FINFIRST: BENCHMARKING SEARCH AGENTS FOR FINANCIAL INFORMATION RETRIEVAL, SOURCING AND TRACEABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial search poses a demanding challenge for LLM agents, requiring not only correct final answers but also responses grounded in temporally valid information, authoritative sources, and correctly aligned units and definitions. Yet existing benchmarks predominantly evaluate only the final answer correctness, obscuring both the causes of error and completeness of the supporting evidence. Our analysis revealed that 17.48% of correctly answered tasks lack a complete and verifiable evidence chain. To address this gap, we introduce **FINFIRST** (**F**inancial **I**nformation **R**etrieval, **S**ourcing and **T**raceability), a financial-search benchmark that jointly evaluates answers and evidential support through atomic rubrics. FINFIRST comprises 123 expert-authored tasks spanning a graduated difficulty spectrum, derived from aggregate patterns of real-world financial scenarios and informed by an 18-field taxonomy, a 6-axis coverage blueprint, and a registry of 138 financial sources. Task construction involved contributions from over 50 finance experts and a 6-stage quality-control pipeline. Each task is paired with an evidence-grounded reference package decomposed into atomic criteria across three dimensions: raw-information acquisition, source verification, and computation and answer formation. We evaluate 15 model configurations under a unified tool setting: Claude-Opus-5 achieves the highest Loose Pass rate at 87.61%, while GPT-5.6-Sol attains the highest Strict Pass rate at 71.54%, with computation and answer formation consistently lagging behind raw-information acquisition across all systems. FINFIRST thus retains final-answer correctness as the primary objective while making the supporting research process measurable, verifiable, and diagnosable.

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). Reliable search in finance rarely reduces to locating a single webpage or matching a number. Instead, an agent must identify authoritative, time-valid sources; retrieve information for the correct entity and reporting period; align units and financial definitions; reconcile discrepancies across sources; perform requisite calculations; and ground conclusions in verifiable evidence. Errors at any stage can yield plausible yet unsupported answers. Evaluating financial search agents thus requires examining not only final answers but also the supporting evidence and intermediate derivations presented alongside them.

Existing benchmarks capture complementary aspects of this workflow. Financial QA datasets such as ConvFinQA (Chen et al., 2022) and FinanceBench (Islam et al., 2023) provide the relevant evidence as input and thus primarily evaluate reasoning over supplied information rather than end-to-end retrieval. More recent agentic benchmarks target broader financial tasks. For example, FinDeepIndicator (Yang et al., 2026) evaluates the staged construction of financial indicators from predefined templates. FinSearchComp (Hu et al., 2025) is most closely aligned with our setting, comprising 635 expert-curated financial tasks spanning time-sensitive retrieval, historical lookup, and complex investigation. However, its binary final-answer evaluation cannot identify whether a failure arises from source selection, data extraction, definition alignment, or calculation; nor can it distinguish a well-supported

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of representative financial benchmarks. **Agentic**: requires active information retrieval or external tool use. **Time-sensitive**: correct answers are tied to a specific time point and may change as new information emerges. **Demand-aligned**: reflects authentic financial needs or real-world professional workflows. **Expert-authored**: designed and verified by domain experts in finance. **Multi-source**: integrates and reconciles evidence from multiple heterogeneous sources. **Evidence scoring**: assesses the quality and relevance of supporting sources and evidence. **Atomic assessment**: decomposes tasks into rubric-specific criteria rather than evaluating final-answer correctness.

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)	✓	✓	✓	✓	✓	✓	✓

answer from one that merely matches the reference. Table 1 summarizes the differences among representative financial benchmarks.

We introduce FINFIRST, a financial benchmark that evaluates final responses through atomically decomposed rubrics covering both answers and their supporting evidence and derivations. These rubrics span three capability groups: (1) *Raw-information acquisition*, assessing whether the response reports the required facts and figures with the correct entity, period, unit, and definition; (2) *Source verification*, assessing whether the cited sources satisfy rubric requirements regarding identity, authority, relevance, and temporal scope; and (3) *Computation and answer formation*, assessing the calculations, reasoning, completeness, and instruction-following demonstrated in the response.

The construction of FINFIRST follows a three-phase pipeline. First, we derive a taxonomy of 18 fields and a coverage blueprint spanning six dimensions from aggregate patterns of real-world financial search demand. In parallel, three senior financial experts compile a source routing map spanning 138 information sources, each annotated along dimensions of authority, task applicability, and geographic coverage. Second, over 50 finance experts author tasks under the guidance of the coverage blueprint, iteratively adjusting and rebalancing to fill under-represented categories. Guided by the source routing map, experts additionally construct a reference package for each question, comprising the final answer, supporting sources, key definitions and data versions, calculation procedures, alternative answers, and numerical tolerances. Each reference package is then decomposed into atomic scoring items across *raw information acquisition*, *source verification*, and *computation and answer formation*. Third, every candidate task undergoes a six-phase quality control process encompassing scope review, independent re-solving, cross-validation and adjudication, rubric auditing, model-based stress testing, and a final consistency check. This pipeline ultimately yields 123 authentic, diverse, and expert-validated tasks with graduated difficulty, ranging from single-source lookups to complex multi-step research requiring planning, cross-source synthesis, and computation.

We evaluate 15 model configurations under a unified toolset comprising Search, Visit, and Python execution. The results yield three main findings. First, even the strongest agents fall short of perfection: Claude-Opus-5 achieves the highest Atomic and Loose Pass scores at 87.59% and 87.61%, respectively, while Strict Pass drops to 69.11%; GPT-5.6-Sol attains the highest Strict Pass rate at 71.54%. The gap between these two metrics suggests that high aggregate scores frequently reflect partially correct research processes rather than consistently complete solutions. Second, *computation and answer formation* consistently lags behind *raw-information acquisition* as a systematic bottleneck across all systems, indicating that successfully retrieving relevant data does not guarantee correct calculation, synthesis, or reporting. Third, comparable 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*. Additionally, the Unsupported-Correct Rate (UCR) metric reveals that, on

average, 17.74% of responses satisfying all answer-layer criteria nonetheless fail at least one evidence-layer criterion. Together, these findings demonstrate the diagnostic value of atomic assessment in pinpointing where financial search agents succeed and where they fall short.

Our contributions are threefold:

- We focus on agentic search tasks in financial domains and extend evaluation granularity from binary answer-correctness to atomic rubrics that jointly evaluate answers and supporting evidence by developing an 18-field taxonomy, a six-axis coverage blueprint, and a routing map of 138 authoritative financial sources grounded in real-world financial scenarios.
- We release FINFIRST, an expert-authored benchmark of 123 tasks with graduated difficulty, constructed by more than 50 finance experts through a six-stage quality-control pipeline, yielding an overall acceptance rate of 9.78%.
- We systematically evaluate 15 model configurations under a unified basic-tool setting, finding that 17.48% of answer-correct responses fail at least one evidence-layer criterion, and that similar aggregate scores can conceal substantially different performance profiles.

2 DATA CONSTRUCTION

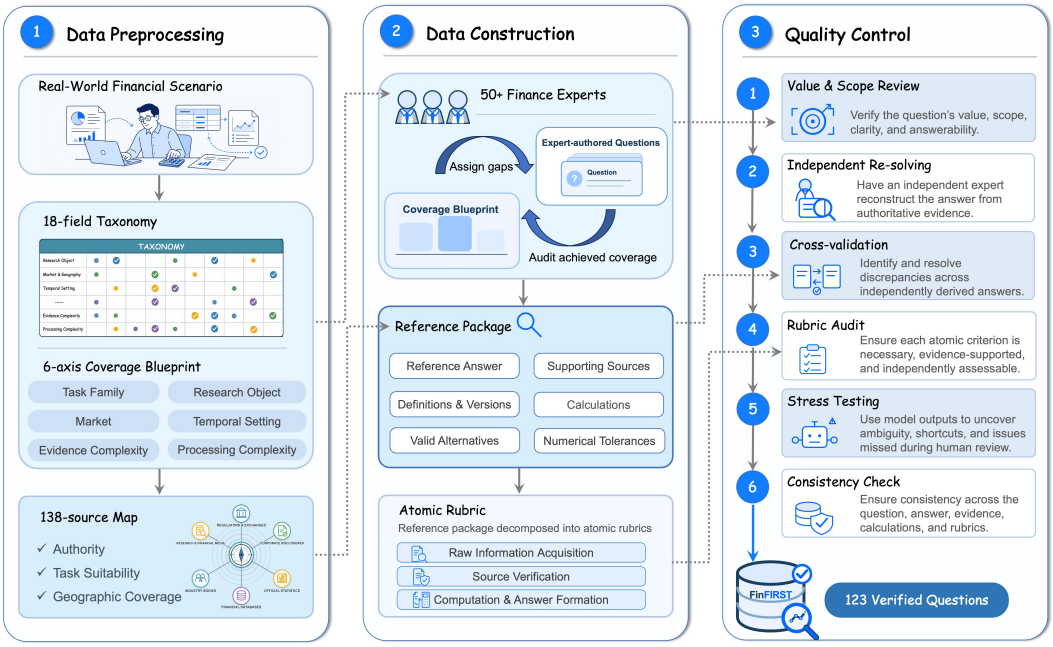

Figure 1: Overview of FINFIRST construction pipeline: Data Preprocessing derives a taxonomy, coverage blueprint, and source map from real-world financial scenarios; Data Construction engages over 50 finance experts to iteratively author tasks, balance coverage, construct evidence-grounded reference packages, and define atomic rubrics; and Quality Control applies a 6-stage filtering procedure, yielding a final set of 123 verified questions.

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 diagnose where an agent succeeds or fails within the research process. As illustrated in Figure 1, these principles are operationalized through three phases: data preprocessing, expert-led data construction, and six-stage quality control—outlined below with further details in Appendix A and an advanced-research example in Appendix B.

2.1 Data Preprocessing

Demand analysis and taxonomy design. We distilled recurring financial needs from real-world financial scenarios into an 18-field taxonomy covering four semantic groups: task structure, research content, processing requirements, and source and retrieval intent. The baseline distribution reflects aggregate demand patterns, while professionally important yet less frequent scenarios are retained when supported by accessible and reproducible evidence. Full taxonomy details are in Appendix A.

Coverage blueprint. Since the full Cartesian product of the taxonomy would yield numerous implausible or redundant tasks, we define a coverage blueprint along six dimensions: task structure, research object, market, temporal setting, evidence complexity, and processing complexity. The latter two dimensions jointly determine task difficulty, ranging from direct or limited-step retrieval to multi-source research requiring definition alignment, calculation, and synthesis.

Financial source map. In parallel, finance experts compile a routing map spanning 138 sources across major markets and research objects, annotated with source identity, accessibility, authority, task suitability, applicable entities and metrics, and geographic coverage. The map prioritizes statutory, official, and first-party evidence while allowing corroborated secondary sources where appropriate. It serves as a retrieval aid rather than an answer key, with task-specific evidence recorded separately in each reference package.

2.2 Expert-Led Data Construction

Question authoring. More than 50 finance experts author questions based on the taxonomy and coverage blueprint, while continuously reviewing the candidate pool to fill coverage gaps and revise or remove redundant entries. Each question specifies the relevant entity, time constraint, definition, unit, output format, and numerical precision, while questions whose difficulty primarily stems from ambiguity, inaccessible evidence, or artificial complexity are excluded.

Reference-package construction. For each candidate question, experts construct an evidence-grounded reference package containing the reference answer, authoritative sources and precise evidence locations, required facts and numerical inputs, relevant definitions and data versions, intermediate calculations or synthesis steps, valid alternatives, and numerical tolerances. For time-sensitive tasks, the package additionally records the reference cutoff and evidence version to preserve reproducibility after the underlying information is updated or restated.

Atomic rubric construction. Each reference package is decomposed into atomic criteria that can be independently evaluated across three capability dimensions: *raw-information acquisition*, which evaluates the retrieval of required facts and values with the correct entity, period, unit, currency, and definition; *source verification*, which evaluates the authority, relevance, and temporal validity of the supporting evidence; and *computation and answer formation*, which evaluates calculations, reasoning, completeness, and instruction following demonstrated in the response.

2.3 Quality Control

Each candidate undergoes the six-stage quality control process: Value and Scope Review, Independent Resolution, Cross-Validation and Adjudication, Rubric Review, Model-based Stress Testing, and Final Consistency Check. Candidates that fail any stage are revised and reassessed or removed, with further procedural details provided in Appendix A.

2.4 Data Statistics

After quality-control filtering, FinFIRST ultimately 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, market and geography, research object, conditional filtering, language, computation, and source specification and count.

The benchmark is bilingual, comprising 60.2% Chinese and 39.8% English tasks, with prompts and reference answers averaging 187.02 and 90.93 characters per task, respectively. Difficulty is distributed across 31.7% easy, 42.3% medium, and 26.0% hard questions. Rubrics span *raw-information acquisition* (59.5%), *source verification* (17.0%), and *computation and answer formation* (23.5%),

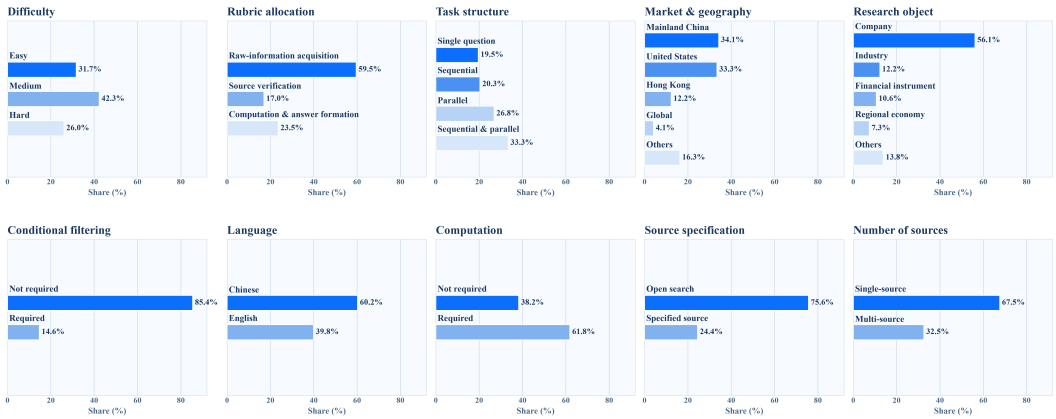

Figure 2: FINFIRST composition across ten dimensions, with long-tail categories grouped as *Others*.

reflecting the benchmark's dual emphasis on information retrieval and evidence-supported answer generation. The final composition balances broad coverage with substantial process complexity: over four-fifths of the questions contain multiple related subproblems; 61.8% require explicit computation; 32.5% draw on multiple sources; 14.6% involve conditional filtering; and 75.6% of questions mandate open search without a specified source. Companies constitute the largest research-object category, with mainland China and the United States contributing comparable shares of geographic coverage.

3 EXPERIMENT

3.1 EVALUATION SETUP

We evaluate 15 model configurations on all 123 tasks in FINFIRST using a unified ReAct (Yao et al., 2023) harness with a shared toolset comprising Search, Visit, and Python execution, thereby minimizing variation from 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 to $T = 1.0$ and report results from a single designated run. Appendix C details the tool interfaces.

3.2 MODELS

The evaluated systems comprise 5 proprietary models, namely GPT-5.6-SOL, CLAUDE-OPUS-5, GEMINI-3.7-FLASH, QWEN3.8-MAX and QWEN3.8-FLASH, along with 10 open-weight ones: DEEPSEEK-V4-PRO-0813, GLM-5.3, GLM-5.3-FLASH, GLM-5.2, KIMI-K3, DEEPSEEK-V4-FLASH-0731, QWEN3.8-27B, MINIMAX-M3, HUNYUAN3, and LING-3.0-FLASH-FIN. Detailed model cards are provided in Appendix D.

3.3 EVALUATION METRICS

Each question q comprises 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:

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

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

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

Atomic is the unweighted pass rate over all 701 criteria. **Loose Pass** grants partial credit proportional to expert-assigned weights, aggregated over the full 12,300-point benchmark. **Strict Pass** counts a task as fully 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*. Group-level scores are 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 a question, its atomic rubric items, and the corresponding model 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. Final scores are computed deterministically as the weighted sum of satisfied items. To validate judge reliability, we randomly sampled 50 evaluation instances for expert annotation. Item-level agreement between GLM-5.1 and the human reference labels yielded Cohen's Kappa κ (Cohen, 1960) = 0.816. The full judge prompt is provided in Appendix E, Table 7.

4 Results and Analysis

4.1 Overall Performance

Table 2 reports performance and tool-use statistics across all 15 model configurations. Claude-Opus-5 achieves the highest scores in both Atomic (87.59%) and Loose Pass (87.61%), while GPT-5.6-Sol leads in Strict Pass rate at 71.54% with 88 of 123 questions fully satisfied. Yet even for top-performing 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 largely correct solutions still frequently leave at least one requirement unsatisfied.

The three rubric groups reveal a consistent downstream bottleneck: *computation and answer formation* underperforms *raw-information acquisition* across all models, with the gap ranging from 5.27 percentage points (GPT-5.6-Sol) to 21.80 points (DeepSeek-V4-Flash). Rubric decomposition further distinguishes models with similar aggregate performance: GLM-5.3 and Kimi-K3 achieve comparable Loose Pass rates (80.61% vs. 80.83%), yet GLM-5.3 excels at *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%).

Tool-use volume does not correlate monotonically with 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 highest single-question peak at 313 calls. Turn counts therefore serve as efficiency indicators only when comparing systems at comparable correctness levels, as higher volumes may reflect broad exploration, or inefficient search rather than superior research outcomes.

4.2 Unsupported-Correct Rate

Binary final-answer correctness alone fails to distinguish a well-supported result from one lacking complete evidence chains. To quantify this distinction, we partition each decomposable question into two layers: an answer layer and an evidence layer. "Correct" refers to satisfaction of all answer-layer

Table 2: Overall, rubric-level, and tool-efficiency results across 15 high-performing models on FINFIRST. ↓ indicates that lower values denote higher efficiency conditional on comparable task correctness. † indicates incomplete results: GPT-5.6-SOL has 117 valid responses, 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 corresponding 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
DEEPSEEK-V4-FLASH	72.90	71.37	48.78	77.49	71.65	55.69	16.49	21.96	106
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
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	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

rubric criteria, rather than an independent assessment of final-answer-only correctness: for calculation questions, it comprises the *computation and answer formation* criteria; for retrieval questions, it comprises the *raw-information acquisition* criteria. All remaining criteria constitute the evidence layer. An answer is marked 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 but with partially creditable supporting content; **Q3**, correct but with incomplete evidence; and **Q4**, correct with all evidence-layer criteria satisfied.

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

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

where the denominator covers all correct-answer attempts. Lower UCR values indicate that such responses more consistently satisfy all evidence-layer requirements in the final response. Across the 1,770 decomposable attempts, 1,150 satisfy all answer-layer criteria, with 201 falling into Q3, yielding a micro-averaged UCR of 17.48% and an unweighted macro-average across models of 17.74%. This suggests that final-answer-only evaluation would overstate fully traceable success for nearly one in six correct attempts. GPT-5.6-SOL achieves the lowest UCR at 8.70%, while GEMINI-3.7-FLASH records the highest at 34.62%. LING-3.0-FLASH-FIN attains a UCR of 10.00%, ranking third lowest overall and lowest among the evaluated open-weight models, indicating relatively consistent satisfaction of evidence-layer requirements among its answer-layer-correct responses.

4.3 SLICE ANALYSIS

Figure 3 reports Strict Pass across four dataset dimensions, with missing outputs counted as failures. Task complexity yields the clearest trend: cross-model averages decline monotonically from 60.51% on easy questions to 52.95% on medium and 47.29% on hard, with thirteen models following this pattern. Source multiplicity produces a comparably consistent effect, as single-source questions average 57.27% versus 46.83% for multi-source questions, with GLM-5.3 being the sole exception. Regarding language dimensions, English questions achieve higher Strict Pass rates than Chinese ones, yet the direction and magnitude of this gap vary across models. Market-level performance is more heterogeneous: Hong Kong leads with 64.89%, followed by the US (56.75%), mainland China (52.22%), and other markets including global tasks (45.33%). Within this dimension, US-developed models hold a clear advantage on US questions (67.48% vs. 54.07%), whereas China-developed

Table 3: 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, credited support	Q3: Correct, partial support	Q4: Correct, complete support	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
DEEPSEEK-V4-FLASH	13	35	12	58	5	17.14
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
QWEN3.8-27B†	7	29	17	65	5	20.73
MINIMAX-M3	20	41	12	45	5	21.05
HUNYUAN3	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

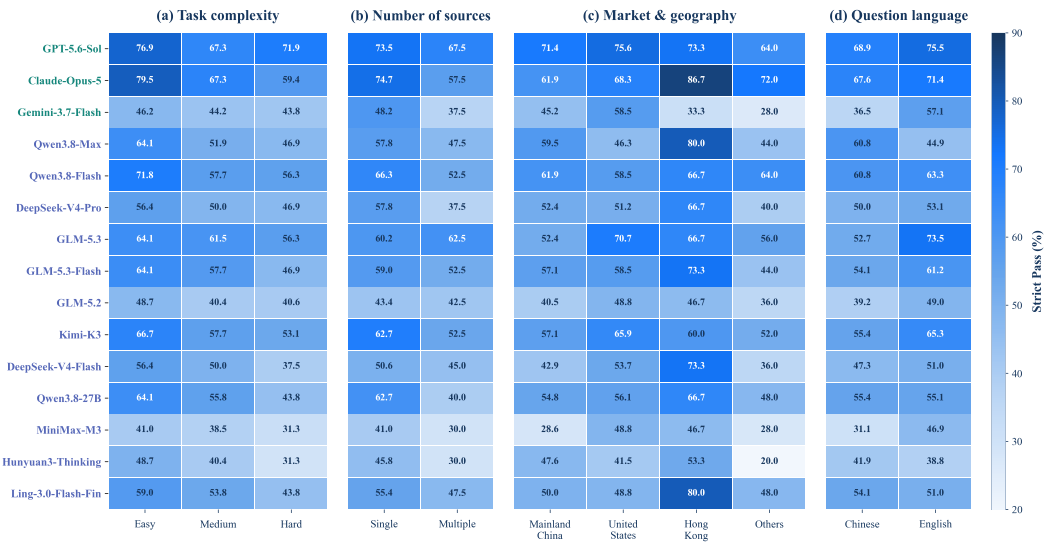

Figure 3: Per-model Strict Pass across task complexity, source count, market and geography, and language. Model names are shown in **Green** for US-developed models and **Blue** for China-developed models.

models show a narrow edge on Hong Kong questions (65.00% vs. 64.44%), potentially attributable to regional corpus coverage, financial terminology, and market conventions.

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, thereby reducing evaluation to retrieval-augmented generation (RAG) rather than measuring end-to-end performance across 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's EDGAR; its atomic rubrics improve upon exact-match scoring, yet the published judge evaluates answer content rather than the quality of the retrieval or reasoning. 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. 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.

These benchmarks offer complementary approaches to financial-agent evaluation. FinFIRST combines broad, demand-informed financial search coverage with atomic criteria organized into raw-information acquisition, source verification, and computation and answer formation under a unified tool environment. Its diagnostic contribution lies in separately assessing the facts, source information, and derivations presented in final responses, enabling traceable and fine-grained financial evaluation.

6 Conclusion

This work aims to advance traceability in financial search tasks by extending evaluation beyond final answer matching to encompass the evidence and intermediate derivations. We introduce FinFIRST—an expert-authored benchmark comprising 123 tasks grounded in real-world financial scenarios. Its construction integrates an 18-field taxonomy, a 6-axis coverage blueprint, a registry of 138 financial information sources, contributions from over 50 financial 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 presented supporting content.

Our evaluation of 15 model configurations under a unified tool environment reveals that current agents remain far from reliably completing financial search tasks. Even the top-performing Claude-Opus-5 achieves the highest Loose Pass score, yet its Strict Pass rate drops by 18.50 percentage points, exposing a persistent gap between partial correctness and end-to-end task completion. Across all models, *computation and answer formation* consistently lags behind *raw-information acquisition*, with 17.48% of responses satisfying all answer-layer criteria fail at least one evidence-layer criterion. These findings demonstrate that final-answer accuracy alone can obscure incomplete evidence and distinct capability bottlenecks. We hope this work lays the groundwork for more measurable and diagnosable agent evaluation, and that the rubrics, dataset, and analytical perspectives will attract broader academic interest in advancing evaluation methodology for financial LLM agents.

AI USE STATEMENT

For FINFIRST, we used LLMs as search agents in our experiments, for model-based stress testing, and as an automated rubric-based judge. Although LLM tools were involved throughout the quality control process, all outputs underwent rigorous manual inspection and adjustment by domain experts. The automated judge was additionally validated against annotations provided by domain experts in finance and we computed its agreement with human annotations to ensure reliability. In addition, we use LLM tools to polish the paper, while the ideas and core content of the paper are entirely developed and written by us. All of our uses of LLM tools comply with the ICLR 2027 policy.

ETHICS STATEMENT

Over 50 finance professionals contributed to FINFIRST construction. Participation was voluntary, with contributors informed of the research purpose and consenting to use of their work for benchmark construction and LLM evaluation. No personal information is included in the released dataset, and any identifying information collected for coordination purposes was removed prior to release. The tasks were derived from aggregate patterns of financial information needs rather than identifiable user queries, and were grounded in publicly accessible sources. FINFIRST is intended solely for research and evaluation purposes and should not be construed as financial or investment advice. We reviewed the licenses and terms of all artifacts used in this study and identified no known conflicts.

REPRODUCIBILITY STATEMENT

We release all 123 benchmark tasks publicly, including their taxonomy labels, evidence-grounded reference packages, and atomic evaluation rubrics. The judge model, GLM-5.1, is explicitly identified, with its complete evaluation prompt provided in Appendix E. The data construction process, quality-control procedures, experimental setup, evaluation metrics, and judging methodology are detailed in Sections 2 and 3; implementation specifics—including tool interfaces, evaluated model identifiers, and additional details—appear in Appendices C, D, and E. All models are evaluated under identical conditions: the same 123 tasks, ReAct-style evaluation harness, tool environment, execution constraints, and failure-handling rules, with temperature fixed at $T = 1.0$. Note that temporal changes in web content, search results, financial disclosures, and hosted model APIs may cause minor variations in reproduced results.

REFERENCES

Anthropic. Claude Opus 5 system card. Official system card, 2026. URL 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. URL 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 (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300. URL 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 (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6279–6292, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.421. URL https://aclanthology.org/2022.emnlp-main.421/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. URL 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. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

DeepSeek-AI. DeepSeek V4 technical documentation and model card. Official technical documentation, April 2026a. URL 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. URL 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. URL 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. URL 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 (eds.), *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)*, pp. 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. URL 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. URL https://arxiv.org/abs/2509.13160.

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

inclusionAI. Ling-3.0-Flash-Fin model card. Openrouter model card, 2026. URL 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. URL https://arxiv.org/abs/2311.11944.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL 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. URL 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. URL 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. URL 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. URL 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 (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 710–725, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL 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. URL 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. URL https://arxiv.org/abs/2501.18062.

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

OpenAI. GPT-5.6 system card. OpenAI Deployment Safety Hub, 2026. URL 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. URL https://arxiv.org/abs/2506.14028.

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

Qwen Team. Qwen3.8-Flash model documentation. Qwen Cloud documentation, August 2026b. URL 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. URL 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 (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2322–2335, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.148. URL https://aclanthology.org/2022.emnlp-main.148/.

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. URL https://arxiv.org/abs/2504.12516.

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. URL 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. URL 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. URL 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. URL https://arxiv.org/abs/2210.03629.

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

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

Z.ai. GLM-5.3-Flash: Frontier intelligence, flash cost. Official research release, August 2026c. URL 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. URL https://arxiv.org/abs/2608.11683.

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 (eds.), *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)*, pp. 3277–3287, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

A Data Construction Details

A.1 Data Preprocessing

Taxonomy design. The 18-field taxonomy is organized into four semantic groups. *Task structure* captures question decomposition, dependencies among subquestions, 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 required number of sources, and the retrieval intent, including data retrieval, fact retrieval, evidence localization, and concept identification. These groups provide an organizational summary rather than a rigid four-tuple: every question is annotated under a uniform schema, and individual fields may be multi-valued or inapplicable. The resulting representation supports stratified sampling, coverage auditing, and fine-grained capability analysis.

Coverage blueprint. The coverage blueprint focuses on six dimensions that directly affect 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 retrieval from multi-source research. Processing complexity ranges from direct extraction and conditional filtering to financial calculation and multi-step synthesis. Easy questions generally require direct or limited-step retrieval; medium questions additionally require alignment, verification, or straightforward calculation; and hard questions combine multiple research steps, sources, definitions, periods, calculations, or synthesis requirements.

Source-map composition. The source 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. Each source is annotated with its identity, access path, description, authority level, applicable entities and metrics, task suitability, and geographic coverage. For each task, experts jointly consider the research object, metric, market, temporal requirement, and verification objective when selecting candidate sources. The source map only supports routing; the exact documents, evidence locations, and acceptable alternatives for a task are specified in its reference package.

A.2 Expert-Led Data Construction

Question authoring. Questions are newly authored by finance experts and labeled using the same schema employed during demand analysis. Throughout construction, the distribution of candidate questions is compared with the target coverage blueprint. Experts are directed to address documented gaps, while redundant or overrepresented candidates are revised or removed. Each prompt must define the research target with sufficient precision, including the entity, temporal range or cutoff, financial or statistical definition, unit, expected output, and numerical precision whenever these attributes affect the answer.

Reference-package construction. Each item-specific reference package contains:

- the reference answer;
- authoritative documents or URLs and precise evidence locations, such as the filing, section, page, table, or data-series identifier;
- the facts and numerical inputs required to answer the question;
- the relevant definitions, reporting periods, units, currencies, consolidation scopes, and data versions;
- the intermediate calculations or synthesis steps connecting the evidence to the final answer; and
- acceptable alternative sources, equivalent definitions, and numerical tolerances.

For time-sensitive tasks, the package also records the reference cutoff, publication or filing date, access date, and evidence version, allowing the reference answer to be reproduced after the underlying information has been updated, revised, or restated.

Atomic-rubric construction. Every atomic criterion must be explicitly required by the question, supported by the reference package, and sufficiently specific to evaluate without unstated assumptions. Each criterion belongs to exactly one of the three capability groups. When financially equivalent solutions exist, the rubric records valid alternative evidence paths, equivalent definitions, and explicit numerical tolerances. This design rewards valid intermediate work without treating a single reference path as the only acceptable solution, while preventing unsupported answers from receiving full credit merely because they match the final value.

A.3 Quality Control

Value and scope review. Reviewers verify that each 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 each question from scratch by 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 is compared with the original reference package. 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 every atomic criterion is 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 absent from the prompt, and any revision to the question, answer, evidence, calculation, or tolerance must be propagated across all related 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 identify residual ambiguity, outdated evidence, omitted alternatives, brittle scoring conditions, and unintended shortcuts. Model outputs are used only for issue detection and difficulty calibration; 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 annotations describe the same research task and evidence requirements.

B CASE STUDY

Table 4 presents a comparison of the portfolio option-adjusted spreads of two target-maturity corporate bond ETFs. The reference answer uses 50.77 basis points for the State Street My2027 Corporate Bond ETF and 32.27 basis points for the iShares iBonds Dec 2027 Term Corporate ETF, both dated July 29, 2026. Their difference is $50.77 - 32.27 = 18.50$ basis points. The question explicitly requests fund identification, a common observation date, issuer-source attribution, both portfolio-level spread values, the subtraction, and a final answer with the required direction, unit, and precision. The example illustrates three distinct assessment dimensions. *Raw-information acquisition* assesses fund identification and the reported input values. *Source verification* assesses the stated date and source attribution against the illustrative reference package. *Computation and answer formation* assesses the calculation and final presentation. A response reporting only the correct numerical difference therefore leaves explicit task requirements unsatisfied. These judgments concern the final response; they do not independently verify retrieval actions or the contents of external webpages.

C 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.

C.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.

Table 4: A constructed ETF spread-comparison example with explicit task requirements and seven scoring criteria. This example is not included in the 123-task benchmark.

Question. Using primary issuer disclosures published on or before July 31, 2026, compare the portfolio option-adjusted spreads of the investment-grade State Street My2027 Corporate Bond ETF and the iShares iBonds Dec 2027 Term Corporate ETF. Use the latest observation date available for both funds within that publication cutoff. Identify both funds by name, state the common observation date, and report each fund's own portfolio-level spread in basis points, rather than its benchmark spread or another portfolio metric. For each input, identify the issuer and the disclosure containing the value. Show the subtraction of the iShares spread from the State Street spread, and state which fund has the wider spread and by how much, in basis points rounded to two decimal places.

Illustrative reference evidence. The reference package stipulates two hypothetical primary-source records: `S1`, a State Street fund-characteristics disclosure reporting a portfolio option-adjusted spread of 50.77 bps for the investment-grade State Street My2027 Corporate Bond ETF; and `S2`, an iShares portfolio-characteristics disclosure reporting 32.27 bps for the iShares iBonds Dec 2027 Term Corporate ETF. Both records are stipulated to have been published by July 31, 2026, with July 29, 2026 as the latest common observation date available within that cutoff.

Reference answer. The latest common observation date within the cutoff is July 29, 2026. The investment-grade State Street My2027 Corporate Bond ETF has a portfolio option-adjusted spread of 50.77 bps, attributed to State Street's fund-characteristics disclosure (`S1`). The iShares iBonds Dec 2027 Term Corporate ETF has a portfolio option-adjusted spread of 32.27 bps, attributed to iShares' portfolio-characteristics disclosure (`S2`). The difference is $50.77 - 32.27 = \mathbf{18.50}$ bps. The State Street fund's portfolio spread is therefore wider by 18.50 basis points.

Capability	Atomic criterion	Points
Source verification	States July 29, 2026 as the latest common observation date available within the July 31, 2026 publication cutoff, and applies that date to both inputs.	10
Source verification	Attributes the State Street input to State Street's fund-characteristics disclosure and the iShares input to iShares' portfolio-characteristics disclosure, corresponding to `S1` and `S2`. Naming only an issuer without identifying the disclosure does not satisfy this item.	7
Raw-information acquisition	Identifies both requested funds by name, distinguishing the investment-grade State Street My2027 Corporate Bond ETF from the State Street My2027 High Yield Corporate Bond ETF. Unambiguous equivalent fund names are accepted.	15
Raw-information acquisition	Reports 50.77 bps as the State Street fund's own portfolio option-adjusted spread, rather than a benchmark spread or another portfolio metric.	15
Raw-information acquisition	Reports 32.27 bps as the iShares fund's own portfolio option-adjusted spread, rather than a benchmark spread or another portfolio metric.	15
Computation and answer formation	Explicitly shows the subtraction of the iShares spread from the State Street spread, yielding $50.77 - 32.27 = 18.50$ bps. An equivalent calculation in consistently converted units is accepted. A correct difference without the requested calculation does not satisfy this item.	30
Computation and answer formation	Presents the final spread difference in basis points to two decimal places and correctly states that the State Street fund has the wider spread. Numerical correctness of the difference is assessed by the preceding item.	8
Total		**100**

Scoring note. Each criterion corresponds to an explicit requirement in the question and is evaluated as a binary item without within-item partial credit. The calculation item accepts equivalent units, whereas the final-presentation item requires basis points to two decimal places. The source labels `S1` and `S2` identify hypothetical records in this illustrative reference package; they are not external citations and need not appear verbatim in the response. Source-related judgments assess response compliance with the stipulated reference evidence, not independent verification of historical disclosures or retrieval actions.

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

[1] `https://serpapi.com/`

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

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

C.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.

D 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.

E Judge Prompt Template

Table 7 provides the Judge Prompt Template; Chinese queries use a corresponding Chinese version. For each instance, the placeholders are populated with the evaluation date, question, atomic rubric items, and model response. The judge observes only the final response—not hidden reasoning or

[2] https://jina.ai/

Table 6: Model cards for the systems evaluated in FINFIRST. A locked icon denotes an API-only proprietary model, whereas an unlocked icon denotes a model with publicly downloadable weights.

MODEL	PROVIDER	CHECKPOINT	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-ai/DeepSeek-V4-Pro-0813`	🔓	1.6T / 49B active	DeepSeek-AI, 2026a
DEEPSEEK-V4-FLASH-0731	DeepSeek-AI	`deepseek-ai/DeepSeek-V4-Flash-0731`	🔓	285B / 13B active	DeepSeek-AI, 2026b
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
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	Tencent	`tencent/Hy3`	🔓	295B / 21B active	Hunyuan Team, 2026
LING-3.0-FLASH-FIN	InclusionAI	`inclusionAI/Ling-3.0-flash-Fin`	🔓	124B / 5.1B active	inclusionAI, 2026

tool trajectories—and does not independently access cited pages. Source-verification and traceability scores therefore reflect whether the response satisfies the corresponding rubric requirements, not whether cited URLs are accessible, cited pages support the reported values, or relevant documents were available before the cutoff. Instructions to "verify" refer to checking the response against the supplied rubric. Recorded trajectories are analyzed separately for tool-use statistics and qualitative case studies and do not affect the main rubric scores.

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, set the `passed` field to the JSON boolean `true` if the item is fully satisfied, and to `false` otherwise.
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; original item numbers and point values have been removed from the rubric text. Output `id` values are assigned by input order, starting at `1`. The `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.