web
BrowseComp [26]	General	✗	EN	1,266	Web
AssistantBench [31]	General	✗	EN	214	Web
ExpertLongBench [13]	General	✗	EN	Various	Offline
DeepResearchBench [2]	General	✗	EN,zh-CN	2,500	Web
ScholarSearch [36]	Academic	✗	EN	223	Offline
ResearchBench [9]	Academic	✗	EN	678	Web
FinSearchComp [6]	Finance	✗	EN, zh-CN	635	Web
FinResearchBench [14]	Finance	✗	EN	Various	Web
<b>FINDEEPRSEARCH</b>	Finance	✓	EN, zh-CN, zh-HK, BI	15,808	Web

researchers, is assembled to design the analytical structure. Subsequent to the generation of the reference report, the details of each report undergo verification by financial experts.

- **Cross-source Data Validation.** The acquisition of financial data for the benchmark construction is drawn from a multi-source framework, including proprietary financial databases, official corporate websites, and established financial portals with API services. To ensure the integrity and accuracy of critical data, such as financial tables, key financial ratios, stock prices, and market indices, a cross-source validation protocol is implemented. Under this protocol, an individual data point is incorporated into the dataset only if it is corroborated by a minimum of two independent sources. In instances where discrepancies arise, a manual review is conducted by financial experts to arbitrate and determine the final value. This cross-source validation approach mitigates the risk of systematic errors and inconsistencies, thereby safeguarding the high quality of the benchmark.
- **Rigorous Structure Guided.** The construction of the benchmark is guided by a comprehensive and rigorous analytical structure designed by financial experts. The clearly defined and unambiguous grading items within this structure facilitate high-quality result generation and systematic verification.
- **Two-rounds Expert Verification.** To ensure the high quality of the benchmark, financial experts conduct a two-round verification process that includes both section-based error correction and report-based consistency checks. This approach guarantees that each grading item is reviewed by minimum two financial experts.

### 3.3 Statistic and Analysis

As shown in Table 1, the FINDEEPRSEARCH dataset comprises 64 companies spanning 10 industries across 8 financial markets in 4 languages. Each company’s analysis is structured hierarchically into 6 major sections, further subdivided into 18 subsections. For quantitative assessment, 247 data points are selected from each analysis and incorporated into a scoring system totaling 350 marks. Consequently, the entire benchmark encompasses 15,808 individual grading items.

To ensure a fair evaluation, we select 8 companies from each financial market. For each market, Table 2 summarizes the character count statistics for the reference analytical report of each company and for the answers of each grading item.

### 3.4 Comparison with Other Benchmarks

Table 3 presents a comparative analysis of FINDEEPRSEARCH against existing deep research benchmarks, highlighting its key advantages. Our benchmark differentiates itself in three key aspects: 1) Whereas existing benchmarks do not require analytically structured outputs, FINDEEPRSEARCH mandates that responses adhere to a rigorous and predefined structure. 2) It offers superior multilingual coverage; while most related works are

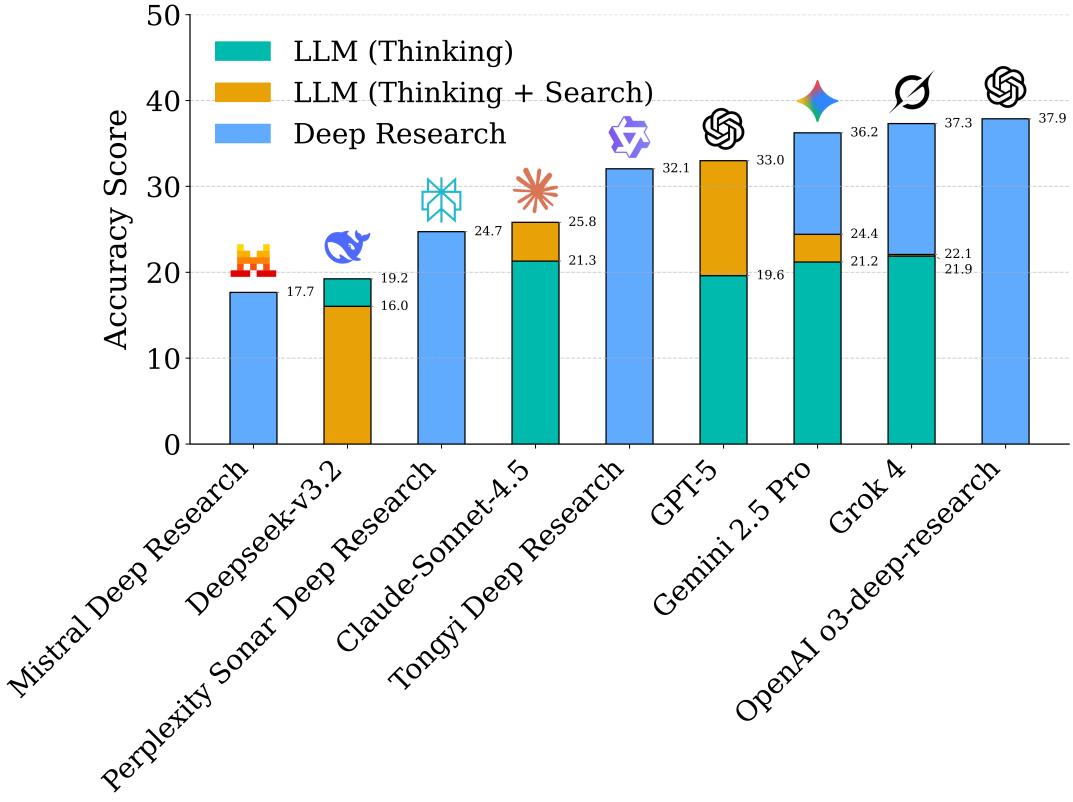


Fig. 3. An evaluation of representative methods on FINDEEPRSEARCH w.r.t *Information Precision*.

limited to English, and others like DeepResearchBench [2] and FinSearchComp [6] include only English and Simplified Chinese, our dataset encompasses four languages: English, Simplified Chinese, Traditional Chinese (Hong Kong), and Bahasa Indonesia. 3) With 15,808 data items for scoring, FINDEEPRSEARCH significantly surpasses the scale of prior benchmarks, enabling a more comprehensive and robust evaluation.

## 4 Experiments

In this section, we introduce the experimental setup and present the comprehensive analysis of the experimental results.

### 4.1 Compared Methods

With the proposed FINDEEPRSEARCH, we conduct experiments with 16 methods from 3 different groups. The selected methods exhibit considerable diversity in their underlying model families.

- **LLM with Thinking (T).** OpenAI GPT-5 (T) [20], Claude-Sonnet-4.5 (T) [16], Gemini 2.5 Pro (T) [18], Deepseek-v3.2 (T) [17], and Grok 4 (T) [28].
- **LLM with Thinking + Search (T+S).** OpenAI GPT-5 (T+S) [20], Claude-Sonnet-4.5 (T+S) [16], Gemini 2.5 Pro (T+S) [18], Deepseek-v3.2 (T+S) [17], and Grok 4 (T+S) [28].

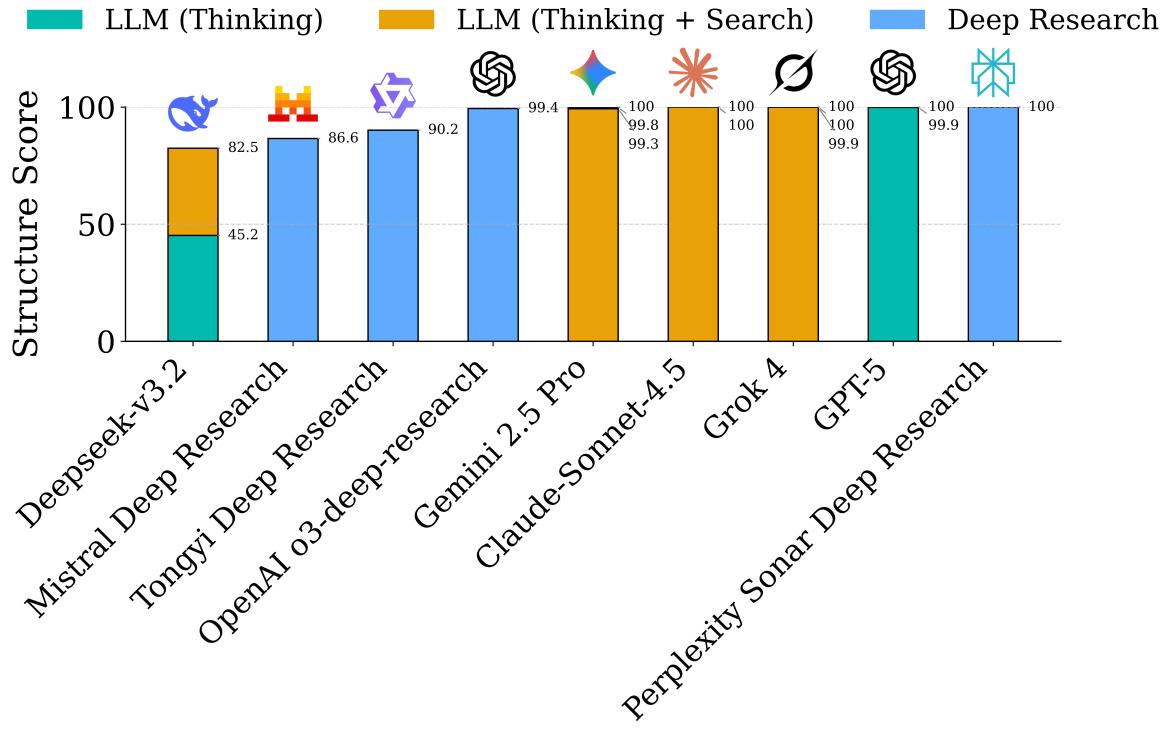


Fig. 4. An evaluation of representative methods on FINDEEPRESEARCH w.r.t *Structural Rigor*.

- **Deep Research.** OpenAI o3-deep-research [12], Gemini 2.5 Pro Deep Research [1], Grok 4 DeepSearch [28], Perplexity Sonar Deep Research [21], Tongyi Deep Research [24] and Mistral Deep Research [11].

## 4.2 Main Results

We present a comparable analysis of different methods on our proposed FINDEEPRESEARCH, assessing their performance in terms of *Information Precision* and *Structural Rigor*.

• **Information Precision.** Figure 3 illustrates a comparison of all methods with respect to *Information Precision*. We make the following key findings: 1) Among all methods, OpenAI’s o3-deep-research achieves the highest performance with an accuracy score of 37.9. It is closely followed by Grok-4 DeepSearch, which ranks a competitive second with a score of 37.3. 2) Deep Research (DR) methods generally perform better than the other two types. This is clearly seen in the top five results, where four are DR methods, and the remaining slot is held by OpenAI GPT-5 (T+S), their most advanced LLM. This result demonstrates the superiority of DR methods in solving high-standard and complex analysis tasks like our proposed FINDEEPRESEARCH. 3) LLMs relying solely on deep reasoning perform poorly, which underscores the necessity of search capability to retrieve external knowledge for effectively addressing the challenges in FINDEEPRESEARCH. 4) All methods face significant challenges in addressing FINDEEPRESEARCH, as evidenced by the highest score achieved being only 37.9 out of a maximum of 100, clearly indicating the persistent difficulty of the benchmark.

• **Structural Rigor.** Figure 4 shows the performance of all methods with respect to *Structural Rigor*. We make the following observations: 1) Most of the methods can produce the analytical results strictly following the

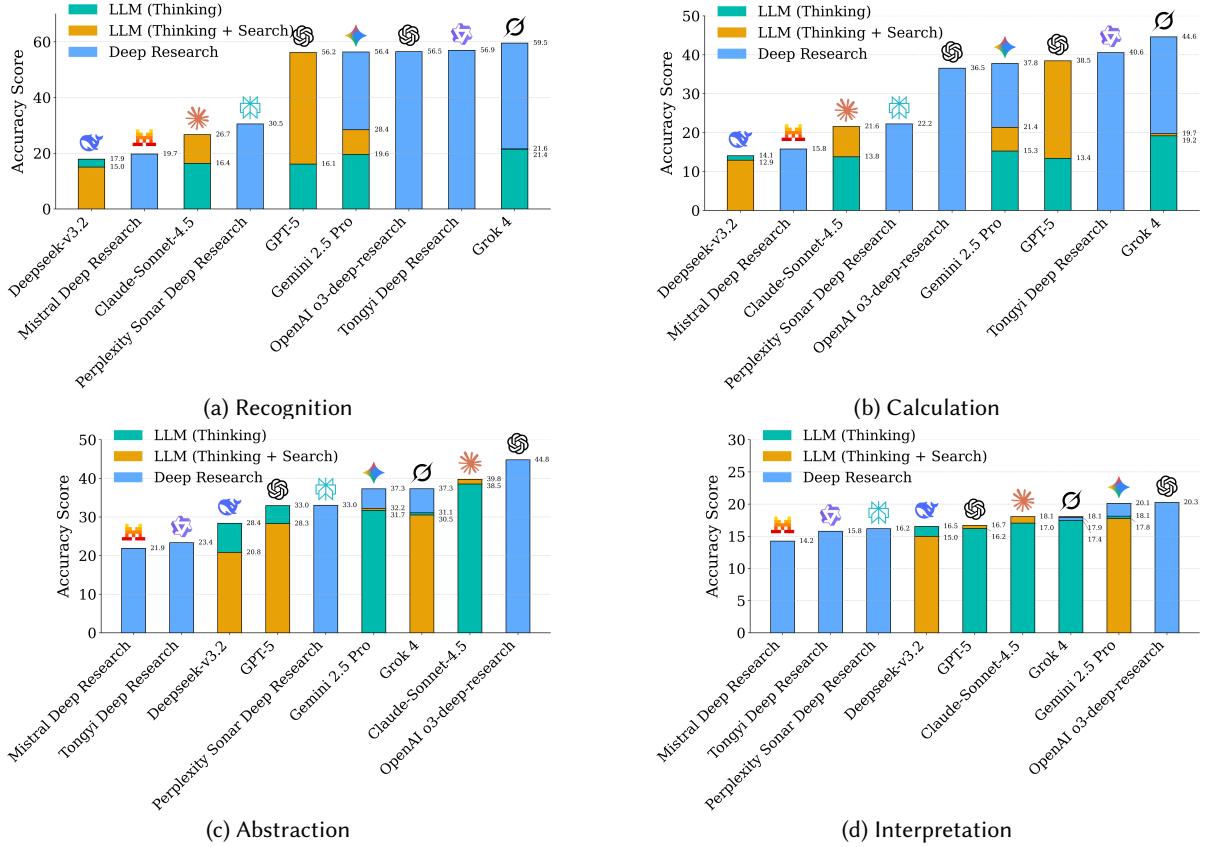


Fig. 5. Performance analysis across four different capabilities.

predefined hierarchical structure. Of all evaluated methods, 7 of them generate outputs with perfect formatting, and 5 of them contain only minor formatting errors. The findings suggest that advanced LLMs have developed the capability to follow complex instructions, such as the hierarchical structure in this study, which is fundamental for successfully executing rigorous research tasks. 2) From Figure 3 and Figure 4, we can observe that methods that perform poorly in structure following generally exhibit suboptimal results in generating accurate information. For instance, DeepSeek-v3.2 (T), DeepSeek-v3.2 (T+S), and Mistral Deep Research rank as the bottom three in *Structural Rigor*, and also show the weakest performance in *Information Precision*.

### 4.3 In-depth Analysis

- Performance Analysis Across Markets.** Table 4 reports the model performance in accuracy score across 8 financial markets and reveals the following findings: 1) Deep Research methods exhibit superior performance across all eight evaluated markets, significantly surpassing the results of “Thinking” and “Thinking + Search” approaches. For instance, OpenAI o3-deep-research leads five markets (*i.e.*, US, UK, CN, AU, ID), whereas Grok 4 DeepSearch secures the top rank in both HK and SG, and Gemini 2.5 Pro Deep Research achieves the highest performance in the MY market. Notably, Open-sourced Tongyi Deep Research demonstrates competitive performance against proprietary DR methods and outperforms most “Thinking” and “Thinking + Search” approaches.

Table 4. The performance analysis across 8 financial markets. The values reported in the table denote the normalized accuracy score. The best and second-best methods are indicated with **bold** and underline, respectively.

Method	US	UK	CN	HK	AU	SG	MY	ID
<b><i>LLM (Thinking)</i></b>								
Gemini 2.5 Pro (T)	19.9	21.0	17.6	20.8	24.4	24.2	25.1	16.5
Deepseek-v3.2 (T)	19.7	17.7	17.3	18.4	20.9	21.0	23.8	15.0
Claude-Sonnet-4.5 (T)	22.2	19.9	19.1	21.7	23.0	22.7	24.7	17.0
Grok 4 (T)	23.2	24.0	16.9	18.4	25.8	24.3	25.0	17.4
OpenAI GPT-5 (T)	18.1	18.7	16.6	17.6	22.6	23.6	23.3	16.3
<b><i>LLM (Thinking + Search)</i></b>								
Gemini 2.5 Pro (T+S)	22.9	20.7	20.4	24.7	26.4	27.6	27.5	20.9
Deepseek-v3.2 (T+S)	10.9	14.9	16.8	16.5	20.4	17.7	21.0	10.0
Claude-Sonnet-4.5 (T+S)	27.8	23.0	25.7	20.3	27.4	28.5	30.4	23.4
Grok 4 (T+S)	23.7	22.4	17.8	19.4	27.2	24.6	25.0	16.4
OpenAI GPT-5 (T+S)	37.4	36.9	20.8	29.3	35.6	<u>42.5</u>	32.3	29.1
<b><i>Deep Research</i></b>								
Perplexity Sonar Deep Research	21.0	23.7	22.4	25.0	28.8	26.9	26.9	23.0
Mistral Deep Research	13.5	16.1	14.0	13.6	22.2	21.1	23.7	17.1
Tongyi Deep Research	32.1	27.8	27.8	29.5	36.1	35.6	37.3	30.3
Gemini 2.5 Pro Deep Research	<u>37.6</u>	34.1	30.8	<u>36.0</u>	36.0	38.9	<b>39.8</b>	<u>36.6</u>
Grok 4 DeepSearch	34.5	39.0	<u>33.4</u>	<b>36.4</b>	39.3	<b>46.7</b>	37.9	31.3
OpenAI o3-deep-research	<b>42.5</b>	<b>43.0</b>	<b>34.7</b>	30.2	<b>41.7</b>	33.6	<u>38.3</u>	<b>38.9</b>

These findings collectively affirm the superiority of DR methods for addressing such high-standard and complex research analysis. 2) A significant performance gap exists across markets. Specifically, CN and HK present a tougher challenge, evidenced by their peak scores of only 34.7 by OpenAI o3-deep-research and 36.4 by Grok 4 DeepSearch, compared to easier markets like SG (46.7), UK (43.0), and US (42.5). This difficulty may stem from the increased complexity that methods encounter when processing languages other than the Latin languages, including Simplified Chinese and Traditional Chinese. 3) None of the markets are fully addressed by existing methods. Even in the best-performing SG market, Grok 4 DeepSearch achieves only 46.7 in the accuracy score. This sizable distance from the perfect score of 100 indicates substantial headroom for future advances in the field.

#### • Performance Analysis Across Capabilities.

Figure 5 presents a comparative performance analysis of the models across the four critical capabilities: Recognition, Calculation, Abstraction, and Interpretation. We make the following key findings: 1) Grok 4 DeepSearch ranks first for the Recognition and Calculation capabilities, while OpenAI o3-deep-research achieves the highest performance in the Abstraction and Interpretation capabilities. 2) DR methods generally outperform the methods in the other two categories for Recognition and Calculation capabilities. Comparably, the performance gap narrows significantly across all models for Abstraction and Interpretation. 3) Performance across the four capabilities demonstrates a clear difficulty spectrum, with Recognition being the most effectively addressed capability, achieving the highest score of 59.5. Calculation and Abstraction exhibit comparable performance, peaking at 44.6 and 44.8, respectively. Conversely, Interpretation is currently the most difficult, evidenced by its maximum score of only 20.3. This significant lag suggests that improving Interpretation capability should be a promising focus for future research. 4) All four capabilities remain far from a perfect score. The highest

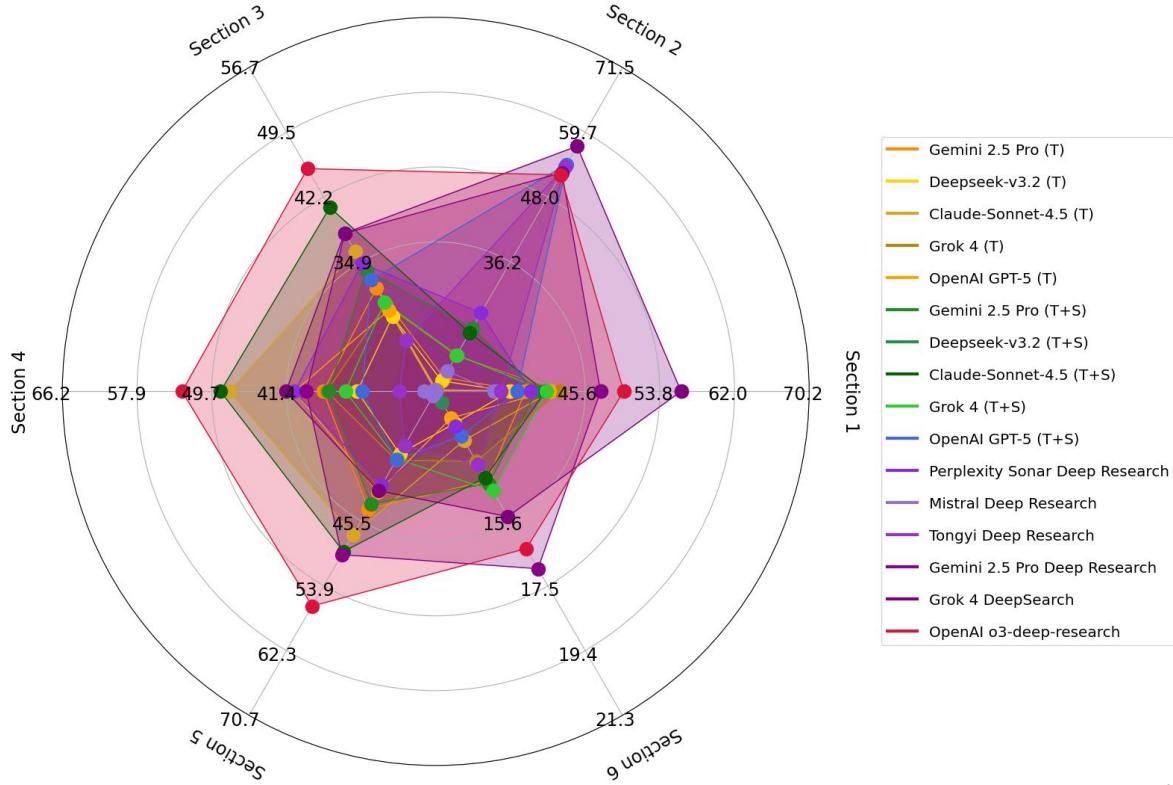


Fig. 6. Performance analysis across sections. To ensure comparability across different sections, the values reported represent the normalized accuracy score, capped at 100.

performance achieved, 59.5 by Grok 4 DeepSearch in Recognition, indicates that there is substantial room for overall improvement.

**• Performance Analysis Across Sections.** Figure 6 shows the performance analysis across different sections. We make the following observations: 1) Performance varies significantly across the sections. Sections 1, 2, 4, and 5 show moderate success with the best normalized accuracy score above 50, and Section 3 reaches 45.4, while Section 6 presents a severe challenge, with a maximum accuracy score of only 17.0. This low performance of Section 6 is attributed to the demands for integrated analysis of diverse inputs, including financial statements, stock prices, news, market indices, and currency exchange rates. 2) Deep Research methods exhibit clear advantages. Specifically, Grok-4 Deep Research achieves the highest scores in Sections 1 and 2, OpenAI o3 Deep Research attains the best performance in Sections 3, 4, and 5, while Gemini 2.5 Pro Deep Research leads in Section 6. This demonstrates the superior efficacy of the Deep Research methods in significantly enhancing report quality compared to the other two methodologies. 3) Across all sections, LLM (Thinking + Search) methods generally demonstrate superior performance over LLM (Thinking) methods, highlighting the importance of retrieval for this task. This superiority is most evident in sections 2 and 3.

#### 4.4 Case Study

We evaluate the three approaches, *i.e.*, "Thinking (T)", "Thinking + Search (T+S)", and "Deep Research (DR)", to characterize their performance boundaries. In Table 5, we use tick marks (✓) to denote good cases where metrics

Table 5. Case Study. “T”, “T+S” and “DR” represent “LLM (Thinking)”, “LLM (Thinking + Search)” and “Deep Research”.

Type	Grading Items	Example
T (✗) T+S(✓) DR (✓)	Income before Income Taxes Total Liabilities Shareholders Equity ...	Income before Income Taxes: 1,721 Millions USD OpenAI GPT-5 (T): N/A OpenAI GPT-5 (T+S): 1,721 in millions USD OpenAI o3-deep-research: 1,721 Millions USD
T (✗) T+S(✗) DR (✓)	Profitability and Earnings Quality Market Position Core Values ...	Profitability and Earnings Quality: Profitability declined in 2024, with a decrease in net profit margin and gross margin. The net profit also decreased significantly. OpenAI GPT-5 (T): - Profitability compressed due to raw-material inflation but remained ... earnings quality supported by cash conversion and low reliance on non-recurring items... margins trailed the prior year's levels. OpenAI GPT-5 (T+S): Margin compression ... Earnings quality supported by positive operating cash flow (US\$52.6m) despite headwinds. OpenAI o3-deep-research: Profitability slipped in 2024, with net income down 26% . Earnings quality remained decent (all profit derived from core operations), but margins were squeezed. Gross profit decline and higher expenses ...
T (✗) T+S(✗) DR (✗)	Annualized Volatility Log Excess Return Maximum Drawdown ...	Annualized Volatility: 17.40% OpenAI GPT-5 (T): N/A OpenAI GPT-5 (T+S): N/A OpenAI o3-deep-research: 20.10%

are generally accurately retrieved or calculated and cross marks (✗) to indicate bad cases where results are mostly inaccurate or unavailable. Results are organized into three approaches based on performance patterns:

- **T (✗), T+S(✓), DR (✓):** T produces bad results while T+S and DR produce good results. This performance gap arises because the T approach operates solely on the model’s parametric knowledge without access to external information sources, limiting its ability to accurately extract financial data from relevant data sources and subsequently compute metrics. In contrast, both T+S and DR methods leverage external retrieval capabilities to access requisite financial data.
- **T (✗), T+S(✗), DR (✓):** Both T and T+S produce bad results while DR produces good results. The superior performance of DR can be attributed to its agent-driven architecture, which facilitates iterative retrieval and reasoning capabilities. The iterative nature of DR allows for self-correction and refinement through multiple reasoning cycles, leading to higher quality outputs compared to the single-pass reasoning employed by T and T+S approaches.
- **T (✗), T+S(✗), DR (✗):** All three approaches yield unsatisfactory results. These shortcomings arise primarily in metrics that require historical adjusted closing price data over extended horizons (typically one year), which are difficult to obtain with precision. A further source of error stems from the demands for integrated analysis of diverse data sources, including financial statements, stock prices, news, market indices, and currency exchange rates.

## 5 Related Work

### 5.1 Deep Research Agents

LLMs have recently shown strong capabilities in reasoning and problem solving, motivating the development of DR agents that autonomously explore the web and generate research reports [4, 23, 32]. Among early agents, ReAct [30] is among the earliest to couple reasoning traces with environment actions, enabling interleaved reasoning-and-acting for open-ended tasks. Building on this idea, Search-R1 uses reinforcement learning to decide when and how to issue search queries for multi-hop question answering [8], and MMSearch-R1 extends this line by incorporating multimodal search for joint text–image reasoning [27]. While effective, these methods generally do not produce well-structured and comprehensive research reports.

To address this gap, recent DR agents integrate planning, multi-round retrieval, and evidence-grounded synthesis in a dynamic loop [19, 21, 29, 34]. For example, the Gemini 2.5 Pro Deep Research agent [1] plans

research, performs broad-coverage retrievals, and synthesizes a structured report end-to-end after reinforcement-learning-driven fine-tuning. OpenAI Deep Research [12] provides a ChatGPT-based workflow that interactively clarifies queries, browses the live web, analyzes retrieved content with built-in tools, and produces source-grounded, citation-rich summaries. Qwen Deep Research [22] employs dynamic research blueprinting and concurrent task orchestration to improve autonomous planning and adaptive execution. Despite these advances, rigorously evaluating the structured research outcomes generated by DR agents remains a major challenge, as there is still no consensus on how to measure both their structural completeness and information accuracy.

## 5.2 Benchmarks for Deep Research Agents

Benchmarking DR agents has become a critical avenue for assessing their ability to plan, retrieve, and synthesize evidence into structured research reports [25, 31]. General-purpose benchmarks typically evaluate agents on open-domain problems requiring long-horizon reasoning, factual grounding, and iterative synthesis [9, 13, 25, 26]. Among them, DeepResearch Bench [2] spans diverse academic disciplines and employs structured frameworks such as RACE and FACT to measure report comprehensiveness, instruction following, and citation fidelity; ExpertLongBench [13] targets expert-level long-form outputs through checklist-based rubrics.

Domain-specific benchmarks, in contrast, focus on emphasizing professional expertise, time sensitivity, and finer-grained evaluation [6, 14]. For instance, FinSearchComp [6] emphasizes financial analyst workflows: retrieving real-time market data, performing historical lookups, and conducting multi-period investigations with expert-annotated tasks and a rigorous multi-stage QA process. FinResearchBench [14], evaluates financial research agents by extracting logic trees from their reports and assessing performance across 70 expert-curated questions spanning 7 key task types. Yet even these domain-specific efforts remain limited: most confine evaluation to short-form answers or coarse global report scores and rarely assess full-length research reports in critical analysis scenarios for structural completeness, evidence reconciliation, and fine-grained factual accuracy. To the best of our knowledge, this work is the first to propose *HisRubric*, a novel evaluation framework and FINDEPRESEARCH benchmark for rigorously assessing deep research agents in financial analysis.

## 6 Conclusion

In this paper, we introduce *HisRubric*, a novel framework to evaluate the ability of DR agents to conduct high-quality and rigorous financial analysis, by defining and measuring the core qualities of *Structural Rigor* and *Information Precision*. We construct a new benchmark, FINDEPRESEARCH, covering 64 companies across 8 markets and 4 languages. Our experiments suggest that even top-performing DR agents struggle to consistently balance a coherent analytical structure with factual accuracy. This imbalance remains the primary barrier to their deployment in high-stakes applications. Future work can extend our framework to other domains, such as legal and clinical research, and explore how novel agent architectures might narrow this performance gap. In summary, we contend that a dual evaluation of rigor and precision is a crucial step towards building the next generation of reliable DR agents for professional, real-world tasks.

## 7 Contributions

- **Project Leader:** Fengbin Zhu, Chao Wang, and Tianhui Tan.
- **Major Contributors:** Xiang Yao Ng, Ziyang Liu, Chang Liu, Xianwei Zeng, Xuan Yao, and Min Xu.
- **Secondary Contributors:** Zixuan Wang, Pengyang Shao, Jing Wang, Xin Lin, Junfeng Li, Jingxian Zhu, and Yang Zhang.
- **Advisors:** Wenjie Wang, Fuli Feng, Richang Hong, Huanbo Luan, Ke-Wei Huang, and Tat-Seng Chua.

## References

- [1] Dave Citron. 2025. Deep Research is now available on Gemini 2.5 Pro Experimental. <https://blog.google/products/gemini/deep-research-gemini-2-5-pro-experimental/> (2025).
- [2] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv preprint arXiv:2506.11763* (2025).
- [3] Lisa Milici Gaynor, Andrea Seaton Kelton, Molly Mercer, and Teri Lombardi Yohn. 2016. Understanding the relation between financial reporting quality and audit quality. *Auditing: A Journal of Practice & Theory* 35, 4 (2016), 1–22. doi:10.2308/ajpt-51453
- [4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
- [5] Sariyama Herath and Norah Albarqi. 2017. Financial reporting quality: A literature review. *Journal of Business Management and Commerce* 2 (2017), 1–14.
- [6] Liang Hu, Jianpeng Jiao, Jiashuo Liu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, Fei Hu, et al. 2025. FinSearchComp: Towards a Realistic, Expert-Level Evaluation of Financial Search and Reasoning. *arXiv preprint arXiv:2509.13160* (2025).
- [7] Jeffrey Ip and Kritin Vongthongsri. 2025. *deepeval*. <https://github.com/confident-ai/deepeval>
- [8] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. In *Second Conference on Language Modeling*.
- [9] Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248* (2025).
- [10] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.
- [11] Mistral Team. 2025. TLe Chat dives deep. <https://mistral.ai/news/le-chat-dives-deep>. Accessed: 2025-10-07.
- [12] OpenAI Team. 2025. Introducing deep research. <https://openai.com/index/introducing-deep-research/>. Accessed: 2025-10-07.
- [13] Jie Ruan, Inderjeet Nair, Shuyang Cao, Amy Liu, Sheza Munir, Micah Pollens-Dempsey, Tiffany Chiang, Lucy Kates, Nicholas David, Sihan Chen, et al. 2025. ExpertLongBench: Benchmarking Language Models on Expert-Level Long-Form Generation Tasks with Structured Checklists. *arXiv preprint arXiv:2506.01241* (2025).
- [14] Rui Sun, Zuo Bai, Wentao Zhang, Yuxiang Zhang, Li Zhao, Shan Sun, and Zhengwen Qiu. 2025. FinResearchBench: A Logic Tree based Agent-as-a-Judge Evaluation Framework for Financial Research Agents. *arXiv preprint arXiv:2507.16248* (2025).
- [15] Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025. AI-Researcher: Autonomous Scientific Innovation. *arXiv preprint arXiv:2505.18705* (2025).
- [16] Claude Team. 2025. Introducing Claude Sonnet 4.5. <https://www.anthropic.com/news/clause-sonnet-4-5>. Accessed: 2025-10-07.
- [17] DeepSeek Team. 2025. Introducing DeepSeek-V3.2-Exp. <https://api-docs.deepseek.com/news/news250929>. Accessed: 2025-10-07.
- [18] Gemini Team. 2025. Gemini 2.5 Pro. <https://deepmind.google/models/gemini/pro/>. Accessed: 2025-10-07.
- [19] Kimi Team. 2025. Kimi-Researcher: End-to-End RL Training for Emerging Agentic Capabilities. [https://moonshotai.github.io/Kimi-Researcher/?utm\\_source=chatgpt.com](https://moonshotai.github.io/Kimi-Researcher/?utm_source=chatgpt.com).
- [20] OpenAI Team. 2025. Introducing GPT-5. <https://openai.com/index/introducing-gpt-5/>. Accessed: 2025-10-07.
- [21] Perplexity Team. 2025. Introducing perplexity deep research. <https://www.perplexity.ai/ja/hub/blog/introducing-perplexity-deep-research>.
- [22] Qwen Team. 2025. Deep research (Qwen-Deep-Research). <https://www.alibabacloud.com/help/en/model-studio/qwen-deep-research>.
- [23] Qwen Team. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning. <https://qwenlm.github.io/blog/qwq-32b/>
- [24] Tongyi Team. 2025. Tongyi DeepResearch: A New Era of Open-Source AI Researchers. <https://tongyi-agent.github.io/blog/introducing-tongyi-deep-research/>. Accessed: 2025-10-07.
- [25] Haiyuan Wan, Chen Yang, Junchi Yu, Meiqi Tu, Jiaxuan Lu, Di Yu, Jianbao Cao, Ben Gao, Jiaqing Xie, Aoran Wang, et al. 2025. DeepResearch Arena: The First Exam of LLMs' Research Abilities via Seminar-Grounded Tasks. *arXiv preprint arXiv:2509.01396* (2025).
- [26] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516* (2025).
- [27] Jinning Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. 2025. MMSearch-R1: Incentivizing LMMs to Search. *arXiv preprint arXiv:2506.20670* (2025).
- [28] xAI Team. 2025. Grok 4. <https://x.ai/news/grok-4>. Accessed: 2025-10-07.
- [29] xAI Team. 2025. Introducing Grok DeepSearch. <https://x.ai/news/grok-3>. Accessed: 2025-04-06.
- [30] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

- [31] Ori Yoran, Samuel Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. 2024. AssistantBench: Can Web Agents Solve Realistic and Time-Consuming Tasks?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 8938–8968.
- [32] Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471* (2025).
- [33] Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2025. Deep Research: A Survey of Autonomous Research Agents. *arXiv:2508.12752 [cs.IR]* <https://arxiv.org/abs/2508.12752>
- [34] Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, et al. 2025. From Web Search towards Agentic Deep Research: Incentivizing Search with Reasoning Agents. *arXiv preprint arXiv:2506.18959* (2025).
- [35] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474* (2023).
- [36] Junting Zhou, Wang Li, Yiyuan Liao, Nengyuan Zhang, Tingjia Miao, Zhihui Qi, Yuhua Wu, and Tong Yang. 2025. ScholarSearch: Benchmarking Scholar Searching Ability of LLMs. *arXiv:2506.13784 [cs.IR]* <https://arxiv.org/abs/2506.13784>

## A Industry Distribution

Table 6 reports the cross-market industry composition. The largest sectors are Communications (12), Consumer Staples (10), Energy (10), and Industrials (10). The remaining sectors are Consumer Discretionary (7), Health Care (6), Real Estate (3), Utilities (3), Technology (2), and Materials (1). Entries denote the number of companies in each industry–market cell; row totals are sector sizes and column totals sum to eight companies per market.

Table 6. The company distribution with varying industries

Industry	US	UK	CN	HK	AU	SG	MY	ID
	🇺🇸	🇬🇧	🇨🇳	🇭🇰	🇦🇺	🇸🇬	🇲🇾	🇮🇩
Communications	0	2	2	2	2	0	2	2
Consumer Discretionary	3	0	4	0	0	0	0	0
Consumer Staples	0	1	1	0	1	2	2	3
Energy	2	3	0	3	0	0	0	2
Health Care	2	0	0	0	2	2	0	0
Industrials	0	2	1	2	0	3	2	0
Materials	0	0	0	0	1	0	0	0
Real Estate	0	0	0	1	0	1	0	1
Technology	1	0	0	0	1	0	0	0
Utilities	0	0	0	0	1	0	2	0

## B Implementation Details

Table 7 lists the method configurations evaluated in our benchmark. Any settings not listed were left at their default values for the corresponding method. No public API is available for Mistral Deep Research, Gemini 2.5 Pro Deep Research, or Grok 4 Deep Research; for these methods, we collected results via their official web interfaces between September 29 and October 3, 2025.

## C Hierarchical Structure

This section documents the hierarchical design of our markdown-based research specification. To standardize the output, we create a comprehensive template, the full structure of which is shown in Figure 7. To guide generative models in populating this template, we develop a primary prompt, shown in Figure 8, which governs the overall research workflow and output constraints for all six sections.

This main prompt integrates detailed specifications for each section. For example, Figure 9 details the specific schema and rules for Section 1 (Company Overview). Together, this structured template and detailed prompting strategy ensure reproducibility, comparability across periods, and strict conformance to the required hierarchical output. The prompt is also adaptable; for models lacking native search capabilities (e.g., the “Thinking” method), we make minor modifications to accommodate their behavior.

Table 7. Correspondence between benchmark aliases, API identifiers, and API settings

Benchmark Alias	API Identifier	API Setting
<b><i>LLM (Thinking)</i></b>		
Gemini 2.5 Pro (T)	gemini-2.5-pro-preview-05-06	thinking_budget=-1
Deepseek-v3.2 (T)	deepseek-v3.2-exp	reasoning.enabled=True
Claude-Sonnet-4.5 (T)	claude-sonnet-4-5-20250929	thinking.type=enabled,thinking.budget_token_s=10000
Grok 4 (T)	grok-4-0709	all defaults
OpenAI GPT-5 (T)	gpt-5-2025-08-07	reasoning.effort=high
<b><i>LLM (Thinking + Search)</i></b>		
Gemini 2.5 Pro (T+S)	gemini-2.5-pro-preview-05-06	thinking_budget=-1,tools=[google_search]
Deepseek-v3.2 (T+S)	deepseek-v3.2-exp	reasoning.enabled=True,plugins=[exa(max_results=8)]
Claude-Sonnet-4.5 (T+S)	claude-sonnet-4-5-20250929	thinking.type=enabled,thinking.budget_token_s=10000,tools=[web_search_20250305]
Grok 4 (T+S)	grok-4-0709	search_parameters.mode=on
OpenAI GPT-5 (T+S)	gpt-5-2025-08-07	reasoning.effort=medium,tools=[web_search]
<b><i>Deep Research</i></b>		
Perplexity Sonar Deep Research	sonar-deep-research	reasoning.effort=high
Tongyi Deep Research	tongyi-deepresearch-30b-a3b	temperature=0.6,top_p=0.95,presence_penalty=1.1
OpenAI o3-deep-research	o3-deep-research-2025-06-26	tools=[web_search_preview,code_interpreter]

```

## Section 1: Company Overview
### S1.1: Basic Information
| Field | Value |
| :-- | :-- |
| Company Name | |
| Establishment Date | |
| Headquarters Location | |
### S1.2: Core Competencies
| Perspective | {FY} | {FY_1} |
| :-- | :-- | :-- |
| Innovation/Product Advantages | |
| Brand Recognition | |
| Reputation Ratings | |
### S1.3: Mission & Vision
| Field | Value |
| :-- | :-- |
| Mission/Vision Statement | |
| Core Values | |

## Section 2: Financial Performance
### S2.1: Income Statement
| Field | {FY} | {FY_1} | {FY_2} | Multiplier | Currency |
| :-- | :-- | :-- | :-- | :-- | :-- |
| Revenue | | | | |
| Cost of Goods Sold | | | | |
| Gross Profit | | | | |
| Operating Expenses/Income | | | | |
| Net Profit | | | | |
| Income before income taxes | | | | |
| Income tax expense (benefit)| | | | |
| Interest Expense | | | | |
### S2.2: Balance Sheet
| Field | {FY} | {FY_1} | {FY_2} | Multiplier | Currency |
| :-- | :-- | :-- | :-- | :-- | :-- |
| Total/Current/Non-Current Assets | | | | |
| Total/Current/Non-Current Liabilities | | | | |
| Shareholders' Equity | | | | |
| Retained Earnings | | | | |
| Total Equity and Liabilities | | | | |
| Inventories | | | | |
| Prepaid Expenses | | | | |
### S2.3: Cash Flow Statement
| Field | {FY} | {FY_1} | {FY_2} | Multiplier | Currency |
| :-- | :-- | :-- | :-- | :-- | :-- |
| Net Cash Flow from Operations/Investing/Financing | | | | |

```

```

| Net Increase/Decrease in Cash | | | | |
| Dividends | | | | |
### S2.4: Key Financial Metrics
| Field | {FY} | {FY_1} | {FY_2} |
| :-- | :-- | :-- | :-- |
| Gross/Operating/Net Profit Margin | | | |
| Current/Quick Ratio | | | |
| Debt-to-Equity | | | |
| Interest Coverage | | | |
| Asset Turnover | | | |
| Return on Equity/Assets | | | |
| Effective Tax Rate | | | |
| Dividend Payout Ratio | | | |
### S2.5: Operating Performance
| Field | {FY} | {FY_1} | {FY_2} |
| :-- | :-- | :-- | :-- |
| Revenue by Product/Service | | | |
| Revenue by Geographic Region | | | |

## Section 3: Business Analysis
### S3.1: Profitability Analysis
| Perspective | Answer |
| :-- | :-- |
| Revenue & Direct-Cost Dynamics | |
| Operating Efficiency | |
| External & One-Off Impact | |
### S3.2: Financial Performance Summary
| Perspective | {FY} | {FY_1} |
| :-- | :-- | :-- |
| Comprehensive Financial Health | | |
| Profitability and Earnings Quality | | |
| Operational Efficiency | | |
| Risk Identification and Early Warning| | |
| Future Financial Performance Projection | | |
### S3.3: Business Competitiveness
| Perspective | {FY} | {FY_1} |
| :-- | :-- | :-- |
| Business Model | | |
| Market Position | | |

## Section 4: Risk Factors
### S4.1: Risk Factors
| Perspective | {FY} | {FY_1} |
| :-- | :-- | :-- |
| Market/Operational/Financial/Compliance Risks| | |

```

```

## Section 5: Corporate Governance
### S5.1: Board Composition
| Name | Position | Total Income |
| :-- | :-- | :-- |
| | | |
### S5.2: Internal Controls
| Perspective | {FY} | {FY_1} |
| :-- | :-- | :-- |
| Risk Assessment Procedures | | |
| Control Activities | | |
| Monitoring Mechanisms | | |
| Identified Material Weaknesses/Deficiencies | | |
| Effectiveness | | |

## Section 6: Market Performance
### S6.1: Stock Performance
| Field | {CY} | {CY_1} |
| :-- | ----:| ----:|
| Lowest/Highest Adjusted Closing Price | | |
| Total Log Return | | |
| Log Excess Return | | |
| Maximum Drawdown | | |
| Annualized Volatility | | |
### S6.2: News Sentiment Analysis
| Field | {CY} | {CY_1} |
| :-- | :-- | :-- |
| Top 1/2/3 Positive Window Date/Summary | | |
| Top 1/2/3 Negative Window Date/Summary | | |
### S6.3: Market Reaction to News
| Field | {CY} | {CY_1} |
| :-- | :-- | :-- |
| Top 1/2/3 Positive Window Date/CAR/Summary | | |
| Top 1/2/3 Negative Window Date/CAR/Summary | | |
### S6.4: Price-to-Earnings (P/E) Ratio
| Field | Value as of {DATE} |
| :-- | :-- |
| Adjusted Closing Price | |
| Diluted EPS & P/E Ratio | |

```

Fig. 7. Complete hierarchical structure for 6 main sections, 18 subsections and 18 markdown tables

Search and analyze the annual reports({LANGUAGE} version) from FY{FY\_1}{FY\_A} and FY{FY} for {COMPANY} listed on {STOCK\_MARKET}. Assuming today's date is 2025-09-20, generate a research report on this company. The report should be in markdown format and must follow the given structure below and include all the tables below.

As defined in the `Research Scope` of each section, annual reports are sufficient for conducting analyses from Section 1 to Section 5. However, to complete Section 6, you should diligently search for additional information such as news, stock prices, or any relevant data that can support your research. Further requirements can be derived from the `Research Scope` of each section.

`Research Language`: Please make sure your answer is written in {LANGUAGE}. Note that always leave the section headers and table headers in English.

`Research Output Format`: Your response must consist ONLY of the section and subsection headers and the completed markdown tables as defined below. No text outside of tables is permitted. All analyses and summaries must be written within table cells, using detailed content (typically 4-8 sentences or structured bullet points).

```
## Section 1: Company Overview
*[Details and table structure omitted for brevity]*
## Section 2: Financial Performance
*[Details and table structure omitted for brevity]*
## Section 3: Business Analysis
*[Details and table structure omitted for brevity]*
## Section 4: Risk Factors
*[Details and table structure omitted for brevity]*
## Section 5: Corporate Governance
*[Details and table structure omitted for brevity]*
## Section 6: Market Performance
```

*[Details and table structure omitted for brevity]*
---

Fig. 8. Research Task Prompt for Sections 1–6

```

## Section 1: Company Overview
This section provides a concise overview of the company, including its basic information, industry background, key strengths, and strategic direction.

`Research Scope`: Focus on the FY{FY} annual report for FY{FY} data, and the FY{FY_1} annual report for FY{FY_1} data for {COMPANY} listed on {STOCK_MARKET}.

### S1.1: Basic Information
This subsection provides fundamental information about the company's identity.

Create a table with Markdown format with Field and Value headers for the following items:

1. Company Name
2. Establishment Date
3. Headquarters Location (City and Country)

| Field | Value |
| :---- | :---- |
| Company Name | |
| Establishment Date | |
| Headquarters Location | |

### S1.2: Core Competencies
This section provides information about the company's core competencies. Create a summary in the table below for each perspective, offering readers insight into the company's competitive strengths and unique value propositions.

| Perspective | {FY} | {FY_1} |
| :---- | :---- | :---- |
| Innovation Advantages | | |
| Product Advantages | | |
| Brand Recognition | | |
| Reputation Ratings | | |

### S1.3: Mission & Vision

```

This section provides information about the company's purpose and long-term goals. Create a summary in the table below for each perspective in the single cell, offering readers a clear understanding of the company's strategic direction and aspirations.

Field	Value
-----	-----
Mission Statement	
Vision Statement	
Core Values	

Fig. 9. Section 1 specification: Company Overview—scope, subsections, and table schemas