BIGFINANCEBENCH: A WORKFLOW-GROUNDED BENCHMARK FOR FINANCIAL-RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial-research answers are decision-relevant only when another analyst can audit how they were produced: which source was chosen, which period and accounting definition were used, which assumptions were made, and how the calculation was performed. Existing finance benchmarks largely evaluate isolated subskills or final answers, leaving the auditable derivation itself under-measured. We introduce BIGFINANCEBENCH, a 928-item expert-authored benchmark of open-ended financial-research tasks in which each item pairs a ground-truth reference answer with a point-weighted rubric that decomposes the derivation into independently checkable steps. BIGFINANCEBENCH is *workflow-grounded* in that it rates the whole derivation, rather than just the final output. Across 36,241 rubric points, the benchmark supports partial-credit evaluation and localization of failures across the analyst workflow. Evaluating ten current frontier and open-weight agents, we find substantial headroom: the best system reaches only 58.8% rubric score, final-answer accuracy is a useful but lossy proxy for derivation quality, and model capability varies non-uniformly across financial workflows.

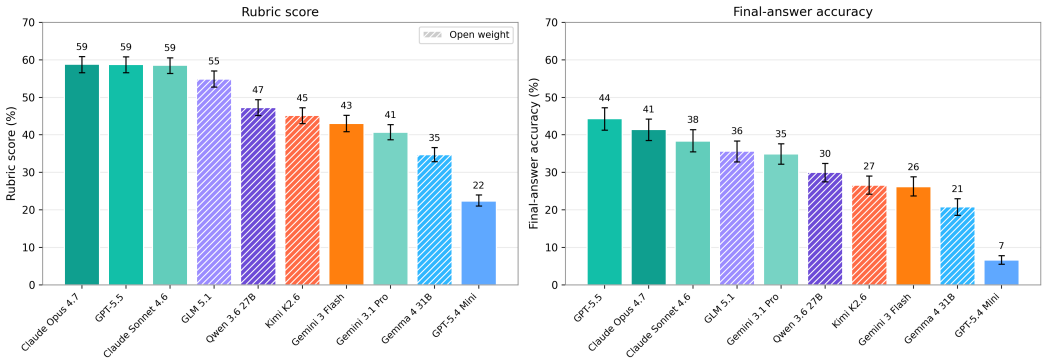


Figure 1: Model performance on BIGFINANCEBENCH: point-weighted rubric score (left) and final-answer accuracy (right). Error bars show 95% bootstrap confidence intervals over questions; hatched bars are open-weight systems.

1 INTRODUCTION

Evaluation is shifting from static knowledge tests toward economically valuable professional work, graded by the standards practitioners apply to one another (Miserendino et al., 2025; Patwardhan et al., 2025; Vidgen et al., 2025; Wang et al., 2025; Mazeika et al., 2025; Akyürek et al., 2025). Financial research is a natural target for this agenda. Analyst work informs investment decisions, diligence, reporting, governance, and regulatory review; an incorrect source, missed adjustment, or unsupported assumption can materially alter conclusions. Yet financial research is not merely a final-answer task. Consider an analyst asked whether management-adjusted EBIT is overstated if capitalized software development is treated as a real cost. A reviewer must be able to verify the filing, fiscal period, GAAP-to-non-GAAP bridge, treatment of amortization, and adjustment formula behind

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.

the reported figure. In practice, the value of the answer depends on whether another analyst can audit that derivation. Doing that work requires an analyst to identify and disambiguate entities, find primary sources, extract line items, normalize conventions, apply accounting adjustments, and compose a defensible number. Recent work has separately developed rubric-based grading for open-ended answers and training signals (Deshpande et al., 2025; Arora et al., 2025; Gunjal et al., 2025; Huang et al., 2025), making it possible to evaluate such work beyond the level of a single final answer.

Existing finance benchmarks provide many of the domain ingredients for this agenda: numerical reasoning over a pre-selected page (Chen et al., 2021; Zhu et al., 2021; Reddy et al., 2024; Zhao et al., 2024), retrieval ranking over filings (Choi et al., 2025b;a), single-fact lookup against documents or structured databases (Islam et al., 2023; Kim & Huang, 2026; Wang et al., 2024), holistic NLP-task aggregation (Xie et al., 2024), long-form regulatory explanation (Chen et al., 2024), and end-to-end agentic evaluation against expert-authored questions (Bigeard et al., 2025). These abstractions are valuable because they isolate retrieval, reading, calculation, and domain knowledge. However, they leave two parts of the work under-specified. First, practicing analysts often receive open-ended, multi-source, assumption-dependent questions unlike typical benchmark tasks that pair a query with a single discrete input, such as one page, filing, retrieved chunk, or database row. Second, the work is reviewed as a derivation, not only as a final number: source choices, time periods, line items, adjustments, formulas, and conclusions must be verifiable by another practitioner. A benchmark for economically valuable finance work therefore needs both questions authored by practitioners and grading units that correspond to the auditable steps of the derivation.

These two mismatches motivate a single research question:

> *Can financial-research agents be evaluated on practitioner-authored tasks and at the derivation-level resolution used in expert review?*

We introduce BIGFINANCEBENCH as our answer: a benchmark of 928 full-workflow financial-research questions written by practicing financial analysts and former investment-banking practitioners, each audited by a separate reviewer before inclusion. The questions reflect the open-ended, multi-source, assumption-dependent tasks these annotators perform in practice, requiring systems to chain retrieval, normalization, accounting judgment, and calculation rather than solve a closed-form problem over pre-selected context. We use *workflow-grounded* to mean that tasks preserve this open-ended, multi-source, assumption-dependent structure of analyst work and are graded at the resolution used in expert review. BIGFINANCEBENCH is intended as a frontier evaluation over realistic workflow families, not as a frequency-weighted census of analyst workloads. To grade these questions at the resolution of analyst review, each item pairs a ground-truth reference answer with a point-weighted rubric derived from the analyst workflow. Rubric criteria decompose the derivation into independently verifiable steps, including entity identification, source selection, line-item retrieval, accounting adjustment, formula construction, and final synthesis. At evaluation time, judges apply these criteria to the visible agent trajectory, including tool calls, tool outputs, calculations, and the submitted answer, so the benchmark grades auditable workflow evidence rather than only the final response.

Contributions. We make three contributions:

- **Practitioner-authored, auditable finance workflows.** We introduce 928 practitioner-authored financial-research questions designed to require source selection, accounting judgment, and calculation over public evidence.
- **Derivation-level evaluation.** Each item includes a point-weighted rubric over independently checkable workflow steps, yielding 15,656 criteria and 36,241 total points for partial-credit evaluation of analyst derivations.
- **Empirical diagnosis of current agents.** Across ten frontier and open-weight agents, we show that rubric scores expose partial progress hidden by final-answer accuracy, that much of the remaining between-model separation appears before a clean setup is reached, and that model strengths vary across workflows.

Table 1: At-a-glance positioning of the closest benchmark families. Prior work provides finance-specific tasks, agent harnesses, or rubric-scored professional outputs; BIGFINANCEBENCH combines all three with derivation-level grading of auditable analyst workflows.

Benchmark	Finance tasks	Agent harness	Rubric-scored	Derivation-level
FinanceBench (Islam et al., 2023)	✓	✗	✗	✗
FinRetrieval (Kim & Huang, 2026)	✓	✓	✗	✗
Finance Agent Benchmark (Bigeard et al., 2025)	✓	✓	✓	✗
FinDeepResearch (Zhu et al., 2025)	✓	✓*	✓	✗
FinResearchBench (Sun et al., 2025)	✓	✓*	✓	✗
ProfBench (Wang et al., 2025)	✗	✗	✓	✗
PRBench (Akyürek et al., 2025)	✓	✗	✓	✗
BIGFINANCEBENCH	✓	✓	✓	✓

* Evaluates deep-research agents, but the exact execution harness and tool surface are not fully specified.

2 RELATED WORK

Previous finance datasets. Finance benchmarks have moved from curated financial excerpts toward richer source and tool settings. Early QA datasets evaluate table-and-text arithmetic over earnings reports and filings, with answers expressed as numbers, programs, or normalized spans (Chen et al., 2021; Zhu et al., 2021; Reddy et al., 2024); FinanceMath adds explicit financial concepts and formula use (Zhao et al., 2024). Filing-centered benchmarks study open-book QA, evidence retrieval, filing-type selection, and passage ranking over public company disclosures (Islam et al., 2023; Choi et al., 2025a;b). Structured-data settings evaluate Text2API generation and agent access to fundamentals databases (Wang et al., 2024; Kim & Huang, 2026). Broader finance suites expand the surface area to multi-task financial NLP, long-form regulatory QA, multilingual and multimodal finance reasoning, filings-plus-market-data reasoning, and ReAct-style financial research agents (Xie et al., 2024; Chen et al., 2024; Xue et al., 2024; Agrawal et al., 2026; Bigeard et al., 2025). FinDeepResearch extends this line to multilingual company-report generation under an expert-designed structure with expert-verified references (Zhu et al., 2025). FinResearchBench offers an interesting complementary approach, extracting a post-hoc logic tree of arguments and evidence from each report to score its structure and support relations (Sun et al., 2025). Professional high-finance experience remains uncommon among benchmark contributors, which can limit data quality and motivated BIGFINANCEBENCH's practitioner-led construction.

Economically valuable and professional task evaluation. Outside finance, evaluation has similarly moved from knowledge tests toward work products and professional judging. MMLU and GPQA (Hendrycks et al., 2020; Rein et al., 2023) evaluate broad academic knowledge and graduate-level expert science questions with multiple-choice accuracy. Economically grounded benchmarks evaluate paid software-engineering tasks and proposal selection (Miserendino et al., 2025), deliverables from occupations across major GDP sectors (Patwardhan et al., 2025), and end-to-end remote freelance projects (Mazeika et al., 2025). Rubric-based professional benchmarks evaluate expert-authored prompts with source documents and criterion-level grading: APEX covers investment banking, consulting, law, and medicine; ProfBench covers scientific and business report generation; and PRBench covers high-stakes finance and legal reasoning (Vidgen et al., 2025; Wang et al., 2025; Akyürek et al., 2025). Table 1 summarizes the closest comparison set.

Rubric scoring and harnesses. Recent benchmarks vary along two methodological axes: how the model is run, and how its work is scored. Rubric-scored evaluations decompose open-ended work into criterion-level judgments: MultiChallenge uses binary instance-level questions, HealthBench uses physician-written criteria with positive and negative point values, and APEX, ProfBench, and PRBench use expert-authored criteria for professional tasks (Deshpande et al., 2025; Arora et al., 2025; Vidgen et al., 2025; Wang et al., 2025; Akyürek et al., 2025). Separately, agent harnesses run the model in a loop: the model observes state, chooses whether to call a tool, receives the tool output, and repeats until it submits a final answer or deliverable. SWE-Lancer, GDPval, RLI, BrowseComp, FinRetrieval, and Finance Agent Benchmark instantiate this pattern with shells, editors, browsers,

computer-use interfaces, web search, SEC search, or structured-data APIs (Miserendino et al., 2025; Patwardhan et al., 2025; Mazeika et al., 2025; Wei et al., 2025; Kim & Huang, 2026; Bigeard et al., 2025). Recent tool-use benchmarks also move beyond final answers by evaluating agent trajectories, tool calls, and intermediate process quality (Lu et al., 2025; Nath et al., 2025; Lù et al., 2025). These are close methodological precedents, but they target generic agent behavior or judge training rather than expert financial-research workflows. FinResearchBench reconstructs a logic tree after generation and FinDeepResearch scores a fixed report schema; both are close finance-specific report-scoring precedents, but neither evaluates the agent's retrieval and calculation trajectory. Related training work converts rubrics into reward signals, using prompt-specific criteria for on-policy reinforcement learning and reward-shaping (Gunjal et al., 2025; Huang et al., 2025).

3 THE BIGFINANCEBENCH DATASET

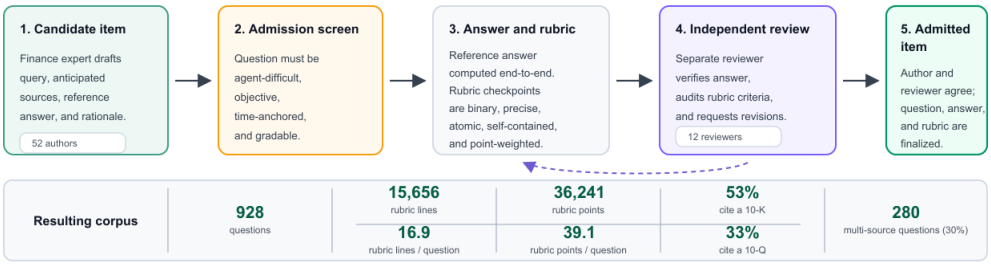


Figure 2: Dataset construction workflow and resulting-corpus statistics for BIGFINANCEBENCH.

3.1 DATASET CONSTRUCTION

The dataset contains 928 items authored by 52 financial-research subject-matter experts, predominantly current and former investment bankers and private equity investors, with smaller representation from equity research and adjacent buy-side roles. A separate panel of 12 reviewers audited the items before inclusion. Each row records the query, anticipated sources, reference answer, and weighted rubric. All item authors and independent reviewers were finance professionals required to have high-finance backgrounds, so item authoring and review drew directly on practitioner experience. The dataset was assembled between September 2025 and March 2026.

Challenging. Candidate questions were crafted to be purposefully difficult for existing agents. Before submission, authors tested each query against multiple frontier agents and systems then recorded the observed failure mode(s). Reviewers accepted questions whose difficulty came from depth of reasoning, including niche expertise, multi-step methodology, or domain-specific judgment. They rejected questions that merely stacked many simple lookups.

Objective. Candidate questions had to be objectively gradable: the answer had to be fixed by expert consensus under a legitimate methodology, anchored to a specific point in time, and expressed with explicit units and rounding rules. Non-obvious assumptions were stated in the query, and rubric criteria were limited to steps reviewers deemed necessary for a perfect answer, not artifacts of a particular solution path.

Authoring. For each item, the author produced a ground-truth reference answer and a point-weighted rubric. Reference answers followed the methodology a practicing analyst would use, with intermediate values carried through the calculation and conventional rounding applied only to the final figure. Rubrics were governed by five authoring rules:

- **Binary.** Each criterion is a yes/no determination, not a nuanced evaluation.
- **Precise.** Each criterion specifies the exact value and time period; the document is not separately tagged unless citing a specific source is itself part of the correct answer.
- **Atomic.** Each criterion isolates a single concept so that partial credit can be awarded fairly.
- **Self-contained.** Given only the candidate answer and a single rubric criterion, the grader can grade the criterion without reading any other criterion.

Table 2: Example BIGFINANCEBENCH items with reference answers and rubric excerpts. Rubrics are clipped if necessary, with the full rubric reported in Appendix Table 5.

Question	Reference answer	Rubric excerpt
If I take Dayforce's management adjusted reported EBIT as is, would it be overstated or understated or the same last year if I think capitalized software expense is a real cost? If so, by how much?	Overstated by $90.1m of excluded capitalized software development costs. Adj EBIT was unburdened by any amortization of capitalized software. Subtracting capitalized software development spend, the adjusted management figure of $410.5m becomes $320.4m.	Full rubric: 30 lines, 91 points; first 10 lines shown. [+1] Identifies DAY as ticker [+2] Identifies Fiscal Year Ended December 31 2024 as the latest year [+4] Identifies $410.5m as mgmt's reported Adj. EBIT for FY24 [+2] Identifies $104.1m as GAAP EBIT for FY24 [+1] Identifies stock-based compensation of $156.6m as a component of the difference between GAAP and non-GAAP EBIT [+2] Identifies amortization of $120.0m as a component of the difference between GAAP and non-GAAP EBIT [+1] Identifies "Other" of $29.8m as a component of the difference between GAAP and non-GAAP EBIT [+2] Identifies that stock-based compensation does not include any amortization of capitalized software added back [+1] Identifies that "Other" includes $19.8m of restructuring expenses [+1] Identifies that "Other" includes $9.0m of fair value adjustments *Clipped: remaining 20 rubric criteria omitted; full entry in Appendix Table 5.*
Udemy (UDMY) Business NDRR vs. ARR Growth: Calculate the implied YoY % Q3'26 ARR and ARR increase for Udemy Business using Q3 2025 financials, assuming the aggregate Net Dollar Retention Rate (NDRR) converges to the current 'Large Customer' benchmark and New Logo ARR contributes a constant $25M. Round the final answer to two decimal places	1.74%	Full rubric: 9 lines, 32 points; all lines shown. [+1] Identifies Q3 '25 Actual ARR as $527.20M [+3] Identifies "Large Customer" NDRR as 97% [+2] Calculates Q3 '26 Retained ARR as (Q3 '25 ARR) * (Large Customer NDRR) [+5] Calculates Q3 '26 Retained ARR as $511.38M [+1] Identifies new logo growth as a $25M [+2] Calculates Q3 '26 ARR as Q3 '26 Retained ARR + new logo growth [+5] Calculates Q3 '26 ARR as $536.38M [+3] Calculates ARR Growth as $9.18M [+10] Calculates YoY % ARR Increase as 1.74%, when rounded to two decimal places

- **Weighted.** Each criterion carries an integer weight on a 1 to 10 scale, with the final answer always carrying a meaningful share of total points.

Review. Each item was routed to a separate finance expert, who independently verified the reference answer and audited the rubric for the five authoring rules and rounding tolerances. Authors revised items until both parties agreed, after which the item was admitted to the benchmark. In the preserved review logs, 81% of items (748 of 928) contain substantive feedback or reviewer-applied edits, consistent with meaningful audit rather than perfunctory sign-off.

Figure 2 summarizes the construction workflow and resulting corpus, Table 2 shows representative items, and Appendices G and E provide summary statistics and a full annotated trajectory; answers are intentionally brief, while the difficulty lies in producing the auditable derivation captured by the point-weighted rubric. Release details are summarized in Appendix A.

Dataset composition. Figure 3 shows that BIGFINANCEBENCH is centered on core public-equity research workflows: earnings quality, M&A, valuation, and operating KPIs. The heatmap indicates that these workflows do not collapse to a single retrieval pattern, since short-answer items still require cross-filing synthesis, forward/scenario reasoning, accounting adjustment, and external market data. The appendix coverage figure (Figure 8) shows broad but uneven sector coverage, with heavier representation of TMT and diversified questions. Different workflows demand different mixes of analyst work. We categorize each rubric criterion into one of three analyst stages (*Retrieval*, *Definition*, *Calculation*) using the regex rules detailed in Appendix I. By rubric points, the stage mix is 29.5% Retrieval, 5.8% Definition, 63.2% Calculation, and 1.5% Other. Appendix Figure 9 reports the corresponding per-workflow stage mix, and Appendix Figure 8 reports sector coverage.

4 BENCHMARK EVALUATION

4.1 SETUP

Agent harness. We evaluate current frontier and open-weight models in a common open-book ReAct-style harness with a 50-step budget. At each step, the model may call one or more tools or submit a final answer. Agents receive only the question; they never see the reference answer or rubric.

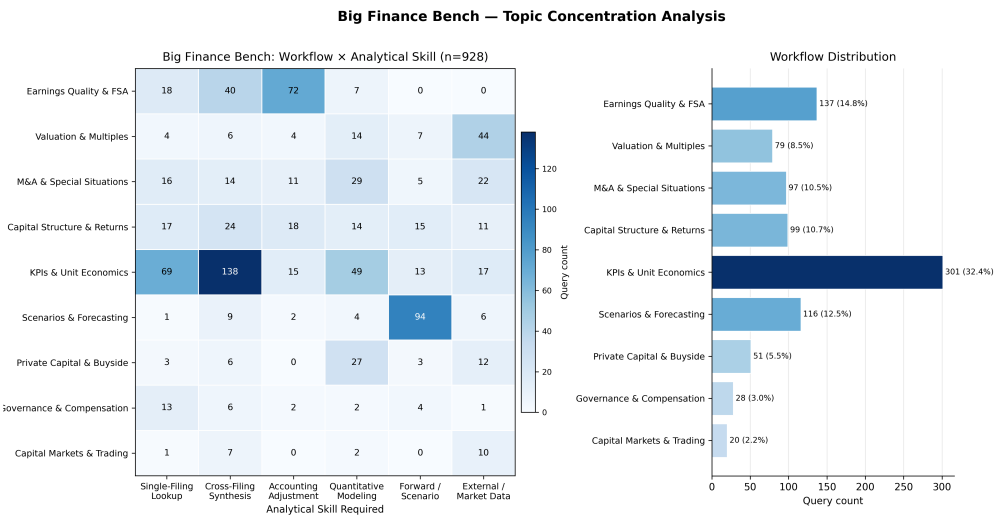


Figure 3: Workflow composition of BIGFINANCEBENCH. (a) Joint distribution of analyst workflow type against the analytical skill each question requires. (b) Marginal distribution over workflow types. See Appendix G for sector coverage and question difficulty indicators.

All model inference uses commercial hosted APIs; open-source models are accessed through Vercel Model Gateway. No on-site compute workers are used for benchmark evaluation.

Tool surface. The tool set is intentionally minimal and public-source only. `web_search` supports broad discovery, `edgar_search` targets SEC filings, `fetch_url` retrieves the selected source, and `python_exec` provides a sandbox for calculation and tabular manipulation; `final_answer` terminates the run. This design tests whether agents can identify and use the relevant public evidence rather than whether they can query a proprietary financial database.

Judging and metrics. For each completed trajectory, two independent judges (Gemini 3.1 Pro Preview and Claude Opus 4.7) receive the question, visible agent trace, reference answer, and full point-weighted rubric, and then grade correctness for each rubric criterion and the final answer. The primary metric is *rubric score*: earned rubric points divided by total rubric points, macro-averaged across questions. We also report *final-answer accuracy*, the fraction of final answers judged correct, and *rubric criterion pass rate* (the unweighted fraction passed). Each model is run for three trials per question. Reported metrics are computed on the paired set of $(\text{question}, \text{trial})$ records that received a valid grade from both judges: they average the two judges within each (question, trial), average trials within question, and then macro-average over questions. Per-model paired coverage ranges from $2{,}776$ to $2{,}783$ of the $2{,}784$ expected, or 99.7% to 100%. Inter-judge agreement is high (Cohen's $\kappa \in [0.952, 0.973]$ on final-answer correctness across the ten models); see Appendix J. On 30 public questions, relative to $n = 3$, aggregate rubric score changes by -0.38 pp at $n = 5$ and -0.23 pp at $n = 10$, while final-answer accuracy changes by 0.00 and $+0.29$ pp (Appendix J). A third judge, GPT-5.5, regrades all $27{,}840$ trajectories, yielding pairwise Cohen's $\kappa = 0.962$–0.972 on final answers and 0.803–0.852 on rubric criteria; the three-judge mean reproduces the original ranking exactly ($\tau = 1.000$). On the full benchmark, point-weighted rubric scores and the unweighted criterion pass rate yield similar model rankings (Kendall's $\tau = 0.90$; Table 4, Appendix O). Additional rollout and grading diagnostics are reported in Appendices J and O.

Evaluation validation. Three complementary checks support the evaluation design. A finance-expert audit of 400 non-error fetches found 86.0% readable and only one parser-garbled; truncated outputs could be refetched with a targeted query or larger token budget. In an independent audit, nine finance professionals re-performed 90 tasks authored by other practitioners; under the same two-judge rubric procedure, their solutions earned a 90.2% two-judge rubric score (an estimated 91.7% after correcting for derivation steps lost when spreadsheet workpapers were flattened to text

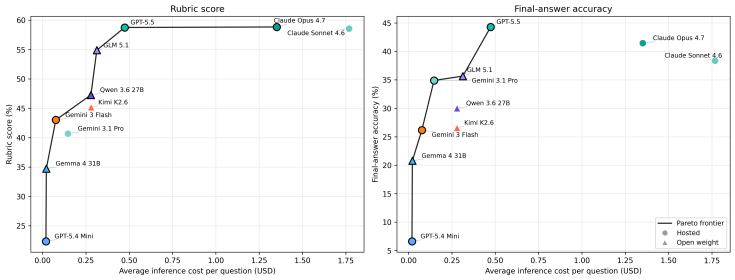


Figure 4: Performance versus cost frontier on BIGFINANCEBENCH. Each point plots a model's score against average inference cost per question, computed from trace token counts and provider list prices. Lines mark the metric-specific Pareto frontiers.

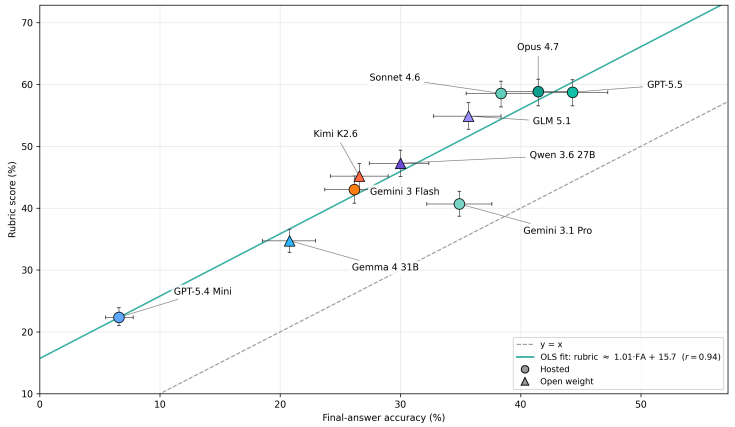


Figure 5: Per-model rubric score against final-answer accuracy. Every system sits above the $y = x$ diagonal (dashed grey), indicating that rubric grading credits intermediate workflow progress that final-answer grading discards. The teal line summarizes the overall trend; Gemini 3.1 Pro is the most visible deviation. Error bars are 95% bootstrap confidence intervals over questions.

for grading; Appendix M) and 98.9% final-answer accuracy. A blind expert regrade of 30 model trajectories found that the two-judge mean differed by only 0.4 pp in aggregate and correlated at Pearson $r = 0.946$ with expert scores. Appendix M reports the full studies.

4.2 RESULTS

BIGFINANCEBENCH remains far from saturated: the headline results figure (Figure 1) shows that the best systems remain below 60% rubric score and below 45% final-answer accuracy. Newer systems evaluated under the same protocol through September 2026 reach at most 67.1% (Appendix B), so substantial headroom remains. The top models are tightly clustered on rubric despite clearer separation in final-answer accuracy (Figure 4; Appendix Table 4). We observe the same hardness at the question level: rubric scoring preserves a broad middle band of partially solved questions, whereas final-answer scoring collapses many of these items into a larger hard tail. Appendix K reports the corresponding drill-downs, including the full heatmap (Figure 14) and source-document conditioning (Figure 13). The performance-cost frontier also depends on the metric, so optimizing for final answers and optimizing for auditable derivations need not select the same model.

Figure 5 shows that rubric and final-answer scores are strongly aligned but systematically separated. Every model sits above the equality line, with a mean rubric-minus-final-answer gap of 15.95 pp, Pearson $r = 0.94$ across models, and 41 of 45 pairwise rankings concordant. Rubric grading therefore preserves most of the ordering that final-answer accuracy produces while adding derivation-level resolution. One case nevertheless shows the metrics are not interchangeable: Gemini 3.1 Pro and

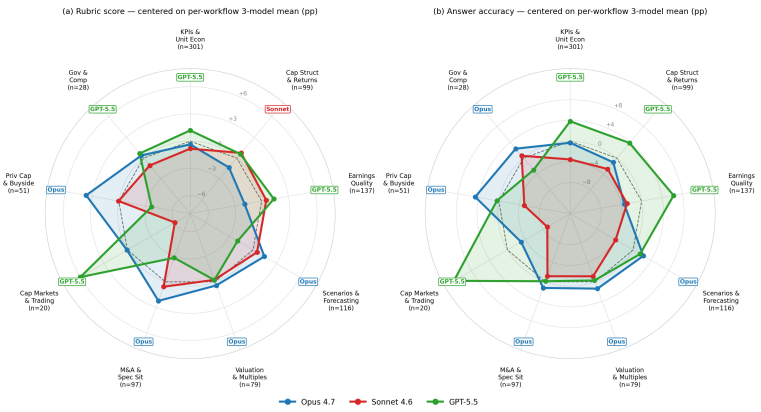


Figure 6: The headline tie among the top-3 closed models is a partition: each leads on different analyst workflows. Per-axis values are the model's mean rubric score (left) or final-answer accuracy (right) *relative to the per-workflow average across the three models*. For visual clarity, we center each workflow axis by the mean of the three models, so the polygons highlight relative specialization rather than shared workflow difficulty. Polygons do not nest in either metric. Absolute-form polygons are reported in Appendix Figure 16.

Gemini 3 Flash invert across the two metrics: Pro is 8.7 pp above Flash on final-answer accuracy, but Flash is 2.3 pp above Pro on rubric score.

Leading models specialize across workflows. The headline tie among the top-3 closed models (Opus 4.7, GPT-5.5, and Sonnet 4.6 within 0.3 pp of each other on rubric, Appendix Table 4) masks different workflow strengths. Among these three models, no model earns the highest rubric on more than 37.7% of questions under fractional tie credit (190 of 928 questions are top-3 ties); each leads on roughly a third of the benchmark. Figure 6 shows the same partition by analyst workflow: the polygons do not nest, indicating orthogonal strengths rather than a single shared frontier.

This non-uniform frontier is also useful operationally: a simple router using observable (workflow, source) features improves over the best single model by 7.6% relative rubric gain and 10.7% relative answer-accuracy gain, while a best-of-10 oracle leaves additional headroom (Appendix L).

Calculation score on clean setup. The naive read from Panel (a) of Figure 7 suggests that models are weakest on calculation. Surprisingly, the opposite is the case: Panel (b) separates two margins. The first is whether a model reaches a clean setup: perfect Retrieval and Definition credit. This entry rate remains strongly model-separated, so stronger systems still complete more of the auditable workflow. The second margin is the point-weighted Calculation score conditional on reaching that setup, where scores compress for most systems. This compression is not just a denominator artifact: Appendix F absorbs question identity and finds much larger question-level variation after clean setup, with a p10 to p90 span of 49.7 pp, than residual between-model variation (5.1 pp). A taxonomy of the residual losses identifies convention and task-interpretation mismatches as the dominant failure, accounting for 38.2% of lost rubric points. Input selection, precision, and rounding account for 25.9%, consistent with models failing to maintain a consistent precision lineage across intermediate values. True arithmetic errors account for 1.3% of lost rubric points, consistent with agents having access to Python. Appendix N reports the methodology and full distribution.

5 Conclusion

BigFinanceBench makes a benchmark-methodology claim: finance agents should be evaluated on auditable analyst workflows, not only on isolated QA, retrieval, or arithmetic. It follows the shift toward economically meaningful professional work and rubric-based grading (Miserendino et al., 2025; Mazeika et al., 2025; Arora et al., 2025; Wang et al., 2025; Akyürek et al., 2025) in a domain where the work product is a number plus its derivation: 928 expert-authored questions, 15,656 rubric

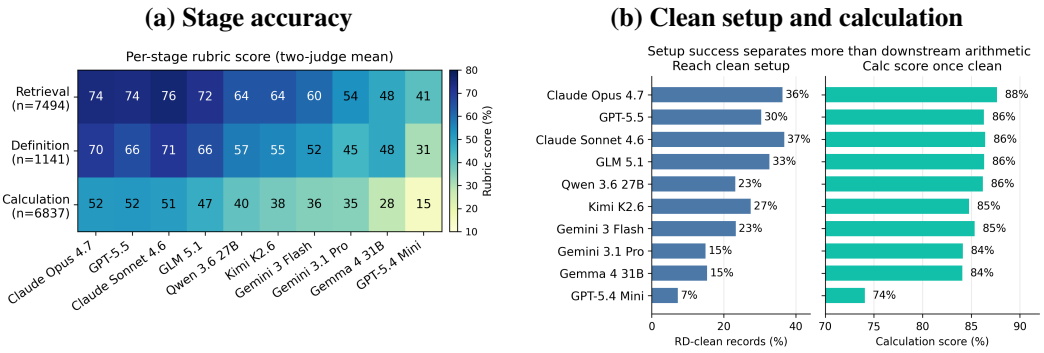


Figure 7: Workflow-stage accuracy and clean-setup calculation. Panel (a) maps each rubric line to a workflow stage by its leading verb and reports the question-level macro-mean within questions containing that stage. Panel (b) reports the share of eligible $(\text{question}, \text{trial})$ records with perfect Retrieval and Definition credit, and the point-weighted Calculation score within that same clean-setup pool.

criteria, and $36{,}241$ rubric points measure whether a system can connect the entity, source, period, definition, assumption, adjustment, and calculation that make a financial answer reviewable.

This capability remains far from solved. As in SWE-Lancer and RLI (Miserendino et al., 2025; Mazeika et al., 2025), current models complete parts of realistic professional tasks but fall short end-to-end: the best systems remain below 60% rubric score and below 45% final-answer accuracy, and the frontier is spiky across workflows. Rubrics complement rather than replace outcome evaluation: as in HealthBench, ProfBench, and PRBench (Arora et al., 2025; Wang et al., 2025; Akyürek et al., 2025), they expose partial progress and, in finance, localize whether a derivation broke at source selection, metric definition, accounting adjustment, or calculation. Final-answer accuracy is a useful but lossy proxy for workflow correctness, retrieval and setup dominate residual failures once arithmetic is instrumented, and capability is jagged across financial workflows, so different models lead on different workflows; how to select the right model for a new financial task remains open.

Limitations. BIGFINANCEBENCH evaluates a static, public-source approximation of analyst work, so scores are not deployment validation for financial decisions, and rubric scores are a high-resolution evaluation instrument rather than literal ground truth: dual judging, expert-authored rubrics, and independent item review mitigate but do not eliminate judge and rubric subjectivity. Models may earn additional rubric credit through unusually verbose traces or by enumerating multiple plausible values without improving the underlying analysis. Judge self-preference remains a limitation despite null-to-negative own-family effects in our three-judge analysis (Appendix O). Additional limitations on scope, judging, and contamination are in Appendix A.1.

Harness code, the released subset, the $1{,}500$ released trajectories, and per-judge grade files from all three judges are available at the anonymous mirrors[1,2]; Appendix A.3 gives the release scope.

AI USE STATEMENT

Generative AI systems are components of the evaluation method, and the paper describes each use. Two LLM judges, Gemini 3.1 Pro Preview and Claude Opus 4.7, grade every trajectory, and GPT-5.5 regrades all trajectories as a third judge (Section 4.1; Appendix O). LLM classifiers assign the workflow and skill labels (GPT-5.5; Appendix H) and the Calculation error taxonomy (GPT-5.6 Terra; Appendix N); the rubric-stage labels come from a regex classifier without an LLM (Appendix I). The evaluated agents are themselves LLMs. In preparing the paper, we used generative AI tools to help draft and edit prose, to write and debug the analysis and plotting code released with the paper, and to run the classification steps above. All questions, reference answers, and rubrics were written by finance practitioners and audited by separate practitioner reviewers; generative AI entered item construction only as a difficulty screen, in which authors tested each candidate question against

[1] https://anonymous.4open.science/r/bigfinancebench-BE84/
[2] https://anonymous-hf.com/a/mebeez9m5rfo/

frontier agents and recorded the observed failure mode (Section 3.1; Appendix C). We did not use generative AI to generate research ideas, to formulate the paper's claims, or to produce synthetic benchmark items. The authors reviewed all AI-assisted text and verified the AI-assisted code against the reported tables and figures. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

BIGFINANCEBENCH items are grounded in publicly available financial materials, chiefly SEC filings, press releases, investor-relations pages, and public market data, and were screened for personally identifiable information and sensitive customer data (Appendix A.3). Item authors and reviewers were finance professionals engaged as paid contractors and compensated at professional hourly rates well above the US minimum wage; they produced finance content rather than serving as study subjects, no personal data was collected beyond professional credentials and authorship attribution, and IRB review was determined not applicable (Appendix C). The released 50-question subset is distributed under CC BY 4.0; the remaining items are withheld to limit training-data contamination, and we will consider bona fide research requests for additional access. Benchmark scores measure a static, public-source approximation of analyst work and are not an endorsement of financial automation or financial, legal, or investment advice. Misuse could support unreliable analysis, overreliance in high-stakes decisions, or market-relevant misinformation, so deployment requires domain-specific validation and expert and compliance review beyond this benchmark (Appendix A.2).

REPRODUCIBILITY STATEMENT

The grading harness code, scripts, tests, Dockerfile, and license, the released 50-question subset with reference answers and rubrics, the per-judge grade files, two human workpaper workbooks, and the result tables are available at a login-free anonymous code mirror[3]. The released subset with Croissant metadata, the traces for all 1,500 released trajectories, the 30 per-judge grade files (ten models $\times$ three judges), the same workpapers, and the result tables are available at a login-free anonymous data mirror[4]. Section 4.1 specifies the harness, tool surface, step budget, model and judge identifiers, trial count, and aggregation rule; Appendix M documents the tool implementations and document parsing, and Appendix J reports stop reasons, inter-judge agreement, and trial reliability. Reported uncertainty is a 95% bootstrap confidence interval over questions (Figure 1; Table 4). The classifiers behind the workflow, skill, and rubric-stage labels are specified in Appendices H and I, and the Calculation error taxonomy in Appendix N. Dataset construction, authoring rules, and the review loop are described in Section 3.1 and Appendix C; release scope, license, and the withheld-set policy are in Appendix A.3.

REFERENCES

Yogesh Agrawal, Aniruddha Dutta, Md Mahadi Hasan, Santu Karmaker, and Aritra Dutta. FinTradeBench: A financial reasoning benchmark for LLMs. *arXiv preprint arXiv:2603.19225*, 2026.

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph, Mohammad Mahmoudi Meymand, et al. PRBench: Large-scale expert rubrics for evaluating high-stakes professional reasoning. *arXiv preprint arXiv:2511.11562*, 2025.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. Finance Agent Benchmark: Benchmarking LLMs on real-world financial research tasks. *arXiv preprint arXiv:2508.00828*, 2025.

[3] https://anonymous.4open.science/r/bigfinancebench-BE84/

[4] https://anonymous-hf.com/a/mebeez9m5rfo/

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. FinTextQA: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6025–6047, 2024.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, 2021.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. FinDER: Financial dataset for question answering and evaluating retrieval-augmented generation. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pp. 638–646, 2025a.

Chanyeol Choi, Jihoon Kwon, Alejandro Lopez-Lira, Chaewoon Kim, Minjae Kim, Juneha Hwang, Jaeseon Ha, Hojun Choi, Suyeol Yun, Yongjin Kim, et al. FinAgentBench: A benchmark dataset for agentic retrieval in financial question answering. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pp. 632–637, 2025b.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, 2025.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, et al. Reinforcement learning with rubric anchors. *arXiv preprint arXiv:2508.12790*, 2025.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Eric Y Kim and Jie Huang. FinRetrieval: A benchmark for financial data retrieval by AI agents. *arXiv preprint arXiv:2603.04403*, 2026.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J Pal, and Siva Reddy. AgentRewardBench: Evaluating automatic evaluations of web agent trajectories. *arXiv preprint arXiv:2504.08942*, 2025.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, et al. Remote labor index: Measuring AI automation of remote work. *arXiv preprint arXiv:2510.26787*, 2025.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. SWE-Lancer: Can frontier LLMs earn $1 million from real-world freelance software engineering? *arXiv preprint arXiv:2502.12115*, 2025.

Vaskar Nath, Pranav Raja, Claire Yoon, and Sean Hendryx. ToolComp: A multi-tool reasoning & process supervision benchmark. *arXiv preprint arXiv:2501.01290*, 2025.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 445–458, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Rui Sun, Zuo Bai, Wentao Zhang, Yuxiang Zhang, Li Zhao, Shan Sun, and Zhengwen Qiu. FinResearchBench: A logic tree based agent-as-a-judge evaluation framework for financial research agents. In *Proceedings of the 6th ACM International Conference on AI in Finance*, 2025.

Bertie Vidgen, Abby Fennelly, Evan Pinnix, Julien Benchek, Daniyal Khan, Zach Richards, Austin Bridges, Calix Huang, Kanishka Sahu, Abhishek Kottamasu, et al. The AI productivity index (APEX). *arXiv preprint arXiv:2509.25721*, 2025.

Mingzhu Wang, Yuzhe Zhang, Qihang Zhao, Junyi Yang, and Hong Zhang. Redefining information retrieval of structured database via large language models. *arXiv preprint arXiv:2405.05508*, 2024.

Zhilin Wang, Jaehun Jung, Ximing Lu, Shizhe Diao, Ellie Evans, Jiaqi Zeng, Pavlo Molchanov, Yejin Choi, Jan Kautz, and Yi Dong. ProfBench: Multi-domain rubrics requiring professional knowledge to answer and judge. *arXiv preprint arXiv:2510.18941*, 2025.

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. *arXiv preprint arXiv:2504.12516*, 2025.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. FinBen: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743, 2024.

Siqiao Xue, Xiaojing Li, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. FAMMA: A benchmark for financial domain multilingual multimodal question answering. *arXiv preprint arXiv:2410.04526*, 2024.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. FinanceMath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12841–12858, 2024.

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 *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, 2021.

Fengbin Zhu, Xiang Yao Ng, Ziyang Liu, Chang Liu, Xianwei Zeng, Chao Wang, Tianhui Tan, Xuan Yao, Pengyang Shao, Min Xu, et al. FinDeepResearch: Evaluating deep research agents in rigorous financial analysis. *arXiv preprint arXiv:2510.13936*, 2025.

A LIMITATIONS, BROADER IMPACTS, AND RELEASE DETAILS

A.1 LIMITATIONS

Four limitations apply. First, BIGFINANCEBENCH is a slice of financial research rather than a complete model of the domain: it is concentrated in public-company, mostly US-listed, English-language source work, which limits conclusions for private-market, non-US, non-English, advisory, diligence, and client-interaction workflows. Second, the benchmark standardizes real analyst work into a static tool-use setting: questions are time-anchored, the harness exposes a fixed public-source tool surface, and agents cannot ask clarifying questions or use proprietary financial databases. This makes systems comparable, but rankings may change under a different interface or live analyst environment. Third, evaluation remains approximate. We mitigate judge noise by grading each trajectory with two independent LLM judges, reporting inter-judge agreement, and manually checking samples for systematic failure modes such as rejection of real-world auditability requirements; nevertheless, judge scores are not ground truth and rubric-criterion credit can differ (Appendix J). Judge self-preference remains a limitation because both headline judges belong to evaluated model families. The three-judge analysis finds no positive aggregate own-family excess after accounting for each judge's overall generosity, with several own-family comparisons negative; this pattern does not rule out model-specific, family-specific, or shared bias (Appendix O). Similarly, rubrics are reviewed by a separate finance expert and intended to contain steps necessary for a perfect answer, but residual subjectivity remains in any finite decomposition of an analyst-correct derivation. Models may also earn additional rubric credit through unusually verbose traces or by enumerating multiple plausible values without improving the underlying analysis. Fourth, the benchmark is vulnerable to contamination and temporal drift. We time-anchor questions to reduce ambiguity from later filings, but public-source facts can enter model training corpora over time.

A.2 BROADER IMPACTS

BIGFINANCEBENCH lowers the barrier for academic and industrial groups to evaluate financial-research agents on auditable workflows, which we view as necessary for safer and more transparent use of such systems. However, benchmark performance should not be read as an endorsement of financial automation or as financial, legal, or investment advice. Misuse could support unreliable analysis, overreliance in high-stakes decisions, market-relevant misinformation, or attempts to influence markets; deployment therefore requires domain-specific validation, expert review, compliance review, and contextual assessment beyond this benchmark.

A.3 RELEASE DETAILS

We release a 50-question subset comprising questions, reference answers, rubrics, the grading harness, the $1{,}500$ released trajectories, and per-judge grade files from all three judges (Gemini 3.1 Pro Preview, Claude Opus 4.7, GPT-5.5) for the $1{,}500$ released trajectories, for academic research. The subset is stratified by workflow and difficulty quartile and preserves full-benchmark model rankings under rubric score (Kendall's $\tau = 0.956$). To test whether the released subset preserves full-benchmark comparisons, we treat every system evaluated after the headline run as a held-out cohort, since the subset was fixed before those systems existed. At the time of the released results file (22 systems), public-subset and full-benchmark rubric rankings agreed at Kendall's $\tau = 0.887$ (12 held-out systems: $\tau = 0.970$); with all 31 systems completed as of September 2026, $\tau = 0.901$ (21 held-out systems: $\tau = 0.924$; Appendix B). Per-model values are in `results/subset_generalization.csv` on the mirrors. All released annotations were author-generated and are distributed under CC BY 4.0. The public subset[5] and evaluation harness[6] are available for academic research. Updated model results will be maintained on a hosted public leaderboard. Items are grounded in publicly available financial materials and were screened for personally identifiable information and sensitive customer data. The full benchmark is withheld at submission time to mitigate benchmark contamination and preserve evaluation validity. We already observe stronger agents searching the web for BIGFINANCEBENCH and its published results during evaluation, making a full public release a direct contamination

[5] `https://anonymous-hf.com/a/mebeez9m5rfo/`

[6] `https://anonymous.4open.science/r/bigfinancebench-BE84/`

channel. We will consider bona fide research requests for additional access under conditions that preserve evaluation validity.

B LEADERBOARD REFRESH

The headline evaluation (Section 4) covers ten systems run in April 2026. We continue to evaluate newly released models under the identical harness, tool set, step budget, three-trial protocol, and two-judge panel, and Table 3 reports every completed full-benchmark run as of September 2026. Thirty-one systems have complete runs. The strongest, Muse Spark 1.1, reaches a 67.1% rubric score and 53.4% final-answer accuracy; nine systems now exceed the best headline score of 58.8%, and the benchmark remains far from saturated. Scores use the same scorer and $1{,}000$-iteration question-level bootstrap as Table 4. The stage, source-type, and cost analyses in Section 4.2 are reported for the headline ten only.

The released 50-item subset was fixed before any of the 21 later systems existed, so they form a held-out cohort for testing whether public-subset rankings track full-benchmark rankings. For these 21 systems, public-subset and full-benchmark rubric rankings agree at Kendall's $\tau = 0.924$ (Spearman $\rho = 0.983$); across all 31 systems, $\tau = 0.901$ ($\rho = 0.978$); for the headline ten, $\tau = 0.956$. The largest rank moves between the full benchmark and the subset are Claude Opus 5 (6th to 12th), Claude Fable 5 (9th to 13th), and GPT-5.5 (11th to 7th), all within the leading cluster whose full-benchmark confidence intervals overlap.

C DETAILED DATASET CONSTRUCTION

The current master release contains 928 items, authored by a panel of 52 subject-matter experts, predominantly current and former investment bankers and private equity investors, with smaller representation from equity research and adjacent buy-side roles, and audited by 12 reviewers. Reviewer involvement is non-trivial: the busiest single reviewer audited over a quarter of the dataset. Each row records the natural-language query, the author's free-text reasoning notes on what makes the question difficult, the anticipated source documents, the reference answer, the weighted rubric, an optional screenshot or hyperlink to a backup source, and the reviewer's name and review log.

Authors were asked to write items that were both *challenging* and *objective*, and these two criteria were operationalized into specific rules that gated admission to the dataset.

Challenging. A query qualified only if a proprietary financial-research agent and at least one other frontier system (Claude, ChatGPT, or Gemini) failed to answer it correctly when the author tested it before submission, and the author recorded the specific failure mode in a free-text notes field. We deliberately distinguished two failure modes and accepted only one. The accepted mode is complexity through *depth of reasoning*: queries that fail because the model lacks the niche expertise, multi-step methodology, or domain-specific judgment to answer correctly (e.g. stripping Ford Credit out of Ford's TEV/EBIT, summing non-employee director shares from a proxy statement, or decomposing a hospital revenue bridge into available patient days, occupancy, and the rate-volume covariance). The rejected mode is complexity through *stacked simplicity*: asking the same simple lookup repeatedly across many entities or periods, where wrong answers come from hallucinations on trivial steps rather than from genuine reasoning failure. Reviewers steered authors away from this pattern, and query diversity was expected to come from varied methodology rather than from changing the company name on a template.

Objective. The objectivity criterion was operationalized as three rules. (i) *Expert consensus*: the reference answer itself had to be objective, meaning any finance expert applying a legitimate methodology would arrive at the same final figure, so open-ended or opinion questions were excluded. For rubric criteria, authors were given a 95-of-100 heuristic to apply when deciding whether a step belonged in the rubric: include a step only if at least 95 of 100 finance experts would agree it belongs in any perfect answer to the query, irrespective of which legitimate methodology the expert applied. This guidance steered authors away from methodology-specific intermediate steps, since many calculations admit multiple valid approaches (e.g., LTM revenue can be derived several ways) and an expert taking a different valid approach would legitimately skip the step. The 95-of-100 standard was a heuristic for authors and reviewers to think with, not an independently tested or audited

Table 3: Full-benchmark results for all 31 systems with completed runs as of September 2026 (928 questions $\times$ 3 trials; two-judge mean of Gemini 3.1 Pro Preview and Claude Opus 4.7). Rubric score shows the 95% question-level bootstrap CI as $\pm$ half-width; Cost/q is the mean provider cost per question-trial in USD; Evaluated is the month the run was launched. † marks the ten headline models of Table 4.

Model	Rubric score	Answer acc.	Cost / q	Evaluated
Muse Spark 1.1	$67.1 \pm 2.0\%$	53.4%	$0.17	Jul 2026
Gemini 3.5 Flash	$65.4 \pm 2.1\%$	42.2%	$0.70	May 2026
Gemini 3.6 Flash	$63.9 \pm 2.1\%$	44.3%	$0.44	Jul 2026
GPT-6 Astra	$63.6 \pm 2.1\%$	49.5%	$0.56	Sep 2026
Gemini 3.7 Flash	$62.6 \pm 2.1\%$	45.3%	$0.12	Sep 2026
Claude Opus 5	$61.8 \pm 2.1\%$	46.0%	$1.71	Jul 2026
Grok 4.5	$60.7 \pm 2.0\%$	41.7%	$0.28	Jul 2026
Claude Fable 5.1	$59.8 \pm 2.0\%$	46.0%	$0.63	Sep 2026
Claude Fable 5	$59.5 \pm 2.1\%$	45.2%	$2.21	Jul 2026
Claude Opus 4.7†	$58.8 \pm 2.1\%$	41.5%	$1.26	Apr 2026
GPT-5.5†	$58.8 \pm 2.1\%$	44.3%	$0.47	Apr 2026
Claude Sonnet 4.6†	$58.5 \pm 2.1\%$	38.4%	$1.68	Apr 2026
Kimi K3	$58.5 \pm 2.0\%$	40.9%	$0.31	Jul 2026
Claude Sonnet 5	$56.1 \pm 2.1\%$	39.4%	$1.84	Jul 2026
GPT-5.6 Sol	$55.7 \pm 2.0\%$	44.2%	$0.19	Jul 2026
Claude Opus 4.8	$55.7 \pm 2.1\%$	40.6%	$1.38	Jun 2026
GLM 5.1†	$54.9 \pm 2.2\%$	35.7%	$0.31	Apr 2026
DeepSeek V4 Pro	$53.3 \pm 2.0\%$	35.6%	$0.04	Jul 2026
GPT-5.6 Terra	$50.6 \pm 2.0\%$	39.6%	$0.11	Jul 2026
GPT-5.6 Luna	$49.9 \pm 2.0\%$	36.1%	$0.04	Jul 2026
MiniMax M3	$48.9 \pm 2.1\%$	31.6%	$0.05	Jul 2026
Qwen 3.6 27B†	$47.3 \pm 2.1\%$	30.0%	$0.28	Apr 2026
Kimi K2.6†	$45.2 \pm 2.1\%$	26.6%	$0.28	Apr 2026
Gemini 3.5 Flash-Lite	$44.8 \pm 2.1\%$	29.0%	$0.04	Jul 2026
Qwen 3.6 Max	$44.2 \pm 2.2\%$	36.8%	$0.49	Jul 2026
Gemini 3 Flash†	$43.0 \pm 2.2\%$	26.2%	$0.08	Apr 2026
Gemini 3.1 Pro†	$40.7 \pm 2.0\%$	34.9%	$0.15	Apr 2026
Nemotron 3 Ultra	$39.2 \pm 2.1\%$	21.9%	$0.64	Jul 2026
Gemma 4 31B†	$34.7 \pm 1.9\%$	20.8%	$0.02	Apr 2026
GPT-5.4 Mini†	$22.3 \pm 1.5\%$	6.6%	$0.02	Apr 2026
Mistral Large 3	$13.0 \pm 1.0\%$	6.5%	$0.02	Jul 2026

threshold. (ii) *Time-anchored*: the query had to reference a specific point in time so the correct answer would not change months later; phrasings like "the latest quarter" or "the last three fiscal years" were disallowed. (iii) *Binary verifiability*: the response had to be gradable yes-or-no against a clear rubric, which in practice meant final answers were currencies, multiples, percentages, or basis points with units and rounding rules stated explicitly in the query itself. Any non-obvious assumption needed to answer the query had to be written into the query text rather than left as something the model was expected to infer.

Authoring. For each item the author produced two artifacts. The *perfect answer* is a definitive response written with the methodology a practicing analyst would actually use, calculated end-to-end without rounding intermediate values and applying conventional rounding only to the final figure. The *rubric* is a weighted checklist of binary criteria governed by five authoring rules:

- **Binary.** Each criterion is a yes/no determination, not a nuanced evaluation.
- **Precise.** Each criterion specifies the exact value and time period (e.g., "share count of $1{,}622{,}843{,}689$ as of AMD's Q2'25 reporting" rather than "share count from the latest filing"); the document is not separately tagged unless citing a specific source is itself part of the correct answer.
- **Atomic.** Each criterion covers a single concept; a 5-year CAGR, for example, is split into separate criteria for each input, the formula, and the final result so that partial credit is awarded fairly.
- **Self-contained.** Given only the candidate answer and a single rubric criterion, the grader can grade the criterion without reading any other criterion; phrasings like "uses the above calculation" or "does the addition correctly" were disallowed because they require cross-criterion context.
- **Weighted.** Each criterion carries an integer weight on a 1 to 10 scale, with the final answer always carrying a meaningful share of total points.

Where intermediate calculations involved repeating decimals or irrational values, authors specified an explicit margin of error and verified that any value in the allowed range still rounded to the correct final answer.

Review. Each item was routed to a separate reviewer, a different finance expert who had not authored the item. The reviewer (i) independently performed the underlying calculation to verify the reference answer rather than taking the author's number on trust, and (ii) audited the rubric for objectivity, atomicity, self-containment, value precision, point distribution, and rounding tolerances. Comments were returned to the author, who edited and resubmitted, and the reviewer re-checked. The loop continued until both parties agreed, at which point the item was marked ready for automated grading and the row was admitted to the benchmark evaluation pool; all 928 rows in the current master release reached this state.

The review log preserved on each row reflects how active this loop was: 81% of items (748 of 928) carry substantive feedback or reviewer-applied edits, with the remaining 19% signed off with pure acknowledgment (e.g. "looks good" or "ready to go"). This figure is a lower bound on the true rework rate: the log preserves only the final state of the comment thread, and authors and reviewers often condensed earlier substantive back-and-forth into a single closing "looks good" line once the item was accepted. Early in the project, reviewer feedback focused on structural rubric construction; over time, as authors internalized the format, feedback shifted toward content, domain depth, point weighting, and technical opinions on methodology, and review served less as quality control and more as a second-expert sign-off.

Author engagement and compensation. Authors and reviewers were not anonymous crowdworkers but finance professionals, including current and former investment bankers, private equity investors, equity research analysts, and adjacent buy-side practitioners, all required to have high finance backgrounds, predominantly in banking and private equity. They were engaged through a dedicated staffing partner and through direct outreach across the authors' and existing experts' professional networks in investment banking, private equity, and MBA programs. Contributors were predominantly US-based, with some having ties to other countries, and were compensated at competitive professional hourly rates well above the prevailing US minimum wage, consistent with the ICLR Code of Ethics. The full set of instructions given to authors is reproduced above in this appendix, including the

Table 4: Model results on BIGFINANCEBENCH ($n = 928$ questions $\times$ 3 trials per model; metrics are the mean across two independent judges, Gemini 3.1 Pro Preview and Claude Opus 4.7). "Rubric score" is the macro-averaged fraction of point-weighted rubric credit earned. "Answer acc." is the fraction of final answers judged correct. "Rubric criteria" is the macro-averaged fraction of unweighted rubric-criterion credit earned. 95% bootstrap confidence intervals (over questions) are within ± 2.3 percentage points for rubric score, ± 3.1 for answer accuracy, and ± 2.2 for rubric criteria.

Model	Rubric score	Answer acc.	Rubric criteria
`Claude Opus 4.7`	58.8%	41.5%	66.6%
`GPT-5.5`	58.8%	44.3%	65.3%
`Claude Sonnet 4.6`	58.5%	38.4%	67.3%
`GLM 5.1`	54.9%	35.7%	63.2%
`Qwen 3.6 27B`	47.3%	30.0%	55.3%
`Kimi K2.6`	45.2%	26.6%	53.7%
`Gemini 3 Flash`	43.0%	26.2%	50.7%
`Gemini 3.1 Pro`	40.7%	34.9%	45.3%
`Gemma 4 31B`	34.7%	20.8%	41.6%
`GPT-5.4 Mini`	22.3%	6.6%	30.4%

admission criteria (*challenging* and *objective*), the five rubric authoring rules, the review loop, and worked examples drawn from previously accepted items; no separate participant-facing instructions or screenshots are reproduced here because authoring took place via private contracting rather than on a public crowdsourcing platform.

Human subjects and risk assessment. The data collection effort did not constitute human subjects research in the usual sense: authors produced finance content (queries, reference answers, rubrics) rather than serving as study subjects, and no behavioral, psychological, biometric, or otherwise personal data was collected from them beyond the professional credentials and authorship attribution preserved for provenance. Working conditions were those of ordinary remote contract work, and no disclosure of medical, financial, or other sensitive personal information was solicited. On that basis, IRB review was determined not applicable and not sought; no risks to participants beyond those of ordinary professional employment were identified, and none were disclosed to participants because none were anticipated.

D ADDITIONAL TABLES AND FIGURES

Table 4 reports the headline results with bootstrap confidence intervals and the unweighted criterion pass rate. Tables 5 and 6 give the complete rubrics for the two example items of Table 2.

E EXAMPLE AGENT TRAJECTORY

Table 7 shows a complete successful trajectory for the Udemy question in Table 2. The run is from Claude Opus 4.7 and illustrates a successful EDGAR path: the agent finds the Q3 2025 8-K, follows the SEC filing directory to the press-release exhibit, extracts the ARR and NDRR inputs, and computes the projection. The harness recorded visible assistant text and tool calls, but no private chain of thought (`thinking=off`). Long URLs and query strings are shortened for readability; every tool call in the trajectory is kept on its own line.

F CLEAN-SETUP CALCULATION FIXED EFFECTS

A complementary way to remove question-mix effects is to keep all RD-clean records and absorb question identity directly. Let R_{mqt}, D_{mqt}, and C_{mqt} denote the Retrieval, Definition, and Calculation scores for model m, question q, and trial t, again pooling the two judges at the rubric-point level. We restrict to records with positive possible points in all three stages and with $R_{mqt} = D_{mqt} = 1$. We exclude GPT-5.4 Mini from this fit because its RD-clean pool is small and its conditional Calculation score is a lower-tail outlier; the fixed-effects model asks whether the middle-pack compression

Table 5: Full Dayforce example item from Table 2. Table 6 gives a third complete item.

Question	Reference answer	Full rubric
If I take Dayforce's management adjusted reported EBIT as is, would it be overstated or understated or the same last year if I think capitalized software expense is a real cost? If so, by how much?	Overstated by $90.1m of excluded capitalized software development costs. Adj EBIT was unburdened by any amortization of capitalized software. Subtracting capitalized software development spend, the adjusted management figure of $410.5m becomes $320.4m.	Full rubric: 30 lines, 91 points. [+1] Identifies DAY as ticker [+2] Identifies Fiscal Year Ended December 31 2024 as the latest year [+4] Identifies $410.5m as mgmt's reported Adj. EBIT for FY24 [+2] Identifies $104.1m as GAAP EBIT for FY24 [+1] Identifies stock-based compensation of $156.6m as a component of the difference between GAAP and non-GAAP EBIT [+2] Identifies amortization of $120.0m as a component of the difference between GAAP and non-GAAP EBIT [+1] Identifies "Other" of $29.8m as a component of the difference between GAAP and non-GAAP EBIT [+2] Identifies that stock-based compensation does not include any amortization of capitalized software added back [+1] Identifies that "Other" includes $19.8m of restructuring expenses [+1] Identifies that "Other" includes $9.0m of fair value adjustments [+1] Identifies that "Other" includes $1.0m of receivables securitization fees [+1] Concludes that the "Other" of $29.8m is comprised entirely of restructuring expenses, fair value adjustments, and receivables securitization fees [+2] Concludes that "Other" does not include any amortization of capitalized software development expense [+4] Notes that the amortization added back to get to non-GAAP EBIT includes some acquisition-related intangible assets [+4] Concludes that this amortization of intangibles must be included in the $120.0m of amortization. [+2] Confludes that further searching has to be done to confirm whether the $120.0m includes capitalized software development expense [+4] Identifies amortization of capitalized software development expense of $70.7m [+3] Calculates 120.0 - 70.7 = 49.3m of other unexplained amortization added back [+3] Identifies that the $120.0m of amortization is comprised of $70.7m of amortization of capitalized software and 49.3m of other unexplained amortization [+2] Updates non-GAAP EBIT bridge to show GAAP EBIT, + stock based comp, plus amortization of capitalized software, plus other amortization, plus other, equals non-GAAP EBIT. [+2] Searches for other sources of amortization in the 10-K that explain the 49.3 of unexplained amortization [+2] Searches other SEC filings for FY24 amortization decompositions [+2] Notes that not enough information is provided to confidently identify the source of the remaining $49.3m of amortization [+1] Notes that Ceridian Trade Name depreciation likely comprised a meaningful amount of the other amortization in FY24 [+7] Confirms that reported non-GAAP EBIT figure is not burdened by amortization of capitalized software expense [+7] Identifies FY24 capitalized software expense of $90.1m [+10] Notes that non-GAAP EBIT is overstated by $90.1m of capitalized software development expense [+7] Subtracts capitalized software expense from non-GAAP EBIT [+5] Concludes that the more appropriate non-GAAP EBIT figure is $320.4m [+5] Cites Dayforce 2024 10K for various figures

survives question adjustment:

$$C_{mqt} = \alpha_q + \beta_m + \epsilon_{mqt}.$$

The question fixed effect α_q gives each question its own baseline Calculation score. The model effect β_m is therefore estimated from within-question contrasts: among records where models reached a clean setup on the same question, which model received more Calculation credit? We weight each record by its available Calculation points, matching the point-pooled conditional score used in Figure 7. The graph over models and questions is connected, so the model effects are comparable on one scale. The fitted design uses 2,909 RD-clean observations. The relevant diagnostic is not overall fit quality but the relative scale of question and model variation after setup is clean. The centered question effects have standard deviation 22.1 percentage points and a p10 to p90 span of 49.7 percentage points, while the centered model effects range from -2.3 to $+2.8$ percentage points, a 5.1 percentage-point span. The conditional calculation pool is therefore dominated by question-level difficulty; after comparing models within the same questions, residual model separation is much smaller.

Table 8 shows the result. Before adjustment, the point-weighted conditional Calculation scores for these nine models span 3.6 percentage points (84.1% to 87.6%). After absorbing question identity, the centered model effects span 5.1 percentage points. Thus question composition masks a small amount of residual middle-pack separation, but it does not turn the conditional result back into the large unconditional Calculation gap. Once setup is clean, the middle systems remain close.

G Dataset Summary Statistics

Figure 8 gives the corresponding appendix marginals: sector coverage is broad but concentrated in TMT and diversified items, while most questions specify precision and many remain single-period lookups despite having compact final answers.

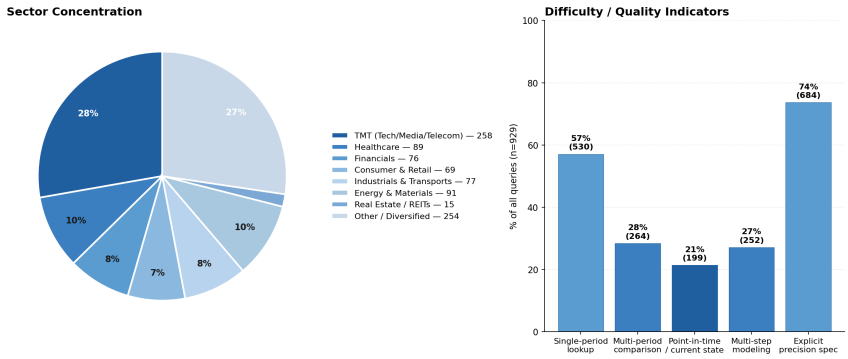


Figure 8: Coverage and difficulty composition of BigFinanceBench. (a) Sector concentration across the 929 authored items; the 928-item benchmark excludes the one item without a reference answer. (b) Per-question difficulty / quality indicators; each item can carry several.

H Workflow / Skill Classification

The workflow and skill axes used by Figure 3 are not authored fields of the dataset; they are reproduced from the question text, reference answer, and rubric by a deterministic LLM classifier. This appendix documents the classifier so the figure is auditable end-to-end.

Taxonomy. The classifier assigns each question exactly one of nine workflow / decision types and exactly one of six analytical skills. The taxonomy strings are pinned in `paper/code/classify_question_workflow_v2.py` and re-imported by the plot script, so the axis labels in Figure 3 are guaranteed to match the enum the classifier was constrained to. The nine workflows are *Earnings Quality & FSA*, *Valuation & Multiples*, *M&A & Special Situations*, *Capital Structure & Returns*, *KPIs & Unit Economics*, *Scenarios & Forecasting*, *Private Capital &*

Buyside, *Governance & Compensation*, and *Capital Markets & Trading*; the six skills are *Single-Filing Lookup*, *Cross-Filing Synthesis*, *Accounting Adjustment*, *Quantitative Modeling*, *Forward / Scenario*, and *External / Market Data*. A one-paragraph definition is supplied to the classifier for each bucket; the full prompt is emitted verbatim by the classifier script.

Model and decoding. The classifier is GPT-5.5 with strict JSON-schema structured output: the workflow and skill enums are the only legal values for the two response fields, so off-taxonomy strings are physically unreturnable. We use the model's default temperature so the structured-output path is honoured uniformly across runs. The system prompt, user prompt, and per-bucket definitions are pinned to a single `PROMPT_VERSION` string that is stamped into the output CSV; mixing prompt versions on resume is refused with an actionable error.

Three-sample voting. Each question is classified three times. The three samples share the same model and temperature, but the order in which the workflow and skill enum values are listed inside the user prompt is shuffled deterministically per (question id, sample index) using SHA-256-derived seeds. Order shuffling is the practical control for recency / primacy bias in prompt enumerations: at the model's default temperature, identical prompts return identical labels, so genuine ambiguity in a borderline question only surfaces if the question is sensitive to the position of the candidate labels. The CSV records all three votes per axis and the modal share; the assigned label is the majority vote, with first-sample tie-break.

Reliability. The classifier produces a complete (workflow, skill) pair for all 928 questions in the released master. Across those 928 questions, the three samples agree unanimously on workflow for 724 (78.0%), reach a 2-of-3 majority for 154 (16.6%), and split evenly across all three votes for 50 (5.4%). The corresponding skill numbers are 711 unanimous (76.6%), 161 majority (17.3%), and 56 tied (6.0%). Borderline questions are therefore concentrated in a small minority of the dataset and are individually identifiable in the published `question_workflow_labels.csv` via the `workflow_agreement` and `skill_agreement` columns.

Reproducing the figure. The code for reproduction will be released together with the paper.

I Rubric-Criterion Stage Classification

Figures 9 and 7 both condition on a per-criterion *stage* label. The labels are produced by a deterministic regex classifier (`paper/code/classify_rubric_stages.py`); we deliberately did not use an LLM here because the rubric prose is highly formulaic and a string-matching rule is auditable criterion-by-criterion. Across all rubric criteria, Retrieval accounts for $7{,}494$ lines (47.9%), Definition for $1{,}141$ (7.3%), Calculation for $6{,}837$ (43.7%), and Other for 184 (1.2%). Point weighting shifts the mix toward Calculation: Retrieval accounts for $10{,}690$ points (29.5%), Definition for $2{,}114$ (5.8%), Calculation for $22{,}892$ (63.2%), and Other for 545 (1.5%). Figure 9 reports the corresponding per-workflow stage mix.

Stages. Each criterion is mapped to one of four stages. *Retrieval* covers the act of locating a specific value in a filing or source (typical leading verbs: *identifies*, *records*, *reports*, *indicates*, *notes*, *states*). *Definition* covers declaring a formula or methodology (*defines*, *formula*, *methodology*, *equation*, *assumes*). *Calculation* covers arithmetic and quantitative transformations (*calculates*, *computes*, *sums*, *adds*, *subtracts*, *multiplies*, *divides*, *adjusts*, *normalises*, *derives*, *estimates*). *Other* is the catch-all residual.

Matching procedure. Each line is processed in two passes. First, a leading adverb in the closed set $\{$*correctly*, *properly*, *accurately*, *appropriately*, *successfully*, *clearly*, *explicitly*, *finally*$\}$ is stripped so "Correctly identifies …" matches as "identifies …". A leading prepositional phrase of the form "For *noun phrase*," is also stripped so "For CFO Jane Doe, reports adoption date …" matches as "reports adoption date …". This handles a Governance & Compensation authoring convention that would otherwise leak the post-comma verb into the residual. Second, the first whitespace-delimited token (lower-cased) is checked against an ordered list of verb sets in the order *Calculation*, *Definition*, *Retrieval*; the first set that contains the token wins. Author typos for high-frequency verbs (e.g.

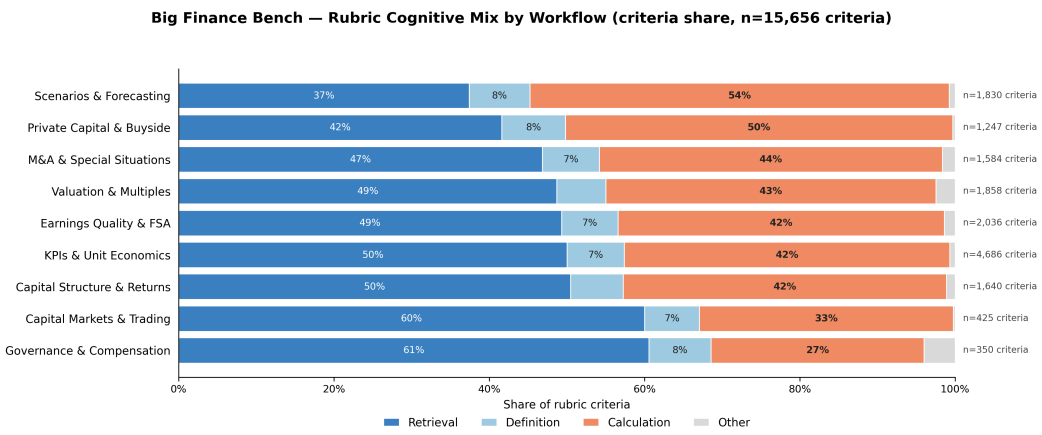


Figure 9: Rubric-criterion stage mix by workflow classified as *Retrieval*, *Definition*, *Calculation*, or *Other*.

idenfities, *identifes*, *calcuates*) are listed alongside the correctly-spelled forms so they do not leak into the residual. Criteria whose leading token is in none of the verb sets land in *Other*.

Distribution. Across the 15,656 rubric criteria in the released master, the rules assign 47.9% of *criteria* to Retrieval, 7.3% to Definition, 43.7% to Calculation, and 1.2% to Other. Weighted by rubric *points* the distribution shifts substantially: Retrieval drops to 29.5% of points while Calculation rises to 63.2%, reflecting the authoring convention that conclusion-style calculation criteria carry the largest `[+N]` weights. The Other residual at 1.2% of criteria is dominated by leading words that appear once or twice in the dataset (e.g. *utilizes*, *includes*, *excludes*); none of the residual leading words exceed 0.13% of criteria individually, so widening the verb sets to absorb them would not change either figure's per-stage shares by 0.1 pp.

Reproduction. `python code/classify_rubric_stages.py` writes one row per `(qid, line_idx)` to `paper/data/rubric_line_stages.csv` (committed alongside the rest of the post-processed CSVs). The classifier is content-based and has no random state; re-running on the same dataset is byte-identical.

J Harness Diagnostics

ReAct stop-reason distribution. Figure 10 breaks down the share of trials terminating in each ReAct stop state: `final_answer` (the model called the terminal tool), `max_steps` (it exhausted the 50-step budget without submitting), or `no_tool_call` (it produced text without calling any tool, so the harness recorded the assistant text as the answer). Two qualitatively distinct failure modes emerge. *Looping*, defined by a high `max_steps` share, accounts for 33% of Kimi K2.6 trials, 34% of Gemini 3 Flash trials, and 21% of GLM 5.1 trials; these systems often run out of budget without converging on an answer. *Skipping the workflow*, defined by a high `no_tool_call` share, is the dominant Gemma 4 31B failure (83% of trials) and the dominant GPT-5.4 Mini failure (46% of trials), and explains why those systems' final-answer accuracy in Table 4 is substantially below their rubric score: when they do submit an answer, it is generated from prior knowledge rather than from the workflow the rubric is designed to check.

Inter-judge agreement. The headline metrics in Table 4 are the mean across two independent LLM judges (Gemini 3.1 Pro Preview and Claude Opus 4.7). Figure 11 reports Cohen's κ on the binary final-answer-correctness verdict between the two single-judge graders as a robustness check on the averaging. Per-model κ ranges from 0.952 (Claude Opus 4.7 and Claude Sonnet 4.6, which tie at the floor) to 0.973 (Gemma 4 31B), all comfortably above the 0.81 "almost perfect" threshold of Landis & Koch (1977). On final-answer accuracy the two judges differ by under 1.3 percentage

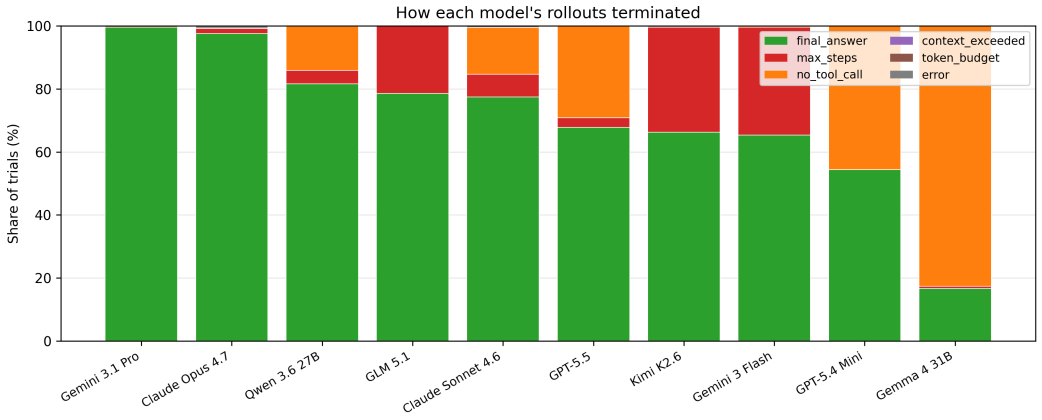


Figure 10: ReAct rollout stop-reason distribution per model. `final_answer` indicates the model called the terminal tool; `max_steps` that it exhausted the 50-step budget; `no_tool_call` that it produced text without invoking any tool. Models are sorted by `final_answer` share descending.

points on every model. On rubric score the disagreement is larger and asymmetric: averaged over (question, trial) pairs, the Opus judge marks +1.9 pp more credit than Gemini for Claude Opus 4.7 and +0.9 pp for Claude Sonnet 4.6, but +5.3 pp for GPT-5.5, which is enough for GPT-5.5 to rank third on rubric under the Gemini judge alone and first under the Opus judge alone. The mean of both judges, which is what the headline metric reports, is the natural way to be robust to that single-judge sensitivity.

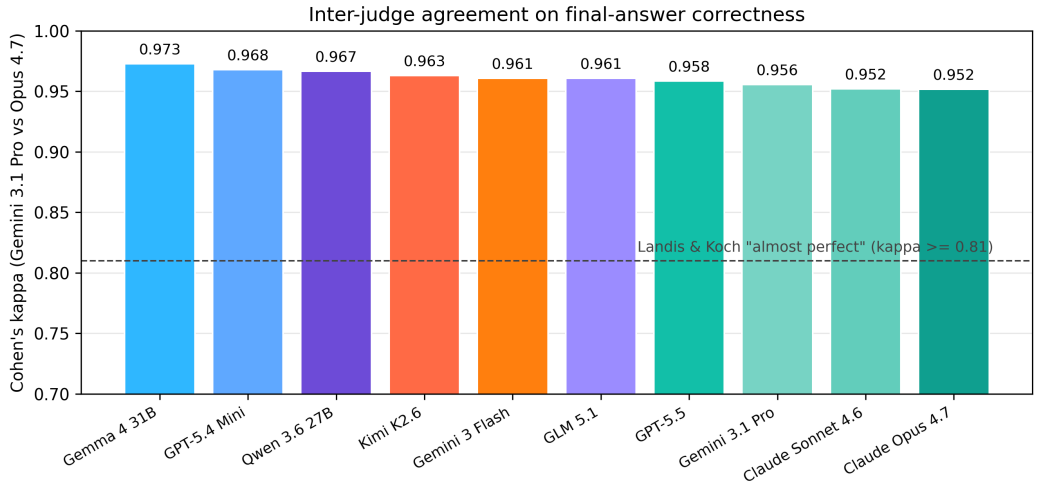


Figure 11: Per-model Cohen's κ between the two judges (Gemini 3.1 Pro Preview and Claude Opus 4.7) on final-answer correctness, paired by (question, trial). All values exceed the 0.81 "almost perfect" threshold of Landis & Koch (1977).

Within-question trial reliability. Inter-judge κ controls for judge noise on a single trial; the complementary axis is rollout noise across the $n = 3$ trials per question for a fixed judge pair. Figure 12 reports the median per-question rubric SD across trials, with the interquartile range as error bars, sorted by ascending median. The two Anthropic systems are most consistent (3.2 pp Opus, 3.5 pp Sonnet); GLM 5.1, GPT-5.5, and GPT-5.4 Mini cluster around 5 pp; the open-weight middle and Gemini 3 Flash sit between 7.3 and 7.8 pp; Gemini 3.1 Pro is the noisiest at 10.6 pp. After the $n = 3$ trial averaging and macro-averaging across the 928-question benchmark, the standard error

contributed by this rollout noise falls below 0.25 pp for every model, so rollout noise is small relative to broad between-model differences.

A direct rollout-count ablation expands 30 public questions from three to five and ten trajectories. The estimates are nested: $n = 3$ uses the original trials, $n = 5$ adds two new trials, and $n = 10$ uses all trajectories; Claude Opus 4.7 and GPT-5.5 grade every trajectory. Relative to $n = 3$, aggregate rubric score changes by -0.38 pp at $n = 5$ and -0.23 pp at $n = 10$, while final-answer accuracy changes by 0.00 and $+0.29$ pp. Across models, the mean absolute rubric-score change is 0.76 pp at $n = 5$ and 0.96 pp at $n = 10$, with maxima of 2.81 and 1.50 pp. For final-answer accuracy, the corresponding means are 1.17 and 1.21 pp and the maxima are 2.11 and 2.44 pp.

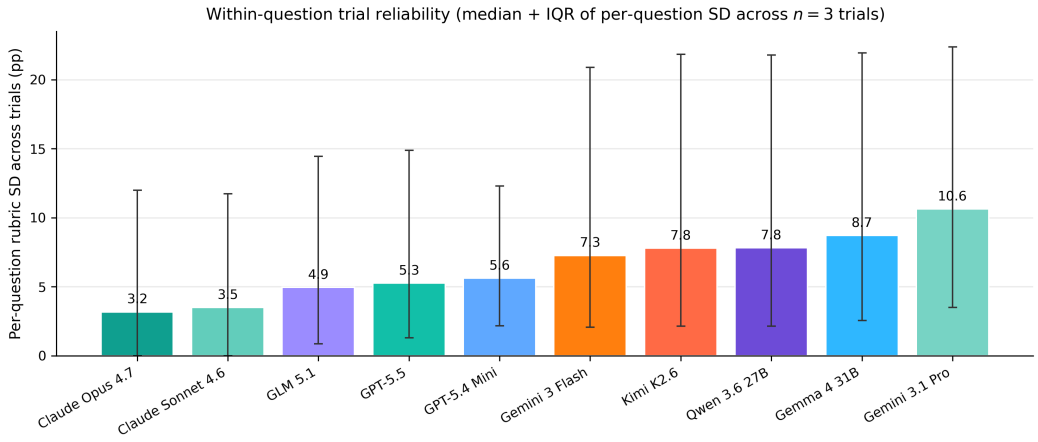


Figure 12: Per-question rubric-score standard deviation across the $n = 3$ trials per question (two-judge mean per trial). Bars are medians across questions; error bars span the 25th to 75th percentiles. Models are sorted by ascending median.

K Conditional Drill-Downs

Source-document conditioning. Different question types are grounded in different SEC filing types (10-K, 10-Q, 8-K, proxy/S-1) or non-SEC sources (press releases, IR pages, web pages, market data). Figure 13 conditions the rubric score on the source bucket the question requires (questions that require multiple source types appear in each bucket they cite). Model rank ordering is largely preserved across buckets: the three top systems (Claude Opus 4.7, GPT-5.5, Claude Sonnet 4.6) are within 3 percentage points of each other in each of the four largest buckets (10-K, 10-Q, 8-K, press release / IR), and within each model the per-bucket rubric across those four buckets varies by under 8 percentage points. The one consistent gap is between 10-K- and 8-K-grounded questions: 8-K rubric scores sit 6 to 9 percentage points below 10-K scores for every evaluated model in the top half of the table (Opus 59 $\rightarrow$ 51, GPT-5.5 60 $\rightarrow$ 54, Sonnet 60 $\rightarrow$ 52, GLM 5.1 56 $\rightarrow$ 47, Qwen 3.6 27B 48 $\rightarrow$ 40), consistent with 8-K events being less structurally standardised than periodic filings. Smaller buckets (proxy/S-1, web) are not large enough to support firm conclusions.

Per-question difficulty structure. Figure 14 shows the full question-by-model score surface under rubric and final-answer scoring.

L Routing-Amenable Frontier

Figure 15 shows achievable rubric and final-answer accuracy at increasing routing granularity, computed across the ten evaluated models. Solid bars are deployable routers that select a model from observable question features (workflow, skill, source bucket); the hatched rightmost rung is a best-of-10 upper bound that selects, per question, the model with the highest score on the metric in question (rubric-best for the rubric number, answer-best for the answer-accuracy number). The upper

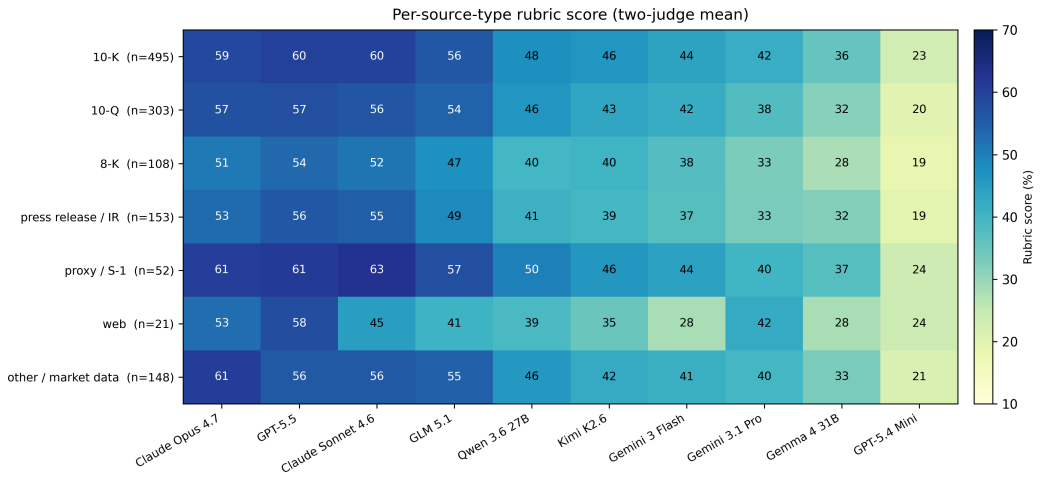


Figure 13: Per-model rubric score conditioned on the source-document type the question requires. Source buckets follow the appendix taxonomy; questions can appear in multiple buckets. Models are sorted left-to-right by overall rubric score; cell values are percentages.

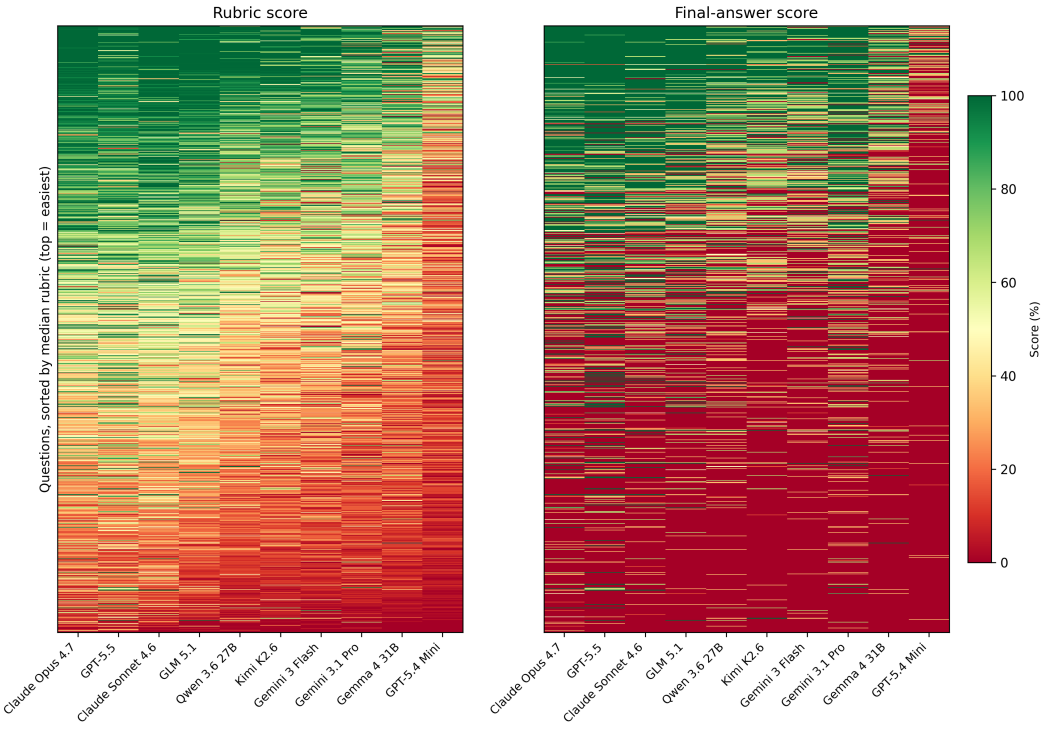


Figure 14: Per-(question, model) score, rubric (left) vs final-answer (right). Each row is one question; rows are shared between panels and sorted by per-question median rubric (top = easiest). Each column is one model, sorted by overall rubric. Cell colour is the per-(question, model) two-judge mean score across $n = 3$ trials; final-answer scores live in $\{0, \frac{1}{6}, \frac{2}{6}, \ldots, 1\}$ and rubric scores are continuous in $[0, 1]$.

bound is not a deployable system since it requires the metric grade itself to identify the per-question winner. Coarse (workflow $\times$ source) routing buys $+4.5$ pp rubric / $+4.7$ pp answer accuracy over the best single model; the best-of-10 upper bound is $+13.2$ pp / $+13.4$ pp. Both metrics agree that the

achievable frontier on BIGFINANCEBENCH is meaningfully wider than any single evaluated model contributes, and that the gap between feasible coarse routers and the best-of-10 upper bound (∼9 pp on each metric) is the headroom available to a learned router that conditions on question features.

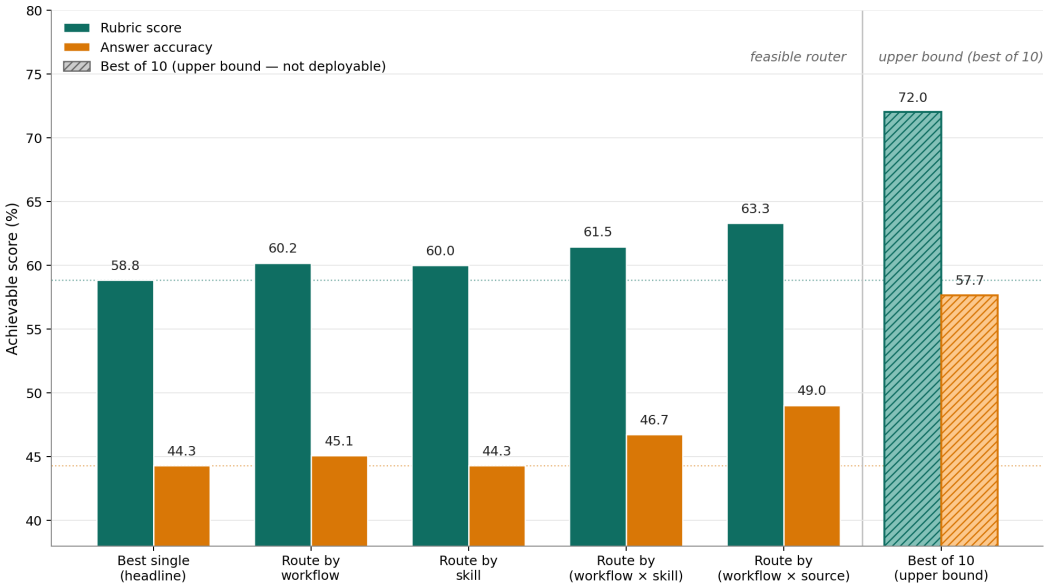


Figure 15: Routing ladder: achievable rubric (teal) and final-answer accuracy (orange) at increasing routing granularity, computed across the ten evaluated models. Solid bars are deployable routers using observable question features; the hatched rightmost rung is a best-of-10 upper bound that requires ground-truth selection and is therefore not deployable. Dotted reference lines mark headline best-single performance.

Figure 16 replots the per-workflow polygons of Figure 6 in absolute mean rubric and final-answer accuracy. Polygons cluster tightly at the absolute scale because the across-workflow variation (some workflows are inherently easier) dwarfs the across-model variation; the centering used in the main-text figure removes that shared workflow-difficulty signal. Concretely, for each workflow and metric we first compute each top-3 model's absolute mean score, then subtract the mean across the three models for that same workflow. The main-text radar therefore reports percentage points relative to the per-workflow 3-model mean, not standard deviations. For plotting only, these signed deviations are shifted by a constant positive radius on a zoomed percentage-point scale; the radial tick labels report the original signed deviations.

M HARNESS AND HUMAN VALIDATION

Tool output formats and document parsing. All models use the same tool implementations. `web_search` returns titles, URLs, and snippets. `edgar_search` resolves tickers to SEC CIKs and returns each filing's form, date, accession number, and primary-document URL. `fetch_url` extracts PDFs with `pymupdf` and HTML with BeautifulSoup, then returns either the five most relevant BM25 chunks for a supplied query or the first 6,000 tokens with truncation marked explicitly. Transient fetch failures are retried automatically.

Fetch reliability in the released traces. Across the 1,500 released trajectories (50 questions × ten models × three trials), agents made 10,012 `fetch_url` calls: 80.1% used targeted BM25 retrieval, 11.8% of successful calls were visibly truncated, and 10.5% returned an error. Fetch failures are associated with lower trajectory scores (Pearson $r = -0.20$), although harder questions may both induce more speculative retrieval and receive lower scores. Retrieval remains the highest-credit stage for every model (Figure 7), indicating that document processing is not the dominant source of benchmark difficulty.



Figure 16: Per-workflow rubric (left) and final-answer accuracy (right) of the top-3 closed models in absolute form, complementing Figure 6. Boxed labels mark the workflow leader; the dashed gray polygon traces the per-workflow 3-model mean (the reference that Figure 6 subtracts to produce its centered values). Polygons cluster tightly at this scale, which is the visual hazard the centering used in the main text avoids.

Independent finance-practitioner audit. As a separate quality-control step, nine finance professionals independently re-performed 90 tasks authored by other practitioners and documented their derivations in spreadsheet workpapers. Applying the same two-judge rubric procedure used for model trajectories to the workpapers flattened to text yields a raw two-judge macro-averaged rubric score of 90.19% over 1,920 rubric criteria and 3,840 judge decisions, and 98.9% final-answer accuracy. Flattening removes formulas, cell comments, and number formats that carry derivation steps. A manual rubric-line audit of 33 human solutions across six workbooks found 277 lost judge-points, of which 44 (15.9%) were attributable to flattening (32 omitted formulas or formula precision, 10 omitted comments, 2 omitted number formats). Applying this share to the raw loss gives an adjusted score of $1 - (1 - 0.9019)(1 - 0.159) = 91.7\%$; this is a proxy adjustment, not a formula-aware regrade. Per judge, the raw rubric scores are 88.0% for Gemini (89.9% adjusted) and 92.4% for Opus (93.6% adjusted), with final-answer accuracies of 100.0% and 97.8%. Workpaper evidence is available for 89 of the 90 solutions; the remaining workpaper was permission-gated.

Human agreement and manual auditability. One finance expert, blinded to the LLM grades, independently regraded all 387 rubric criteria in 30 released trajectories spanning 21 questions and all ten models. The two-judge score closely reproduces the expert assessment: trajectory-level point-weighted scores correlate at Pearson $r = 0.946$ and Spearman $\rho = 0.942$, with a mean absolute error of 5.6 pp, while the aggregate scores differ by only 0.4 pp (49.0% versus 49.4%). Criterion-level agreement is 89.4% with Gemini ($\kappa = 0.789$) and 88.6% with Opus ($\kappa = 0.767$). Where the two judges agree, the expert confirms 98.1% of their 314 decisions ($\kappa = 0.961$); across the remaining 73, the expert sides with Gemini in 38 cases and Opus in 35, supporting the use of their average rather than either judge alone.

Manual fetch audit. A finance expert reviewed a model-balanced sample of 400 non-error `fetch_url` outputs (40 per model). 86.0% were readable: 69.5% contained the relevant passage and 16.5% were off-target. A further 10.3% were truncated before the target and 3.5% were empty or boilerplate; only one output (0.3%) was parser-garbled. Because agents can refine the retrieval query, increase the token budget, or retry, most incomplete outputs leave a direct recovery path. The audit therefore points to retrieval quality rather than parser corruption as the principal tool-side limitation.

N FINER-GRAINED CALCULATION ERROR ANALYSIS

The clean-setup analysis (Figure 7 and Appendix F) shows that Calculation credit stalls near 85% even after full Retrieval and Definition credit, despite sandboxed Python access. We therefore perform a trace-aware error analysis over the full benchmark.

Method. We retain every Calculation criterion for which at least one judge withholds credit in trajectories where both judges award full Retrieval and Definition credit. This yields $3{,}261$ missed criteria from 855 trajectories spanning 180 questions. For each trajectory, a GPT-5.6 Terra classifier receives the question, reference and model answers, complete assistant and tool trace, missed criteria, and both judges' line-level explanations. Structured outputs assign each criterion one of eight primary categories and any applicable secondary flags. Lost rubric points equal the criterion weight multiplied by the fraction of judges withholding credit.

Findings. Financial convention and task-interpretation mismatches jointly account for 38.2% of lost rubric points, making alignment with the requested financial quantity the dominant residual failure. Input selection, precision, and rounding follows at 25.9%, consistent with models failing to maintain a consistent precision lineage across intermediate values. Omitted calculations account for 13.0%, reflecting work that was not made explicit enough to grade. True arithmetic errors account for only 1.3% of lost rubric points, consistent with agents having access to Python.

Limits. These classifier labels are descriptive rather than expert-adjudicated root causes, and one upstream choice can affect several downstream criteria. Non-exclusive secondary flags record propagated errors, judge or tolerance disagreements, and cases where the model is more precise than the reference.

O RUBRIC-GRADING ROBUSTNESS

A natural concern for any rubric-based benchmark is that a system could earn credit through grader-targeted output rather than genuine analysis. Agents do not receive the rubric or reference answer at inference time, and judges are instructed to award credit only for positive evidence of an atomic fact or derivation step. These choices reduce direct checklist imitation, but they do not rule out generic attacks such as enumerating many candidate values, adding contradictory derivations, or optimizing against rubric-style graders. The final-answer criterion limits, but does not eliminate, the credit available to such strategies. We did not run a direct adversarial attack against the grader in the current evaluation. We therefore treat adversarial grader optimization as an unresolved validity risk rather than as a tested property of the benchmark.

Rollout stability. The rollout-count ablation that expands 30 public questions from three to five and ten trajectories is reported in Appendix J.

Three-judge robustness. GPT-5.5 independently regrades all $27{,}840$ full-benchmark trajectories, complementing the original Gemini 3.1 Pro and Claude Opus 4.7 grades. Across the three judge pairs, agreement on final-answer correctness is 98.4–98.8% (Cohen's $\kappa = 0.962$–0.972). Agreement on individual rubric criteria is lower at 90.2–92.6% ($\kappa = 0.803$–0.852), confirming greater judge sensitivity in criterion-level credit.

The judges also differ systematically in generosity: their mean rubric scores across models are 44.9%, 47.9%, and 46.2%, respectively. After centering each judge by this mean, the average per-model range is 1.8 pp (maximum 5.7 pp). Relative to the original two-judge mean, GPT-5.5 changes model rubric scores by 0.7 pp on average in absolute value and by at most 1.4 pp. Individual-judge rankings agree at pairwise Kendall's $\tau = 0.822$–0.956, with movement confined to the three leading systems and the two Gemini systems. Crucially, the three-judge mean reproduces the original two-judge ranking exactly ($\tau = 1.000$). Broad performance tiers are therefore stable, while exact ordering among closely matched systems is not invariant to judge choice.

Rubric-weight ablation. Point weights concentrate credit on conclusion-style criteria (Appendix I), so one may ask whether the weighting drives the ranking. The broad ordering is stable: rankings

under the point-weighted rubric score and the unweighted criterion pass rate (both reported in Table 4) agree at Kendall's $\tau = 0.90$ on the full benchmark ($\tau = 0.91$ on the released subset). Weighting does change the order within the leading cluster, so we do not treat the exact top-three ordering as invariant.

Table 6: Full Spotify example item: a third representative BIGFINANCEBENCH item with its complete rubric.

Question	Reference answer	Full rubric
What would be the payouts to Spotify's executive officers if Spotify received a buyout offer at a $800 per share price? Please use the latest disclosed incentive program payments (including options and RSUs), shareholdings, and change of control severance payments in addition to any other compensation schemes as of 9/30/2025. Please round the final answer to the nearest tenth of a billion.	$23.7 billion	Full rubric: 15 lines, 31 points. [+5] Calculates the payout to Spotify's executive officers by summing cash severance plus the value of all equity holdings (including options and RSUs) at the buyout price [+3] Utilizes the treasury stock method to calculate the net cash payment to key executives following cash required to exercise outstanding options [+1] Identifies Daniel Ek as a key executive officer. [+1] Identifies Gustave Soderstrom as a key executive officer. [+1] Identifies Alex Nordstrom as a key executive officer. [+1] Identifies Dustee Jenkins as a key executive officer. [+1] Identifies Christian Luiga as a key executive officer. [+2] Identifies a cash severance of 12x Annual Base Salary for Alex Nostrom, Cristian Luiga, and Gustav Soderstrom. [+2] Identifies Daniel Ek's 21.9 million outstanding share ownership (acceptable within +/- 1%) [+2] Multiplies Daniel Ek's outstanding share ownership by the buyout share price to calculate payments to Daniel Ek. [+1] Recognizes that Daniel Ek is leaving the CEO position at year-end 2025 and thus does not receive cash severence. [+2] Calculates the total payout for Daniel Ek of $23.3 billion (acceptable within +/- 1%) [+2] Calculates the total payout for Gustav Söderström of $223.7 million (acceptable within +/- 1%) [+2] Calculates the total payout for Alex Norström of $148.8 million (acceptable within +/- 1%) [+5] Final calculated grand total payout to the executive team is $23.7 billion (acceptable within +/- 1%)

Table 7: Complete successful Claude Opus 4.7 trajectory for the Udemy question in Table 2. The model submitted 1.74% YoY growth, a $9.18M ARR increase, and $536.38M implied Q3'26 ARR; both judges marked the final answer correct and awarded $32/32$ rubric points ($9/9$ lines).

Step	Visible agent text and tool call	Key returned content	Reward signal
0a	*"I'll research Udemy's Q3 2025 financials to find the relevant metrics."* `edgar_search(ticker=UDMY, form_type=8-K)`	EDGAR returns recent UDMY 8-Ks, including the 2025-10-29 Q3 2025 results filing.	Source discovery.
0b	`web_search` with query: Udemy Q3 2025 earnings Udemy Business ARR NDRR large customer	Result snippets report Udemy Business NDRR of 93% and Large Customer NDRR of 97%.	✓ `[+3]` Large Customer NDRR is 97%.
1	*"Let me get the Q3 2025 details to find the exact ARR figure."* `fetch_url(url=Udemy static Q3 2025 file, query=UB ARR)`	Fetch failed.	No direct credit.
2	`fetch_url(url=Udemy investor-relations release, query=UB ARR and NDRR)`	Fetch failed.	No direct credit.
3	`fetch_url(url=SEC Q3 2025 8-K, query=UB ARR and NDRR)`	Fetches the SEC 8-K shell for the 2025-10-29 Q3 results filing.	Source discovery.
4a	*"I need the press release exhibit. Let me check the filing index."* `fetch_url(url=SEC company 8-K index, query=October 29 2025 press release)`	Returns the SEC filing index containing the Q3 2025 8-K accession.	Source discovery.
4b	`web_search` with query: Udemy Q3 2025 Udemy Business ARR	Supplemental search result points back to Udemy investor materials.	No direct credit.
5	`fetch_url(url=SEC accession directory, query=press release exhibit 99.1)`	Returns the SEC directory for accession `0001607939-25-000138`.	Source discovery.
6	`fetch_url(url=SEC q32025pressrelease.htm, query=UB ARR and NDRR)`	SEC exhibit reports UB ARR of $527.2M and Large Customer NDRR of 97%.	✓ `[+1]` Q3 '25 ARR is $527.20M. Confirms `[+3]`.

7	*"Now let's compute the implied Q3'26 ARR using NDRR = 97% applied to existing ARR + $25M new logo ARR."* `python_exec` with `q3_25_arr=527.2`, `ndrr=0.97`, and `new_logo=25.0`	`Q3'26 ARR: $536.38M`
		`ARR Increase: $9.18M`
		`YoY Growth: 1.74%`
			✓ `[+2]` retained ARR formula.
			✓ `[+1]` new logo ARR is $25M.
			✓ `[+2]` Q3 '26 ARR formula.
			✓ `[+5]` Q3 '26 ARR is $536.38M.
			✓ `[+3]` ARR growth is $9.18M.
			✓ `[+10]` YoY increase is 1.74%.

8	`final_answer` with answer: Q3'26 UB ARR $536.38M; ARR increase $9.18M; YoY growth 1.74%.	Submitted answer includes the retained-ARR intermediate: $527.2M × 0.97 = $511.38M.	✓ `[+5]` retained ARR is $511.38M. ✓ Final answer correct.

Table 8: Question fixed-effects check for Calculation score among Retrieval+Definition-clean records, excluding GPT-5.4 Mini from the fit. "Raw calc" is the point-weighted conditional Calculation score. The fixed-effects column is the centered model effect from $C_{mqt} = \alpha_q + \beta_m + \epsilon_{mqt}$, in percentage points; positive values indicate higher Calculation credit than the average included model after absorbing question identity.

Model	RD-clean records	Raw calc	Question-FE effect
Claude Opus 4.7	440	87.6%	+2.8 pp
GPT-5.5	368	86.3%	+1.4 pp
Claude Sonnet 4.6	444	86.4%	−0.1 pp
GLM 5.1	395	86.3%	+0.5 pp
Qwen 3.6 27B	281	86.2%	−0.1 pp
Kimi K2.6	332	84.8%	−0.8 pp
Gemini 3 Flash	282	85.3%	−0.8 pp
Gemini 3.1 Pro	181	84.1%	−2.3 pp
Gemma 4 31B	186	84.1%	−0.6 pp

Table 9: BIGFINANCEBENCH dataset summary statistics, computed over the 928-item master release.

Statistic	Value
Questions (total)	928
Distinct annotators	52
Distinct reviewers	12
Query length, mean / median / max chars	324 / 281 / 1,479
Answer length, median / p75 / p95 / p99 chars	9 / 34 / 150 / 384
Answers $\leq$ 20 chars	648 (70%)
Answers $\geq$ 200 chars	33 (3.6%)
Rubric criteria per question, mean / median / max	16.9 / 13 / 192
Rubric criteria (total)	15,656
Rubric criteria with explicit `[+N]` weight	99.9%
Rubric points per question, mean / max	39.1 / 472
Rubric points (total)	36,241
Source mentions: 10-K	495
Source mentions: 10-Q	303
Source mentions: press release / IR	153
Source mentions: other / market data	148
Source mentions: 8-K	108
Source mentions: proxy / S-1	52
Source mentions: web	21
Multi-source questions	280 (30%)

Table 10: Trace-aware taxonomy of Calculation criteria missed despite full Retrieval and Definition credit on the full benchmark, weighted by lost rubric points.

Primary category	Share of lost points
Input selection, precision, or rounding	25.9%
Financial convention mismatch	20.4%
Task interpretation	17.8%
Omitted calculation or result	13.0%
Potential human reference/rubric math error	9.4%
Other	8.3%
Financial methodology	3.7%
Pure arithmetic	1.3%

Table 11: Full-benchmark rubric scores by judge ($n = 3$ trials per question; trials are averaged within question, then macro-averaged). Rank range is the lowest and highest rank assigned by any individual judge.

Model	Gemini	Opus	GPT-5.5	3-judge mean	Rank range
`Claude Opus 4.7`	57.9%	59.8%	57.7%	58.5%	1–2
`GPT-5.5`	56.1%	61.4%	57.4%	58.3%	1–3
`Claude Sonnet 4.6`	58.1%	59.0%	57.7%	58.3%	1–3
`GLM 5.1`	53.6%	56.1%	55.2%	55.0%	4
`Qwen 3.6 27B`	45.6%	49.0%	47.2%	47.3%	5
`Kimi K2.6`	44.5%	45.8%	45.7%	45.4%	6
`Gemini 3 Flash`	42.2%	43.8%	43.3%	43.1%	7–8
`Gemini 3.1 Pro`	36.3%	45.1%	39.6%	40.3%	7–8
`Gemma 4 31B`	33.8%	35.7%	34.5%	34.7%	9
`GPT-5.4 Mini`	21.0%	23.7%	23.2%	22.6%	10