FinFIRST: Benchmarking Search Agents for Financial Information Retrieval, Sourcing and Traceability

Wenqing Wang[1,*], **Haitao Xiang**[1,*], **Xinyi Zhao**[1,2,◇], **Mingming Yin**[1], **Ying Zhong**[1], **Zhaoxin Huan**[1], **Qiheng Zhou**[3], **Jin Zhu**[1], **Xiaolu Zhang**[1,†], **Shi Chang**[1], **Jun Zhou**[1]

[1]Ling Team, Inclusion AI, [2]Peking University, [3]China International Capital Corporation Limited

Financial search is a highly demanding task for LLM agents, requiring not only a correct final answer but also temporally valid information retrieval, authoritative source selection, entity and period alignment, unit and definition consistency, and verifiable evidence for all conclusions. Existing benchmarks predominantly evaluate only the final answer, making it difficult to localize errors or assess whether an answer is well-founded. To address this gap, we introduce **FINFIRST** (**F**inancial **I**nformation **R**etrieval, **S**ourcing and **T**raceability), the first financial benchmark to jointly evaluate answers and supporting evidence through atomic rubrics. FINFIRST comprises 123 expert-authored tasks spanning a graduated difficulty spectrum, constructed from aggregate patterns of real-world financial scenarios via an 18-field taxonomy, a 6-axis coverage blueprint, a registry of 138 financial sources, contributions from over 50 finance experts, and 6-stage quality control pipeline. Each task is accompanied by an evidence-grounded reference package decomposed into atomic criteria across three dimensions—raw-information acquisition, source verification, and computation and answer formation—enabling fine-grained evaluation of the complete research process. We evaluate 15 model configurations under a unified tool setting: Claude-Opus-5 achieves the highest atomic score (87.59%), while GPT-5.6-Sol attains the highest strict pass rate (71.54%), with computation and answer formation consistently lagging behind raw-information acquisition as a common bottleneck across all systems. FINFIRST thus retains final-answer correctness as the primary objective while making the supporting research process measurable, verifiable, and diagnosable [a].

Date: September, 2026

Huggingface: https://huggingface.co/datasets/inclusionAI/FinFIRST

[a]This paper presents FINFIRST V1; the forthcoming V2 will expand the benchmark further, with tasks spanning foundational search and advanced financial research.

ANT GROUP

1 Introduction

Search has emerged as a core capability of Large Language Model (LLM) agents (Wei et al., 2025; Gou et al., 2025), and finance provides a particularly demanding setting for this capability (Hu et al., 2025; Bigeard et al., 2025). A successful search in finance rarely means locating a single webpage or matching a number. Instead, an agent must identify authoritative and time-valid sources, retrieve information for the correct entity and reporting period, align units and financial

*Equal contribution. †Corresponding author: <yueyin.zxl@antgroup.com>. ◇Work done during an internship at Ling Team, Inclusion AI.

Table 1 Comparison of representative financial benchmarks. **Agentic**: requires agents to actively retrieve information or use external tools. **Time-sensitive**: contains tasks whose correct answers are tied to a specific point in time and may change as new information becomes available. **Demand-aligned**: ensures tasks reflect authentic financial needs or real-world professional workflows. **Expert-authored**: requires tasks to be designed and authored by domain experts in finance. **Multi-source**: requires integrating and reconciling evidence from multiple heterogeneous sources. **Evidence scoring**: evaluates the quality and relevance of the supporting sources and evidence. **Atomic assessment**: assigns separate scores to individual intermediate criteria rather than relying solely on final-answer evaluation.

Benchmark	Agentic	Time-sensitive	Demand-aligned	Expert-authored	Multi-source	Evidence-scoring	Atomic-assessment
ConvFinQA (Chen et al., 2022)	✗	✗	✗	✓	✗	✗	✗
FinanceBench (Islam et al., 2023)	✗	✗	✗	✓	✗	✗	✗
FinanceQA (Mateega et al., 2025)	✗	✗	✗	✓	✗	✗	✗
BizFinBench (Lu et al., 2025)	✗	✗	✓	✓	✗	✗	✗
Finance Agent Benchmark (Bigeard et al., 2025)	✓	✗	✗	✓	✓	✗	✓
FinSearchComp (Hu et al., 2025)	✓	✓	✗	✓	✓	✗	✗
FrontierFinance (research) (Zhang et al., 2026)	✓	✓	✗	✓	✓	✓	✗
FrontierFinance (computer use) (Krumdick et al., 2026)	✓	✗	✗	✓	✓	✗	✗
FinDeepIndicator (Yang et al., 2026)	✓	✗	✗	✗	✓	✗	✓
FinFIRST (ours)	✓	✓	✓	✓	✓	✓	✓

definitions, reconcile evidence across sources, perform necessary calculations, and support its conclusion with verifiable evidence. Errors at any of these steps can yield a plausible-looking but unsupported answer. Evaluating financial research agents therefore requires examining not only the final answer, but also the evidence and intermediate steps used to derive it.

However, existing benchmarks cover complementary parts of this workflow. Financial QA datasets such as FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022), TAT-QA (Zhu et al., 2021), and FinanceBench (Islam et al., 2023) provide the relevant evidence context, and therefore primarily evaluate reasoning over evidence rather than the end-to-end process that discovers it.

Finance Agent Benchmark (Bigeard et al., 2025) and FrontierFinance (Zhang et al., 2026) extend evaluation to tool use and research-style outputs, but do not directly assess supporting evidence through atomic intermediate criteria. FinDeepIndicator (Yang et al., 2026) evaluates intermediate stages, but focuses on template-derived financial indicator construction. Among these agentic financial benchmarks, FinSearchComp (Hu et al., 2025) is most closely related to our work: it comprises 635 expert-curated questions covering time-sensitive data retrieval, historical lookup, and complex investigation across global and Greater China markets. Nevertheless, its binary final-answer evaluation reveals neither where an error occurs—in source selection, data extraction, definition alignment, or calculation—nor whether a correct response is genuinely supported by appropriate evidence or merely happens to match the reference answer. Table 1 summarizes these differences.

In this work, we introduce **FinFIRST** (Financial Information Retrieval, Sourcing and Traceability), the first financial benchmark to evaluate both generated answers and retrieved evidence through atomically decomposed rubrics. The rubrics span three capability groups: *(1) Raw-information acquisition* examines whether the required facts and figures are retrieved with the correct entity, period, unit, and definition; *(2) Source verification* assesses whether the cited sources are authoritative, temporally valid, and preferably first-party; and *(3) Computation and answer formation* evaluates subsequent calculations, reasoning, completeness, and instruction following. Traceability is enforced throughout by requiring key facts, figures, intermediate calculations, and conclusions to be explicitly grounded in and traceable to the cited evidence. The construction of FinFIRST benefited from professional support provided by the investment banking team at China International Capital

Table 2 An example combining Meta's official disclosures with an industry report, and its atomic evaluation rubric.

Question. Using the latest official and industry data available as of June 30, 2026, and Meta's official disclosures together with the IAB/PwC Internet Advertising Revenue Report, calculate Meta's 2025 United States & Canada advertising revenue as a percentage of 2025 U.S. social media advertising revenue. Report the result as a percentage rounded to two decimal places.

Reference answer. Meta reported quarterly U.S. & Canada advertising revenue of \$18.259, \$20.045, \$21.331, and \$25.643 billion, totaling \$85.278 billion. The IAB/PwC report gives 2025 U.S. social media advertising revenue of \$117.70 billion. Therefore, $\$85.278/\$117.70 \times 100\% = \mathbf{72.45}\%$.

Capability	Atomic criterion	Points
Source verification	Identifies Meta's official 2025 disclosure materials or filing.	10
Source verification	Identifies the 2025 IAB/PwC Internet Advertising Revenue Report.	10
Raw-information acquisition	Extracts Meta's four quarterly U.S. & Canada advertising-revenue figures: \$18.259, \$20.045, \$21.331, and \$25.643 billion.	30
Raw-information acquisition	Extracts 2025 U.S. social media advertising revenue of \$117.70 billion.	20
Computation and answer formation	Correctly aggregates Meta's quarterly revenue: $18.259 + 20.045 + 21.331 + 25.643 = 85.278$ billion.	10
Computation and answer formation	Correctly calculates: $85.278/117.70 \times 100\% = 72.45\%$.	10
Computation and answer formation	Reports the final answer as 72.45%, with the required precision and unit.	10
Total		**100**

Corporation Limited (CICC), lending strong industry-grounded expertise to the development process.

Specifically, FINFIRST is constructed through the three-phase pipeline. First, we derive an 18-field taxonomy and a 6-axis coverage blueprint from aggregate patterns in real financial search demand. In parallel, finance experts compile a routing map of 138 sources, annotating each for authority, task suitability, and geographic coverage. Second, over 50 finance experts author questions guided by the coverage blueprint, with continuous rebalancing to fill under-represented categories. Guided by the source map, they construct a reference package for each question, including the answer, supporting sources, key definitions and data versions, calculations, valid alternatives, and numerical tolerances. Each package is then decomposed into atomic rubric items across *raw-information acquisition*, *source verification*, and *computation and answer formation*. Third, every candidate undergoes 6-stage quality control: scope review, independent re-solving, cross-validation and adjudication, rubric auditing, model-based stress testing, and final consistency checks. This pipeline yields 123 realistic, diverse, and expert-vetted tasks with graduated difficulty, ranging from focused single-source lookups to complex multi-step research requiring planning, cross-source synthesis, and calculation. Table 2 presents an advanced-research example alongside its atomic rubric.

We evaluate 15 model configurations under a unified toolset comprising web search, webpage fetching, and Python execution, employing atomic scores and strict pass rates to distinguish partial progress from complete task success. The results yield three main findings. First, even the strongest agents remain far from perfect: Claude-Opus-5 achieves the highest overall rubric score (87.59%), while GPT-5.6-Sol attains the highest strict pass rate (73.54%). The gap between these two metrics suggests that high aggregate scores often reflect partially correct research processes rather than consistently complete solutions. Second, *computation and answer formation* consistently lag behind *raw-information acquisition* across all evaluated systems, indicating that successfully retrieving

relevant data does not guarantee correct calculation, synthesis, or reporting. Third, similar overall scores can mask distinct capability profiles. For instance, GLM-5.3 and Kimi-K3 achieve nearly identical overall scores (80.61% vs. 80.83%), yet GLM-5.3 substantially outperforms Kimi-K3 in source verification (78.56% vs. 70.70%), whereas Kimi-K3 holds an advantage in raw-information acquisition and computation and answer formation. Together, these findings demonstrate the diagnostic value of atomic assessment in pinpointing where financial research agents succeed and where they fall short.

Our contributions are threefold:

- We introduce and release FINFIRST, a demand-aligned, expert-authored benchmark of 123 tasks with graduated difficulty, accompanied by atomic rubrics that evaluate both answers and evidence for fine-grained assessment.
- We develop an expert-led construction pipeline grounded in real-world financial scenarios, combining an 18-field taxonomy, a 6-axis coverage blueprint, 138 financial sources, more than 50 finance experts, and 6-stage quality control.
- We systematically evaluate 15 model configurations under a unified basic-tool setting, revealing a gap between atomic scores and strict task completion, a consistent bottleneck in computation and answer formation, and distinct capability profiles hidden by similar aggregate scores.

2 Data Construction

The construction of FINFIRST follows four principles: benchmark composition should reflect financial information needs observed in practice; task difficulty should arise from meaningful retrieval, source selection, definition alignment, evidence reconciliation, and computation rather than obscure facts or artificial traps; each reference answer should be reproducible from authoritative, version-correct evidence; and evaluation should reveal where an agent succeeds or fails within the research process. As illustrated in Figure 1, we operationalize these principles through three phases: data preprocessing, expert-led data construction, and six-stage quality control.

2.1 Data Preprocessing

2.1.1 Demand Analysis and Taxonomy Design

We abstract recurring financial information needs into an 18-field taxonomy, organized across four semantic groups. **Task structure** captures question decomposition, subquestion dependencies, and conditional filtering. **Research content** covers the research object, target metric, metric definition or accounting convention, geography or market, temporal-constraint type, and explicit time specification. **Processing requirements** specify the verification mode, computation type, output format, and response language. **Source and retrieval intent** records whether a source is specified, its accessibility and mainstreamness, the number of sources required, and the retrieval intent—data retrieval, fact retrieval, evidence localization, or concept identification. Notably, these four groups constitute an organizational summary rather than a rigid four-tuple: every question is annotated over a uniform 18-field schema, with individual fields may be multi-valued or inapplicable. This representation supports stratified sampling, coverage auditing, and fine-grained capability analysis. Aggregate patterns from real-world financial-use scenarios determine the baseline distribution, while professionally important but less frequent settings are retained when they can be supported by accessible and reproducible evidence.

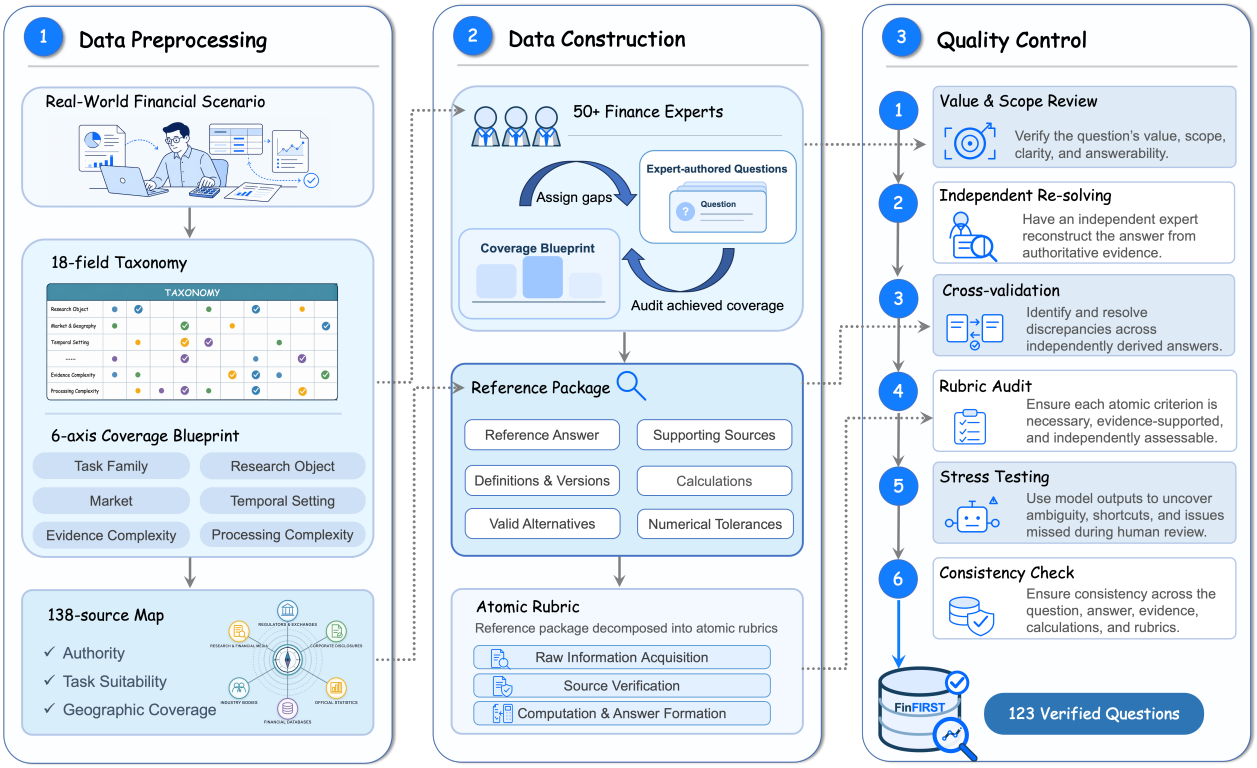

Figure 1 Overview of FINFIRST construction pipeline. The pipeline comprises three stages: data preprocessing derives an 18-field taxonomy, a 6-axis coverage blueprint, and a 138-source map from patterns observed in real-world financial scenarios; data construction engages over 50 finance experts to iteratively author questions, balance coverage, construct evidence-grounded reference packages, and define atomic rubrics; and quality control applies a 6-stage review and filtering procedure, resulting in a final set of 123 verified questions.

2.1.2 Coverage Blueprint

Taking the Cartesian product of all 18 taxonomy fields would yield a large number of implausible or redundant tasks. We therefore define a coverage blueprint over six dimensions that directly shape the research process: **task structure**, **research object**, **market**, **temporal setting**, **evidence complexity**, and **processing complexity**. Evidence complexity distinguishes specified-source retrieval from open search and single-source tasks from multi-source research; processing complexity ranges from direct extraction and conditional filtering to financial calculation and multi-step synthesis. Together, these two dimensions jointly determine the difficulty distribution: easy questions involve direct or limited-step retrieval, medium questions require additional alignment, verification, or straightforward calculation, and hard questions combine multiple research steps, sources, definitions, periods, calculations, or synthesis requirements.

2.1.3 Financial Source Map

In parallel, finance experts construct a routing map of 138 financial sources spanning major markets and research objects. Each source is annotated with its **identity**, **access path**, **description**, **authority level**, **applicable entities and metrics**, **task suitability**, and **geographic coverage**. The map covers regulators and exchanges, statutory and corporate disclosures, official statistics, international and industry organizations, professional financial databases, specialist research providers, and established financial media. Its authority hierarchy prioritizes statutory, official, and first-party

evidence, while allowing established secondary sources when they are appropriate to the task and can be corroborated.

The source map is a routing aid rather than an answer key. For each question, experts jointly consider the research object, metric, market, temporal requirement, and verification objective to identify suitable primary sources and valid alternatives. The exact documents and evidence accepted for an individual question are recorded separately in its reference package.

2.2 Expert-Led Data Construction

2.2.1 Question Authoring

More than 50 finance professionals participate in question construction. Experts author new questions according to the taxonomy and coverage blueprint, with each question assigned structured labels under the same schema used during demand analysis. Coverage is monitored continuously by comparing the distribution of candidate questions against the target; experts are directed to fill documented gaps while redundant or over-represented candidates are revised or removed, creating an iterative loop between authoring and coverage auditing.

Each question must define the research target with sufficient precision. Where relevant to the answer, the prompt specifies the target entity, temporal range or cutoff, financial or statistical definition, unit, expected output, and numerical precision. Questions whose difficulty derives primarily from ambiguity, inaccessible evidence, obscure trivia, or unnecessary computation are revised or excluded.

2.2.2 Reference-Package Construction

For each candidate question, the expert uses the source map to identify task-appropriate primary sources and valid alternatives, then constructs an item-specific reference package. The package contains: reference answer; authoritative documents or URLs with precise evidence locations, such as the filing, section, page, table, or data-series identifier; the required facts and numerical inputs; relevant definitions, reporting periods, units, currencies, consolidation scopes, and data versions; intermediate calculations or synthesis steps; and acceptable alternative sources, equivalent definitions, and numerical tolerances. For time-sensitive items, the package additionally records the reference cutoff, publication or filing date, access date, and evidence version, collectively preserving reproducibility when the underlying information is subsequently updated, revised, or restated.

2.2.3 Atomic Rubric Construction

Finance experts further decompose the reference package into independently assessable atomic criteria, each of which must be required by the question, supported by the reference package, and specific enough to evaluate without unstated assumptions. The criteria span three capability groups: **raw-information acquisition**, covering whether each required fact or value is retrieved with the correct entity, period, unit, currency, and definition; **source verification**, assessing whether the agent uses authoritative, task-relevant, and time- and version-valid evidence; and **computation and answer formation**, measuring the correctness of the required calculations or reasoning and the completeness and format of the final answer. Traceability is enforced across all three groups by linking key facts, numerical inputs, intermediate calculations, and conclusions to supporting evidence. When financially equivalent solutions exist, the rubric records valid alternative evidence paths, equivalent definitions, and explicit numerical tolerances, thereby rewarding correct inter-

mediate work, avoiding the treatment of a single reference path as the only valid solution, and preventing unsupported answers from receiving full credit merely by matching the final value.

2.3 Quality Control

Every candidate undergoes the same six-stage quality-control process; those that fail a stage are revised and reassessed or removed.

- **Value and scope review.** Reviewers verify that the question represents a meaningful financial need with a clear entity, period, definition, source requirement, and answer boundary. Questions based on artificial search traps, immaterial facts, or complexity unrelated to financial research are revised or rejected.
- **Independent re-solving.** A second finance expert solves the question from scratch, independently locating authoritative evidence, verifying document versions and reporting periods, extracting the required inputs, and reproducing the calculation or analysis.
- **Cross-validation and adjudication.** The independent solution from the second finance expert is compared with the original reference package, and any disagreements are localized to their source, version, entity scope, definition, period, unit, currency, calculation, or interpretation and resolved against the strongest available evidence rather than by majority vote.
- **Rubric audit.** Reviewers verify that each atomic criterion is explicitly required by the question, supported by the reference evidence, assigned to the appropriate capability group, and weighted consistently. The rubric must not introduce source, format, or intermediate-result requirements that are absent from the prompt. Any revision to the question, answer, evidence, calculation, or tolerance is propagated to the other components.
- **Model-based stress testing.** Candidate questions are executed by multiple LLMs, including CLAUDE-OPUS-5 (Anthropic, 2026), GPT-5.6-SOL (OpenAI, 2026), and GLM-5.3 (Z.ai, 2026b), to expose residual ambiguity, outdated evidence, omitted sources, brittle scoring conditions, and unintended shortcuts. Model outputs are used only to identify potential issues and calibrate task difficulty; all revisions and acceptance decisions remain under expert control.
- **Final consistency check.** Accepted candidates undergo deduplication, leakage checks, taxonomy normalization, evidence-link and version verification, and cross-component consistency checks. Reviewers confirm that the prompt, reference package, atomic rubric, and taxonomy labels describe the same research task and evidence requirements.

2.4 Data Statistics

After quality-control filtering, FINFIRST retains 123 expert-authored and verified questions, corresponding to an overall item acceptance rate of 9.78%. Figure 2 summarizes the benchmark along ten dimensions: difficulty, rubric allocation, task structure, research object, market and geography, language, computation, conditional filtering, and source specification and count.

The benchmark follows a graded difficulty distribution of 31.7% easy, 42.3% medium, and 26.0% hard questions, with prompts and reference answers averaging 187.02 and 90.93 characters per task, respectively. Rubric points are distributed across raw-information acquisition (59.5%), source verification (17.0%) and computation and answer formation (23.5%), reflecting the benchmark's dual emphasis on retrieving required information and producing evidence-supported answers. The

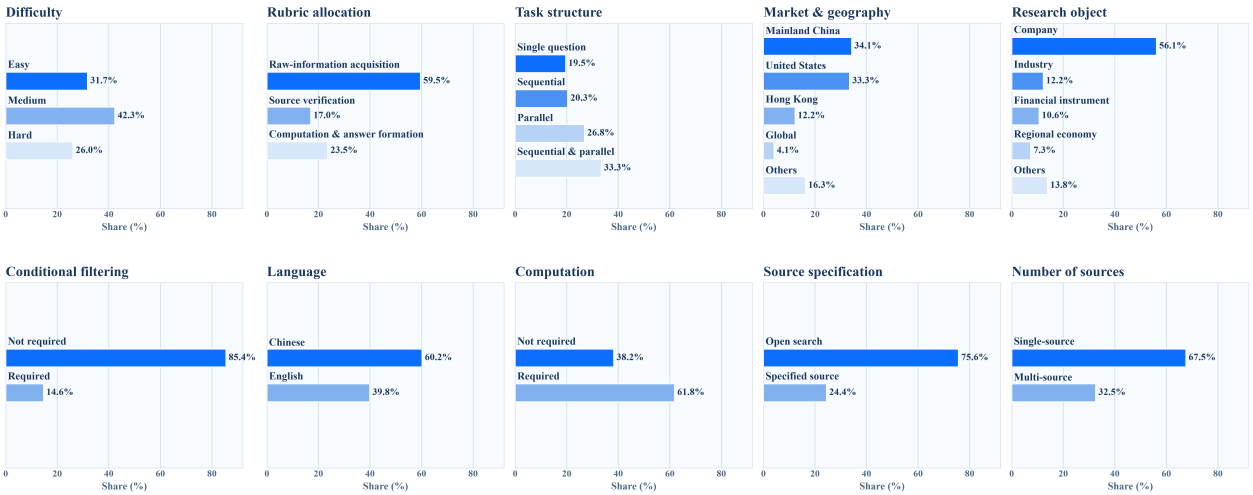

Figure 2 Composition of FINFIRST across dataset and rubric dimensions. Each panel reports category shares within the corresponding dimension. Percentages are rounded, and long-tail categories are grouped under *Others* where applicable.

final composition combines broad coverage with substantial process complexity: more than four-fifths of the questions contain multiple related subproblems, 61.8% require explicit computation, 32.5% draw on multiple sources, and 14.6% involve conditional filtering. While 75.6% of questions require open search without a specified source, the remainder involve designated sources—all of which are publicly accessible. Companies constitute the largest research-object category, with mainland China and the United States contributing comparable shares of geographic coverage. The benchmark is bilingual, comprising 60.2% Chinese and 39.8% English questions.

3 Experiment

3.1 Evaluation Setup

We evaluate 15 model configurations on all 123 questions in FINFIRST under a common agent harness. All models follow the same ReAct-style (Yao et al., 2023) interaction protocol and share an identical toolset—web search, page visit, and Python execution—thereby reducing variation introduced by proprietary search stacks and keeping the focus on each agent's ability to plan searches, retrieve evidence, and construct well-supported answers. We set the output temperature T of 1.0, and results were reported over one designated run for each model configuration. Appendix A details the tool interfaces and their roles in the research workflow.

3.2 Models

A total of 15 model configurations are evaluated: QWEN3.8-MAX (Qwen Team, 2026c), QWEN3.8-27B (Qwen Team, 2026a), QWEN3.8-FLASH (Qwen Team, 2026b), MINIMAX-M3 (MiniMax, 2026), HUNYUAN3-THINKING (Hunyuan Team, 2026), GLM-5.3-FLASH (Z.ai, 2026c), GEMINI-3.7-FLASH (DeepMind, 2026), GLM-5.2 (Z.ai, 2026a), CLAUDE-OPUS-5 (Anthropic, 2026), KIMI-K3 (Kimi Team, 2026), GLM-5.3 (Z.ai, 2026b), DEEPSEEK-V4-PRO (DeepSeek-AI, 2026a), DEEPSEEK-V4-FLASH (DeepSeek-AI, 2026b), and GPT-5.6-SOL (OpenAI, 2026)—together with the finance-specialized LING-3.0-FLASH-FIN (inclusionAI, 2026). All results are produced under the common basic-tool setting described above. Detailed model cards and evaluated identifiers are provided in Appendix B.

3.3 Evaluation Metrics

Each question q contains a set of atomic rubric items $\mathcal{R}_q$, with the total count across all 123 questions summing to $\sum_{q \in \mathcal{Q}} |\mathcal{R}_q| = 701$. Each rubric item i carries an expert-defined weight $w_{q,i}$, normalized within each question to 100 points ($\sum_{i \in \mathcal{R}_q} w_{q,i} = 100$), yielding $123 \times 100 = 12{,}300$ weighted rubric points in total. Given a binary judge decision $z_{q,i} \in \{0, 1\}$ for each item, we evaluate models using three complementary metrics:

$$\text{Atomic} = \frac{\sum_q \sum_{i \in \mathcal{R}_q} z_{q,i}}{\sum_q |\mathcal{R}_q|}, \tag{1}$$

$$\text{Weighted Rubric Score} = \frac{\sum_q \sum_{i \in \mathcal{R}_q} w_{q,i} z_{q,i}}{\sum_q \sum_{i \in \mathcal{R}_q} w_{q,i}}, \tag{2}$$

$$\text{Strict Pass} = \frac{1}{|\mathcal{Q}|} \sum_q \mathbb{I}\left[\forall i \in \mathcal{R}_q,\ z_{q,i} = 1\right]. \tag{3}$$

Atomic is the unweighted pass rate over all 701 criteria, treating every item equally regardless of its assigned weight. **Weighted Rubric Score** (referred to as **Loose** on the internal leaderboard) grants partial credit proportional to expert-assigned weights across the full 12,300-point benchmark. **Strict Pass** counts a question as correct only if all of its atomic criteria are satisfied.

Each atomic criterion is assigned to exactly one of three capability groups—raw-information acquisition, source verification, and computation and answer formation—with group-level scores reported as weighted pass rates over their respective point totals

$$W_g = \sum_{q \in \mathcal{Q}} \sum_{i \in \mathcal{R}_{q,g}} w_{q,i}, \tag{4}$$

where $\mathcal{R}_{q,g} \subseteq \mathcal{R}_q$ denotes the criteria of question q in group g. This yields 7,322 points for raw-information acquisition, 2,085 for source verification, and 2,893 for computation and answer formation, which together sum to the full benchmark total of 12,300.

3.4 Rubric Judge

We utilize GLM-5.1(GLM-5 Team, 2026) as the rubric judge. Given the question, the atomic rubric items, and the model's final response, the judge independently determines whether each item is fully satisfied—verifying numerical values, units, time ranges, financial definitions, sources, calculations, and output requirements, and rejecting any item with missing evidence, incorrect periods, or mismatched definitions. Scores are then computed deterministically from the resulting binary decisions and its weights. To validate judge reliability, we randomly sampled 50 evaluation instances for annotation by eight finance experts; after adjudication, item-level agreement between GLM-5.1 and the human reference labels yielded Cohen's Kappa κ (Cohen, 1960) = 0.816, indicating strong agreement. The full judge prompt is provided in Appendix C, Table 7.

4 Result and Analysis

4.1 Overall Performance

Table 3 reports performance and tool-use statistics for all 15 model configurations. CLAUDE-OPUS-5 achieves the highest Atomic (87.59%) and Loose Pass (87.61%) scores, while GPT-5.6-SOL leads in

Table 3 Overall, rubric-level, and tool-efficiency results on FINFIRST under the common basic-tool setting. ↓ indicates that lower values denote greater efficiency conditional on comparable task correctness, and efficiency statistics are averaged over questions with recorded trajectories. † indicates incomplete result coverage: GPT-5.6-SOL has 117 valid questions, whereas CLAUDE-OPUS-5, GLM-5.3, QWEN3.8-27B, and LING-3.0-FLASH-FIN each have 122. Missing outputs remain in the fixed 123-question denominator, with all associated rubric items marked as failed.

Model	Overall metrics (%)			Rubric dimensions (%)			Tool efficiency		
	Atomic	Loose Pass	Strict Pass	Raw-information	Source verification	Computation & answer	Avg. rounds ↓	Avg. tool calls ↓	Max. tool calls ↓
GPT-5.6-SOL†	85.45	85.92	**71.54**	86.85	88.68	81.58	24.64	62.28	281
CLAUDE-OPUS-5†	**87.59**	**87.61**	69.11	**88.98**	**89.59**	**82.72**	14.80	23.89	153
GEMINI-3.7-FLASH	75.89	76.75	44.72	81.19	66.62	72.80	23.73	23.73	74
QWEN3.8-MAX	78.17	77.40	54.47	80.33	77.75	69.72	32.20	41.47	167
QWEN3.8-FLASH	82.31	81.23	61.79	84.29	83.36	71.93	25.00	35.37	134
DEEPSEEK-V4-PRO	76.03	75.41	51.22	80.09	73.57	64.88	21.78	22.21	100
GLM-5.3†	79.32	80.61	60.98	83.11	78.56	75.77	18.61	24.11	95
GLM-5.3-FLASH	79.60	76.64	56.91	79.50	80.77	66.44	22.43	30.46	135
GLM-5.2	70.61	65.89	43.09	69.75	68.30	54.37	20.34	23.72	87
KIMI-K3	84.45	80.83	59.35	84.84	70.70	77.98	17.59	36.49	313
DEEPSEEK-V4-FLASH	72.90	71.37	48.78	77.49	71.65	55.69	16.49	21.96	106
QWEN3.8-27B†	78.03	77.28	55.28	81.49	76.83	66.92	28.20	37.33	149
MINIMAX-M3	63.05	60.70	37.40	65.51	62.97	46.87	27.45	42.28	177
HUNYUAN3-THINKING	65.76	65.22	40.65	70.43	68.01	50.02	29.24	40.28	189
LING-3.0-FLASH-FIN†	75.89	75.07	52.85	78.43	82.45	61.25	23.08	30.87	119

Strict Pass rate with 88 of 123 questions fully satisfied (71.54%). Even for the strongest systems, the gap between partial credit and complete task success remains substantial: CLAUDE-OPUS-5's Strict Pass rate trails its Loose Pass score by 18.50 percentage points, indicating that a largely correct solution frequently retains at least one unresolved requirement.

The three rubric groups reveal a consistent downstream bottleneck: computation and answer formation underperforms raw-information acquisition for each model; the gap ranges from 5.27 percentage points for GPT-5.6-SOL to 21.80 points for DEEPSEEK-V4-FLASH. Rubric decomposition further distinguishes models with similar aggregate performance: GLM-5.3 and KIMI-K3 achieve comparable Loose Pass scores (80.61% and 80.83%, respectively), yet GLM-5.3 is stronger in source verification (78.56% vs. 70.70%), while KIMI-K3 leads in both raw-information acquisition (84.84% vs. 83.11%) and computation and answer formation (77.98% vs. 75.77%). Among all evaluated systems, LING-3.0-FLASH-FIN stands out in source verification, reaching 82.45%—the highest among open-weight models in the lower block of Table 3.

Tool-use volume is not monotonically related to performance. GPT-5.6-SOL averages 62.28 tool calls per covered question versus 23.89 for CLAUDE-OPUS-5, and KIMI-K3 records the largest single-question maximum at 313 calls; LING-3.0-FLASH-FIN, by contrast, pairs its strong source-verification score with a moderate average of 30.87 calls and a maximum of 119. Call counts should therefore be treated as efficiency indicators only when comparing systems at similar levels of task correctness, as higher volumes may reflect broad exploration, recovery from retrieval failures, or inefficient search rather than superior research outcomes.

4.2 Unsupported-Correct Answers

Final-answer accuracy alone cannot distinguish a well-supported result from one reached without a complete evidence chain. To quantify this distinction, we partition each decomposable question into an answer layer and an evidence layer: computation-and-answer criteria form the former when available, while raw-information criteria serve this role for direct-retrieval questions, and all

Table 4 Answer correctness and evidence completeness. NA denotes questions without independently separable answer and evidence layers. UCR is $Q3/(Q3+Q4)$, with ↓ indicating that lower is better.

Model	Q1: Wrong, no support	Q2: Wrong, partial process	Q3: Correct, partial evidence	Q4: Correct, fully traceable	NA	UCR (%) ↓
GPT-5.6-SOL†	7	19	8	84	5	**8.70**
CLAUDE-OPUS-5†	1	26	9	82	5	9.89
GEMINI-3.7-FLASH	9	31	27	51	5	34.62
QWEN3.8-MAX	8	30	17	63	5	21.25
QWEN3.8-FLASH	8	26	11	73	5	13.10
DEEPSEEK-V4-PRO	10	36	13	59	5	18.06
GLM-5.3†	9	25	14	70	5	16.67
GLM-5.3-FLASH	10	31	9	68	5	11.69
GLM-5.2	17	36	14	51	5	21.54
KIMI-K3	10	19	19	70	5	21.35
DEEPSEEK-V4-FLASH	13	35	12	58	5	17.14
QWEN3.8-27B†	7	29	17	65	5	20.73
MINIMAX-M3	20	41	12	45	5	21.05
HUNYUAN3-THINKING	17	42	12	47	5	20.34
LING-3.0-FLASH-FIN†	7	41	7	63	5	10.00
Average	10.20	31.13	13.40	63.27	5.00	17.74

remaining criteria constitute the latter. An answer is deemed correct only when all answer-layer criteria pass, and fully supported only when every evidence criterion additionally passes. This yields four outcomes: **Q1**, incorrect with no credited support; **Q2**, incorrect with partial research progress; **Q3**, correct but incompletely supported; and **Q4**, correct and fully traceable.

For model m, we define the **Unsupported-Correct Rate** as

$$\mathrm{UCR}_m = \frac{N_m(Q3)}{N_m(Q3) + N_m(Q4)}, \tag{5}$$

where the denominator contains all correct-answer attempts. UCR directly captures the value of the evidence-aware atomic rubrics in FINFIRST: a lower value indicates that correct answers are more consistently accompanied by complete, verifiable support. Across the 1,770 decomposable attempts, 1,150 answers are correct, of which 201 fall into Q3, yielding a micro-averaged UCR of 17.48% and an unweighted mean across models of 17.74%—meaning that final-answer-only evaluation would overstate fully traceable success for nearly one in six correct-answer attempts. GPT-5.6-SOL achieves the lowest UCR (8.70%) and GEMINI-3.7-FLASH the highest (34.62%). LING-3.0-FLASH-FIN reaches 10.00%—third-lowest overall and below every evaluated open-weight model—indicating that its correct answers are typically backed by complete evidence chains.

4.3 Slice Analysis

Figure 3 reports per-model Strict Pass across four dataset dimensions, with missing outputs counted as failures. Task complexity yields the clearest trend: cross-model averages decline from 60.55% on easy questions to 52.90% on medium and 47.71% on hard, with 14 of 15 models following this monotonic decline. Source multiplicity produces a comparably consistent effect, as single-source questions average 58.54% versus 46.94% for multi-source questions, with only GLM-5.3 reversing the pattern. Market-level performance is more heterogeneous: Hong Kong questions record the highest average (64.81%), followed by the United States (56.74%), mainland China (52.22%), and other markets (45.25%). Within this dimension, U.S.-developed models hold a clear advantage on U.S. questions (67.47% vs. 54.06% for China-developed models), whereas China-developed models show a narrow edge on Hong Kong questions (65.01% vs. 64.00%), suggesting partial regional specialization rather than uniform home-market dominance. Finally, English questions

Figure 3 Per-model Strict Pass across task complexity, source count, market and geography, and language. Model names are shown in **Green** for U.S.-developed models and **Blue** for China-developed models.

outscore Chinese questions on average (57.14% vs. 51.71%), though model-level exceptions—notably QWEN3.8-MAX and LING-3.0-FLASH-FIN—indicate that this aggregate gap partly reflects differences in market and task composition rather than language alone.

5 Related Work

5.1 Financial QA Benchmarks

Early financial benchmarks primarily evaluate language understanding and numerical reasoning within a bounded evidence setting. FinQA (Chen et al., 2021) and ConvFinQA (Chen et al., 2022) pair questions with pre-selected passages and tables from annual reports, and additionally annotate executable reasoning programs or conversational reasoning chains. TAT-QA (Zhu et al., 2021) focuses on arithmetic reasoning over supplied tabular and textual evidence, while FinanceBench (Islam et al., 2023) evaluates retrieval and question answering over a fixed corpus of public filings. Broader suites extend coverage to sentiment analysis, text classification, forecasting, multilinguality, and multimodality: FLUE (Shah et al., 2022), PIXIU/FLARE (Xie et al., 2023), FinBen (Xie et al., 2024), MultiFinBen (Peng et al., 2025), and CFBenchmark (Lei et al., 2023) aggregate diverse financial NLP tasks from multiple sources. FinEval (Guo et al., 2025) additionally covers Chinese financial knowledge, security, and agent-oriented questions; CPA-KQA (Kuang et al., 2025) introduces skill-level cognitive diagnosis using professional accounting examinations; and FinDABench (Liu et al., 2025) targets financial data analysis and visualization. To improve professional realism, FinanceQA (Mateega et al., 2025) employs tasks written and verified by domain practitioners, while BizFinBench (Lu et al., 2025) grounds its Chinese tasks in queries drawn from a real-world financial application.

Despite their diversity, these benchmarks share a fundamental limitation: they assume that relevant evidence has already been provided or bounded, and thus cannot measure end-to-end performance

spanning the full pipeline—from discovering and validating sources to extracting structured information and generating the final answer.

5.2 Agentic Financial Benchmarks

Recent benchmarks have increasingly shifted toward evaluating financial agents that retrieve and synthesize external information. Finance Agent Benchmark (Bigeard et al., 2025) comprises 537 expert-authored questions spanning nine research categories, and provides a model-agnostic evaluation harness with access to web search and SEC EDGAR; its atomic rubrics represent an improvement over exact-match scoring, though the published judge assesses answer content rather than the quality of the underlying retrieval or reasoning process. FinSearchComp (Hu et al., 2025) contains 635 questions constructed and quality-controlled by 70 finance professionals, covering time-sensitive data fetching, historical lookup, and complex investigation tasks, yet assigns each response a binary final-answer score. FrontierFinance (Zhang et al., 2026) advances long-form investment research evaluation through 220 timestamped expert queries and 11,543 source-attributed criteria across six investor workflows, while also comparing systems built on different tool harnesses; however, its grading remains focused primarily on the content of final responses. A separate benchmark under the same name (Krumdick et al., 2026) extends this line to long-horizon computer-use tasks that produce financial models and client-ready artifacts. FinDeepIndicator (Yang et al., 2026) takes a more diagnostic approach by independently scoring formula specification, data collection, indicator calculation, and answer generation, but its scope is confined to template-derived indicator-construction questions.

Taken together, existing benchmarks in this line share three limitations: they either emphasize final-answer content without systematically diagnosing the full evidence-to-answer chain, conflate model capability with the effects of heterogeneous tool environments, or offer process-level evaluation only within a narrow task family. None simultaneously combines broad, demand-informed financial search coverage with provenance-aware grading, stage-level process diagnosis, and a common configurable evaluation harness.

5.3 General Web Search Benchmarks

Domain-general benchmarks provide complementary views of browsing and tool use across diverse task types. GAIA (Mialon et al., 2023) tests browsing, multimodality, reasoning, and tools on conceptually simple but real-world questions. BrowseComp (Wei et al., 2025) and BrowseComp-ZH (Zhou et al., 2025) stress multi-step web search for short, time-invariant answers, while BrowseComp-Plus (Chen et al., 2025) adds a fixed human-verified evidence corpus, controlled retrievers, and citation-retrieval metrics. WebWalkerQA (Wu et al., 2025) evaluates traversal through website hierarchies, and AssistantBench (Yoran et al., 2024) evaluates realistic and time-consuming open-web tasks. SimpleQA (Wei et al., 2024) and Humanity's Last Exam (Phan et al., 2025) test short-form factuality and expert-level knowledge, respectively, without treating end-to-end web research as their primary object of evaluation. Mind2Web 2 (Gou et al., 2025) uses tree-structured rubrics and judge agents to assess complex, time-varying answers and source attribution; however, it evaluates final retrieved information rather than intermediate interactions and avoids tasks that explicitly require complex reasoning or external computation.

The central limitation of these benchmarks for our setting is their lack of financial specialization: they do not encode authoritative financial source hierarchies, reporting-period and definition alignment, strict numerical accuracy, or the evidence standards needed for auditable investment

research.

6 Conclusion

We introduced FINFIRST, an expert-authored benchmark designed to evaluate financial search agents beyond final-answer matching. Its 123 tasks are grounded in real-world financial scenarios and constructed through an 18-field taxonomy, a 6-axis coverage blueprint, a registry of 138 financial sources, contributions from over 50 finance experts, and a 6-stage quality-control pipeline. Each task includes an evidence-grounded reference package and atomic rubrics, enabling fine-grained assessment of both the answer and the research process supporting it.

Our evaluation of 15 model configurations under a unified tool environment reveals that current agents remain far from consistently completing financial search tasks. CLAUDE-OPUS-5 achieves the highest Atomic and Loose Pass scores, at 87.59% and 87.61%, respectively, yet its Strict Pass rate falls to only 69.11%. Across models, computation and answer formation consistently lags behind raw-information acquisition, and 17.48% of correct-answer attempts lack a complete, verifiable evidence chain. These findings demonstrate that final-answer accuracy alone can obscure incomplete evidence, unresolved requirements, and distinct capability bottlenecks. By making the underlying research process measurable and diagnosable, FINFIRST provides a foundation for developing financial agents whose answers are not only correct, but also well-supported, traceable, and independently verifiable.

References

Anthropic. Claude Opus 5 system card. Official system card, 2026. https://www.anthropic.com/claude-opus-5-system-card.

Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. Finance Agent Benchmark: Benchmarking LLMs on Real-world Financial Research Tasks, 2025. https://arxiv.org/abs/2508.00828.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300. https://aclanthology.org/2021.emnlp-main.300/.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.421. https://aclanthology.org/2022.emnlp-main.421/.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent, 2025. https://arxiv.org/abs/2508.06600.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46, 1960. https://www.semanticscholar.org/paper/A-Coefficient-of-Agreement-for-Nominal-Scales-Cohen/9e463eefadbcd336c69270a299666e4104d50159.

DeepMind. Gemini 3.7 Flash model card. Official model card, August 2026. https://deepmind.google/models/model-cards/gemini-3-7-flash/.

DeepSeek-AI. DeepSeek V4 technical documentation and model card. Official technical documentation, April 2026a. https://fe-static.deepseek.com/chat/transparency/deepseek-V4-model-card-EN.pdf.

DeepSeek-AI. DeepSeek-V4-Flash-0731 model card. Hugging Face model card, July 2026b. https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731.

GLM-5 Team. GLM-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. https://arxiv.org/abs/2602.15763.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge, 2025. https://arxiv.org/abs/2506.21506.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6258–6292, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.318. https://aclanthology.org/2025.naacl-long.318/.

Liang Hu, Jianpeng Jiao, Jiashuo Liu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, Fei Hu, Yali Liao, Zaiyuan Wang, Chenghao Yang, Qianyu Yang, Mingren Yin, Zhiyuan Zeng, Ge Zhang, Xinyi Zhang, Xiying Zhao, Zhenwei Zhu, Hongseok Namkoong, Wenhao Huang, and Yuwen Tang. FinSearchComp: Towards a Realistic, Expert-Level Evaluation of Financial Search and Reasoning, 2025. https://arxiv.org/abs/2509.13160.

Hunyuan Team. Hy3: A reasoning and agent model. Official model repository and model card, 2026. https://github.com/Tencent-Hunyuan/Hy3.

inclusionAI. Ling-3.0-Flash-Fin model card. Openrouter model card, 2026. https://openrouter.ai/inclusionai/ling-3.0-flash-fin:free.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A New Benchmark for Financial Question Answering, 2023. https://arxiv.org/abs/2311.11944.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. https://arxiv.org/abs/2607.24653.

Michael Krumdick, Varshini Reddy, Shivani Chaudhary, William Day, Maarij Ahmed, Hayan Haqqi, Muhammad Ahsen Fahim, Hanzallah Amjad, Ahmad Orakzai, Aqsa Gul, and Chris Tanner. FrontierFinance: A Long-Horizon Computer-Use Benchmark of Real-World Financial Tasks, 2026. https://arxiv.org/abs/2604.05912.

Ziyan Kuang, Feiyu Zhu, Maowei Jiang, Yanzhao Lai, Zelin Wang, Zhitong Wang, Meikang Qiu, Jiajia Huang, Min Peng, Qianqian Xie, and Sophia Ananiadou. From Scores to Skills: A Cognitive Diagnosis Framework for Evaluating Financial Large Language Models, 2025. https://arxiv.org/abs/2508.13491.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, and Pengyu Zhao. MiniMax Sparse Attention. *arXiv preprint arXiv:2606.13392*, 2026. https://arxiv.org/abs/2606.13392.

Yang Lei, Jiangtong Li, Dawei Cheng, Zhijun Ding, and Changjun Jiang. CFBenchmark: Chinese Financial Assistant Benchmark for Large Language Model, 2023. https://arxiv.org/abs/2311.05812.

Shu Liu, Shangqing Zhao, Chenghao Jia, Xinlin Zhuang, Zhaoguang Long, Jie Zhou, Aimin Zhou, Man Lan, and Yang Chong. FinDABench: Benchmarking financial data analysis ability of large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 710–725, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.48/.

Guilong Lu, Xuntao Guo, Rongjunchen Zhang, Wenqiao Zhu, and Ji Liu. BizFinBench: A Business-Driven Real-World Financial Benchmark for Evaluating LLMs, 2025. https://arxiv.org/abs/2505.19457.

Spencer Mateega, Carlos Georgescu, and Danny Tang. FinanceQA: A Benchmark for Evaluating Financial Analysis Capabilities of Large Language Models, 2025. https://arxiv.org/abs/2501.18062.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for General AI Assistants, 2023. https://arxiv.org/abs/2311.12983.

MiniMax. MiniMax M3: Frontier coding, 1m context, native multimodality—all in one model. Official research release, June 2026. https://www.minimax.io/blog/minimax-m3.

OpenAI. GPT-5.6 system card. OpenAI Deployment Safety Hub, 2026. https://deploymentsafety.openai.com/gpt-5-6.

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, Shengyuan Lin, Keyi Wang, Shanshan Yang, Yilun Zhao, Zhiwei Liu, Peng Lu, Jerry Huang, Suyuchen Wang, Triantafillos Papadopoulos, Polydoros Giannouris, Efstathia Soufleri, Nuo Chen, Zhiyang Deng, Heming Fu, Yijia Zhao, Mingquan Lin, Meikang Qiu, Kaleb E Smith, Arman Cohan, Xiao-Yang Liu, Jimin Huang, Guojun Xiong, Alejandro Lopez-Lira, Xi Chen, Junichi Tsujii, Jian-Yun Nie, Sophia Ananiadou, and Qianqian Xie. MultiFinBen: Benchmarking Large Language Models for Multilingual and Multimodal Financial Application, 2025. https://arxiv.org/abs/2506.14028.

Long Phan et al. Humanity's Last Exam, 2025. https://arxiv.org/abs/2501.14249.

Qwen Team. Qwen3.8-27B model card. Hugging Face model card, August 2026a. https://huggingface.co/Qwen/Qwen3.8-27B.

Qwen Team. Qwen3.8-Flash model documentation. Qwen Cloud documentation, August 2026b. https://docs.qwencloud.com/developer-guides/getting-started/vision-models.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. Official model release, August 2026c. https://qwen.ai/blog?id=qwen3.8.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.148. https://aclanthology.org/2022.emnlp-main.148/.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. https://arxiv.org/abs/2411.04368.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents, 2025. https://arxiv.org/abs/2504.12516.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. WebWalker: Benchmarking LLMs in Web Traversal, 2025. https://arxiv.org/abs/2501.07572.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance, 2023. https://arxiv.org/abs/2306.05443.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. FinBen: A Holistic Financial Benchmark for Large Language Models, 2024. https://arxiv.org/abs/2402.12659.

Chaoqun Yang, Fengbin Zhu, Xinyu Lin, Long Bai, Xiaoluan Liu, Ke-Wei Huang, Roger Zimmermann, and Tat-Seng Chua. FinDeepIndicator: Benchmarking Deep Research Agents in End-to-End Financial Indicator Construction, 2026. https://arxiv.org/abs/2608.00764.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. https://arxiv.org/abs/2210.03629.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. AssistantBench: Can Web Agents Solve Realistic and Time-Consuming Tasks?, 2024. https://arxiv.org/abs/2407.15711.

Z.ai. GLM-5.2: Built for long-horizon tasks. Official research release, June 2026a. https://z.ai/blog/glm-5.2.

Z.ai. GLM-5.3: Frontier coding with emergent cyber capabilities. Official research release, August 2026b. https://z.ai/blog/glm-5.3.

Z.ai. GLM-5.3-Flash: Frontier intelligence, flash cost. Official research release, August 2026c. https://z.ai/blog/glm-5.3-flash.

Yuhao Zhang, O. Ozan Koyluoglu, Thejas Venkatesh, Richard Diehl Martinez, Vishank Bhatia, Arash Alidoust, and Ashwin Paranjape. FrontierFinance: A Challenging Benchmark for Measuring Frontier Intelligence of Finance Agents, 2026. https://arxiv.org/abs/2608.11683.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. BrowseComp-ZH: Benchmarking Web Browsing Ability of Large Language Models in Chinese, 2025. https://arxiv.org/abs/2504.19314.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.254. https://aclanthology.org/2021.acl-long.254/.

Table 5 Tools available to every evaluated agent in the common basic-tool environment.

Tool	Primary input	Returned information	Role in the workflow
Web Search	A textual search query	A ranked list of candidate results, including the title, URL, and snippet when available	Discover potentially relevant official filings, company disclosures, statistics, reports, databases, and other public sources
Visit	One or more target URLs	Extracted content and available document metadata from HTML webpages or PDF documents	Inspect a selected source, locate supporting passages or tables, and retrieve evidence used in the answer
Python	Executable Python code	Standard output, standard error, and execution status from an isolated runtime with common numerical and data-analysis libraries	Perform arithmetic, aggregation, unit conversion, filtering, tabular manipulation, and consistency checks

A Tool Environment

Our evaluation environment provides three complementary tools for information discovery, evidence inspection, and numerical analysis. All evaluated models receive the same tool descriptions, interfaces, execution limits, and per-task budgets. The reported leaderboard excludes runs that use additional professional financial tools, so differences in performance are not attributable to unequal access to proprietary databases or model-specific plugins.

A.1 Tool Interfaces

Table 5 summarizes the common interfaces. The schemas are intentionally compact: search discovers candidate sources, visit exposes the contents of a selected source, and Python performs deterministic computation over retrieved information. Search-result snippets are treated only as navigation aids; evidence-based answers are expected to rely on content inspected through the visit tool.

Web Search. The search tool accesses search results via SerpAPI[1]. It accepts a natural-language query and returns ranked candidate pages. Agents may issue multiple calls and reformulate queries as the research process develops. Each result provides sufficient routing information to decide whether the underlying page should be inspected. Because snippets can be incomplete, stale, or detached from their original context, they are not treated as authoritative evidence by themselves.

Visit. The visit tool employs Jina Reader[2] to retrieve and parse the contents of a specified public URL. It supports both webpages and PDF documents and returns machine-readable text together with available structural or document metadata. The tool is used to inspect primary materials, identify exact evidence locations, and verify entities, periods, units, definitions, and data versions. If extraction fails or a page is inaccessible, the failure is returned to the agent, which may search for an alternative public source within the remaining budget.

Python. The Python tool executes code in an isolated environment with common numerical and data-analysis packages. It supports calculations and transformations that would be error-prone to perform manually, including multi-period aggregation, percentage and growth-rate computation, unit normalization, conditional filtering, and tabular checks. The tool does not provide independent

[1] https://serpapi.com/
[2] https://jina.ai/

Table 6 Model cards for the systems evaluated in FINFIRST.

MODEL	PROVIDER	CHECKPOINTS	AVAIL.	#PARAM	REFERENCE
GPT-5.6-SOL	OpenAI	`gpt-5.6-sol`	🔒	–	OpenAI, 2026
CLAUDE-OPUS-5	Anthropic	`claude-opus-5`	🔒	–	Anthropic, 2026
GEMINI-3.7-FLASH	Google DeepMind	`gemini-3.7-flash`	🔒	–	DeepMind, 2026
QWEN3.8-MAX	Alibaba/Qwen	`qwen3.8-max`	🔒	2.4T / 95B active	Qwen Team, 2026c
QWEN3.8-FLASH	Alibaba/Qwen	`qwen3.8-flash`	🔒	–	Qwen Team, 2026b
DEEPSEEK-V4-PRO-0813	DeepSeek-AI	`DeepSeek-V4-Pro`	🔒	1.6T / 49B active	DeepSeek-AI, 2026a
GLM-5.3	Z.ai	`zai-org/GLM-5.3`	🔓	753B / 40B active	Z.ai, 2026b
GLM-5.3-FLASH	Z.ai	`zai-org/GLM-5.3-Flash`	🔓	320B / 18B active	Z.ai, 2026c
GLM-5.2	Z.ai	`zai-org/GLM-5.2`	🔓	753B / 40B active	Z.ai, 2026a
KIMI-K3	Moonshot AI	`moonshotai/Kimi-K3`	🔓	2.8T / 104B active	Kimi Team, 2026
DEEPSEEK-V4-FLASH-0731	DeepSeek-AI	`deepseek-ai/DeepSeek-V4-Flash`	🔓	285B / 13B active	DeepSeek-AI, 2026b
QWEN3.8-27B	Alibaba/Qwen	`Qwen/Qwen3.8-27B`	🔓	27B	Qwen Team, 2026a
MINIMAX-M3	MiniMax	`MiniMaxAI/MiniMax-M3`	🔓	428B / 23B active	MiniMax, 2026; Lai et al., 2026
HUNYUAN3-THINKING	Tencent	`tencent/Hy3`	🔓	295B / 21B active	Hunyuan Team, 2026
LING-3.0-FLASH-FIN	inclusionAI	`Ling-3.0-Flash-Fin`	🔓	124B / 5.1B active	inclusionAI, 2026

financial evidence: all numerical inputs must still be grounded in sources retrieved through the search and visit workflow.

A.2 Controlled Execution

All models are evaluated through the same ReAct-style loop. At each step, an agent may produce a final response or invoke one of the three tools; the resulting observation is appended to the interaction history before the next model call. Tool-call budgets, timeouts, network conditions, and termination rules are held fixed across model configurations. Tool or model-service failures are recorded explicitly. Under the fixed-123-question protocol used in the main results, an unresolved execution or judging failure remains in the denominator and receives zero credit rather than being silently removed.

B Evaluated Models

Table 6 summarizes the model configurations used in our experiments. We report the evaluated API identifier or public checkpoint whenever it is available. Parameter counts follow official model documentation; "active" denotes the parameters activated per token for sparse mixture-of-experts models, and "–" indicates that the provider has not publicly disclosed the value. 🔒 denotes a proprietary or hosted configuration, whereas 🔓 denotes publicly available model weights.

C Judge Prompt Template

Table 7 provides the Judge Prompt Template; Chinese queries use a corresponding Chinese version. The placeholders are populated with the evaluation date, question, atomic rubric items, and model response for each instance. Since the judge observes only the final response—not the agent's hidden reasoning or tool trajectory—source-verification and traceability scores reflect the evidence presented in the final response rather than every latent action taken during retrieval. Recorded trajectories are analyzed separately for tool-use statistics and qualitative case studies, and are not used to overwrite rubric decisions in the main leaderboard.

Table 7 Prompt template for the automatic rubric judge.

Automatic Rubric-Judge Prompt

You are a strict and impartial judge for financial-search evaluation. Given a question and its itemized rubric, determine whether the model response satisfies each rubric item.

Current evaluation date:
`{current_date}`

Evaluation requirements:

1. Base your decisions solely on the itemized rubric. Do not refer to any reference answer or external ground truth.
2. Evaluate every rubric item independently and determine whether it is fully satisfied. Do not output a score for any item; scores are calculated by the evaluation system.
3. Strictly verify numerical values, units, time ranges, measurement definitions and conventions, data sources, calculation procedures, and final-answer requirements.
4. If the model provides only a conclusion but omits a key source or calculation step required by the rubric, mark the corresponding rubric item as not satisfied.
5. If the model uses an incorrect source, year, measurement definition, or convention, mark the corresponding rubric item as not satisfied.
6. The `category` annotation of each rubric item is used only for subsequent attribution analysis and must not affect the strictness with which that item is evaluated.
7. For each rubric item, output only `passed=true` or `passed=false`. Output `true` only if the item is fully satisfied; otherwise, output `false`.
8. Treat facts, numerical values, and dates explicitly stated in the rubric as authoritative benchmark annotations. Do not reject information marked as correct in the rubric based on your own assessment of dates or data-release timing.
9. If the rubric requires an official source, a specific source, or a specific definition, strictly verify that the model response satisfies the corresponding source and definitional requirements.
10. The output must be valid JSON. Do not enclose it in a Markdown code block.

Question:
`{question}`

Itemized rubric:
(JSON; item numbers and point values have been removed from the rubric text, and `category` is used only for attribution analysis)
`{rubric_items}`

Model response:
`{model_answer}`

Return JSON in the following format:

```
{
  "rubric_item_judgements": [
    {
      "id": 1,
      "passed": true,
      "category": "category from the corresponding input item",
      "reason": "primary basis for determining whether the item is satisfied"
    }
  ],
  "summary": "one-sentence summary of the main strengths and deficiencies",
  "deductions": ["major deduction 1", "major deduction 2"]
}
```

Note: Do not output `total_score`, `score`, or `max_score`. The evaluation system calculates the final score for each rubric item from its `passed` value.