INVESTMENT RESEARCH AGENT BENCHMARK: GROUNDING AGENT EVALUATION IN REAL-WORLD TASKS AND LIVE INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI agents enter investment research workflows, their evaluation must reflect diverse professional objectives and changing evidence. This poses three challenges: task grounding, scoring alignment, and reference maintenance. To address these challenges, we introduce the Investment Research Agent Benchmark (IRAB), comprising 101 tasks adapted from a corpus of 252,015 practitioner requests. Focused primarily on Chinese markets, these tasks span 12 research intents, ranging from industry-chain transmission analysis to stock and fund screening. For each task, IRAB pairs a task-specific rubric with dated reference evidence maintained under explicit refresh policies. Each rubric selects and weights applicable criteria from an inventory of 27 reusable, expert-derived criteria. A multiplicative scoring rule then combines data capability, answer quality, and reliability, so that detected factual failures reduce the overall score. Using IRAB, we evaluate 58 hosted configurations of 42 base models. Descriptive comparisons show higher data timeliness and source authority in newer release cohorts, while case analyses reveal persistent factual and verification failures. Repeated-run analysis further identifies variation in cited sources, underscoring the need for evidence preservation. **This work thus contributes a practitioner-grounded benchmark, a reliability-aware evaluation framework, and an empirical characterization of the capabilities and limitations of investment research agents.**

1 INTRODUCTION

AI agents are increasingly entering financial and **investment research** workflows, where they retrieve disclosures, analyze data, and produce reports for practitioners (Luan et al., 2026). Their usefulness depends on coordinating these activities around specific investment research objectives, such as tracing event impacts along value chains or conducting multi-criteria asset screening and evaluation. Such tasks require agents to draw on macroeconomic and corporate financial knowledge, connect economic mechanisms to asset exposures, and substantiate investment-relevant judgments. As agents take on more of this work, evaluation needs to establish how well they satisfy practitioners' research objectives and which capabilities make their outputs reliable.

Existing financial benchmarks offer complementary coverage of financial competence and research workflows. Early work, including PIXIU and FinanceBench, establishes foundations for language understanding, numerical reasoning, and financial data use, while broader suites also incorporate agent-related tasks (Xie et al., 2023; Islam et al., 2023; Xie et al., 2024; Guo et al., 2025). More recent work evaluates financial search, analytical derivations, and research reports through task-specific criteria (Hu et al., 2026; Bigeard et al., 2025; Wang et al., 2026; Zhang et al., 2026), alongside broader assessments of deep research and professional work (Du et al., 2025; Sharma et al., 2026; Vidgen et al., 2025). Table 1 compares their task provenance, evaluation design, reference maintenance, and evidence access. These choices determine which aspects of research practice a benchmark can assess. For investment research, we focus on three connected evaluation challenges.

1. Grounding tasks in practitioner demand. Financial subject matter alone does not establish how well a benchmark reflects practitioner demand. Task construction should account for observed research objectives and deliberately cover less frequent but professionally important requests.

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

Table 1: Comparison of IRAB with related benchmarks for financial research, deep research, and professional work products.

Benchmark	Scope	Request-grounded	Expert rubrics	Penalty / gate	Reference updates	Evidence access
Finance Agent Benchmark Bigeard et al., 2025	US	✗	✓	Check	✗	Web Market Docs
FinSearchComp Hu et al., 2026	CN/Global	✗	✓	✗	✓ Live	Web Market Docs
BigFinanceBench Wang et al., 2026	US	✗	✓	✗	✗	Web Market Docs
FinResearchBench II Luan et al., 2026	CN	✓	✗	✗	✗	NR
FrontierFinance Zhang et al., 2026	US	✗	✓	✗	✗	Web Market Docs
DeepResearch Bench Du et al., 2025	—	✗	✗	✗	✗	Web Market Docs
ResearchRubrics Sharma et al., 2026	—	Partial	✓	Add.	✗	Web Market Docs
APEX Vidgen et al., 2025	—	✗	✓	✗	✗	Web Market Docs
IRAB‡ this paper	CN/Global	✓	✓*	**Mult.**	✓ **Sched.**	Web Market Docs

✓/✗: documented/not documented in the cited version; —/NR: unspecified; CN: China. **Request-grounded**: adapted from observed user requests; Partial: mixed with expert-authored tasks. **Expert rubrics**: expert-written/reviewed rubric sets, excluding sampled validation alone. *Reusable expert criteria with task-specific selections and weights. **Penalty**: Check, contradiction check; Add., negative weights; Mult., multiplicative discount. **Live**: dynamic API references; **Sched.**: scheduled refreshes with archived versions. **Access**: Web, search; Market, financial data; Docs, documents. Colored tags indicate access in at least one configuration; grey, not documented. ‡ IRAB reports in-house runs; reference-harness access may differ (Section 2.6).

2. Connecting task requirements to scoring decisions. Research requests place different demands on evidence collection, comparison, and synthesis. Task-specific criteria and weights must reflect these requirements while retaining shared capability dimensions for comparison. Evaluation must also specify how detected factual failures affect the overall score.

3. Maintaining references as information changes. Reference evidence can become stale as relevant information changes. Evaluation needs dated references and explicit maintenance policies that preserve each task's intended period while making changes to its evidential basis inspectable. Version histories and citation records are also needed to interpret changes across repeated evaluations.

Existing benchmarks address important aspects of these challenges, but leave room for a more integrated treatment of all three (Table 1). We introduce the **Investment Research Agent Benchmark (IRAB)** to address these challenges. We adapt 101 tasks from selected requests in a corpus of 252,015 practitioner requests, covering 12 research intents and balancing observed demand with coverage of less frequent objectives. The tasks primarily cover Chinese markets, with additional US and global tasks. Through practitioner interviews, we derive an inventory of 27 reusable assessment criteria, then select and weight the applicable criteria for each task. Our **live gold** design separates these criteria from dated reference evidence maintained under task-specific refresh policies, with archived versions and source records retained for inspection. The **DQR** (data capability $\times$ answer quality $\times$ reliability) score aggregates rubric judgments, with detected factual failures reducing overall performance through a separate reliability factor.

Evaluation results show that across 58 hosted configurations from 42 base models, descriptive comparisons and case analyses reveal differences in evidence selection alongside persistent verification failures. A separate repeated-run analysis finds substantial variation in cited sources, illustrating the need to retain evidence records when interpreting repeated evaluations.

Our contributions span benchmark construction, evaluation design, and empirical analysis:

1. **A practitioner-grounded benchmark with maintained references.** IRAB connects diverse professional research objectives with maintained, inspectable reference evidence in a

common evaluation resource. It supports assessment grounded in practitioner demand while retaining the temporal context needed to interpret results as information changes.

2. **Task-specific evaluation with explicit reliability penalties.** We combine task-level selection and weighting of reusable expert criteria with a multiplicative reliability discount. This design connects professional expectations to shared capability dimensions and makes the consequences of detected factual failures explicit in the overall score.
3. **An empirical study of research-agent capability and reliability.** Our study illustrates the benchmark's diagnostic value by examining evidence-selection strengths, verification failures, and variation in cited evidence. These analyses support assessment beyond aggregate rankings and identify concrete targets for improving research-agent reliability.

2 The Investment Research Agent Benchmark

IRAB connects practitioner-grounded tasks, task-specific rubrics, and dated reference evidence to specify the research objective, assessment standard, and factual basis of each evaluation.

2.1 Practitioner-grounded tasks

IRAB comprises 101 benchmark tasks adapted from individual requests submitted to a deployed investment research agent. The source corpus contains 252,015 requests collected between April 1 and June 30, 2026. The platform serves institutional practitioners, including buy-side analysts, portfolio managers, and private-equity investors.

We inductively classified the underlying research intents into twelve categories, spanning analytical workflows such as single-company analysis, sector research, event attribution, and strategy backtesting (Figure 1A).

To construct the task bank, we used an LLM as a judge to assist request screening and incorporated expert feedback into task selection. We considered how clearly each request specified its research objective and requirements, together with observed execution time and diversity across research topics. To broaden coverage beyond the most common research needs, we oversampled less frequent intent categories (Figure 1B).

We anonymized and adapted the selected requests for evaluation through abstraction, rewriting, and selective substitution of request-specific details. These revisions retained selected research subjects, temporal scopes, constraints, and analytical objectives.

The 101 tasks primarily cover Chinese A-share and H-share markets, with additional coverage of US and other global exchanges. They span equities, bonds, REITs, commodities, and dedicated fund research.

2.2 Task-specific evaluation

We derived all rubric criteria from field interviews with investment practitioners serving as domain experts. For concrete deliverables, experts identified the decision to support, the necessary evidence, consequential failures, and minimum passing requirements. Each criterion specifies what is assessed; its rubric defines how it is scored through applicability rules, grading scales, anchors, and task-specific references (Appendix A.2).

The 27 criteria form three families. *Data and factuality* assesses the evidence supporting a response and the correctness of its claims and calculations. *Reasoning and insight* assesses how well the analysis addresses the research objective and supports its conclusions. *Delivery and expression* assesses the form, clarity, and usability of the requested deliverable. Appendix A.2 defines each criterion and reports its task coverage.

The rubrics use mixed grading formats (Rao & Callison-Burch, 2026): 18 criteria use ordinal scales, five use nominal categories with predefined utilities, and four use binary judgments. Values are normalized to $[0, 1]$ for aggregation; Appendix A.3 details normalization and distinguishes positive binary criteria from factual-failure penalties.

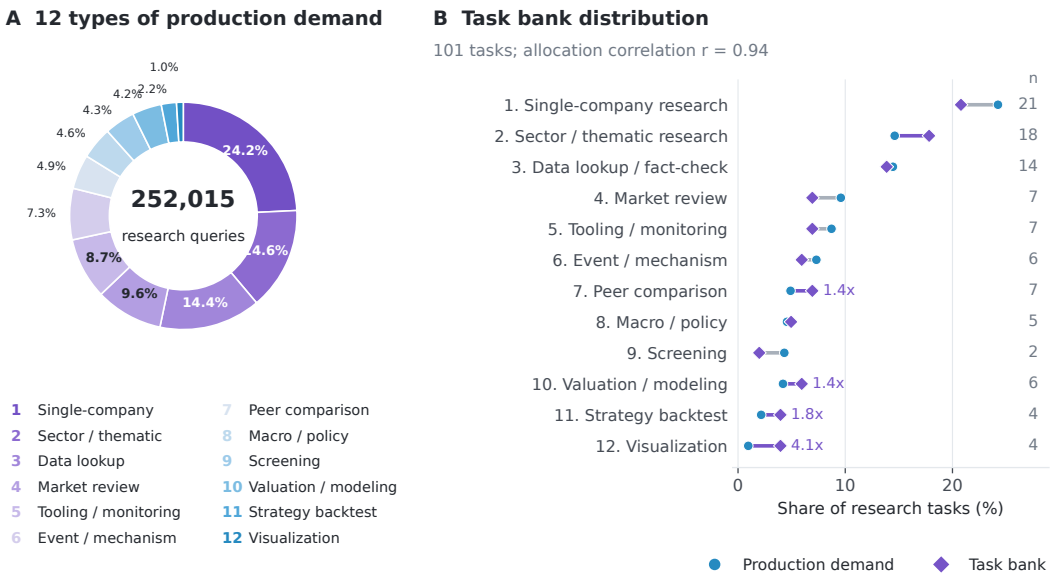

Figure 1: Alignment of IRAB tasks with real-world user intent. **(A)** Distribution of the 12 analytical intent categories across 252,015 practitioner requests. **(B)** Task bank share (diamonds) against observed demand share (circles). Task shares correlate with observed request shares ($r = 0.94$).

Each task-specific rubric combines three to eight applicable criteria, their weights, and reference items specifying evidence, requirements, or failure conditions. Selection determines what is assessed; weights express its relative importance. Table 2 illustrates this distinction: the A-share limit-up lookup weights `answer_correctness` at $+3$ and `data_timeliness` at $+1$, whereas the ASCO task weights `concept_clarity` at $+3$ and `analysis_depth` at $+2$ to assess explanations of technical findings and their investment implications for non-specialists.

Experts reviewed criterion selections and weights, combining explicit task requirements with expectations for the deliverable. For example, the East Money reference requires monthly series for full completeness credit, although the displayed task does not specify frequency (Table 2). Expert review of judge-assigned scores informed anchor calibration (Section 2.5).

2.3 Time-aligned live gold

Task-specific rubrics specify how responses are assessed, while their reference evidence can change. Published answers may enter training corpora and fixed references may diverge from later information (Vu et al., 2024; Ouyang et al., 2025). IRAB therefore separates stable assessment criteria from dated factual material. A task's temporal anchor specifies the period its request addresses; its reference-maintenance policy specifies when that material should be checked or refreshed. These fields are distinct: a historical period does not move with the execution date, but its reported figures can still be revised.

Of the 101 tasks, 69 have reference-refresh schedules and 32 have no scheduled refresh. The frequencies range from daily to yearly (Appendix A.1). These are maintenance counts, not a partition into rolling and fixed temporal anchors. A reference without a refresh schedule can still receive a correction. The task index records its as-of date, frequency, next-refresh date, and archived-version dates separately (Appendix A.4). Reference versions are matched automatically by code for each evaluation.

Retrieval agents assemble factual targets and assessment anchors from professional data services, corporate filings, and the web, retaining dated provenance for inspection. Dates and completed refreshes alone do not establish correctness or suitability for a particular execution.

Table 2 links three requests to selected criteria and dated reference items; Appendix A.5 discusses their assessment requirements.

Table 2: Three IRAB tasks with their selected criteria and sample reference items. The tasks are English translations of the adapted Chinese benchmark tasks. Bracketed numbers are criterion weights; negative weights mark penalty criteria (Section 2.4). The reference column shows sample items for criteria marked •; each selected criterion has reference items. Adapted task specifications and expert-reference expectations are shown separately.

Task	Selected criteria	Reference items (selected criteria)
ROLLING How many A-share stocks hit limit-up this past Monday? Exclude ST stocks and first-day listings; report the 10% and 20% tiers separately; name the stock with the longest run of consecutive limit-ups and how long it ran.	• [+3] answer_correctness [+2] data_quality [+1] data_timeliness [+1] user_need_coverage [−1] data_inconsistency	Gold as of 2026-09-14; refreshed every Monday. answer_correctness • 55 stocks closed limit-up on Monday, 2026-09-14, after the stated exclusions. • Longest streak: Mindong Electric Power (000993), four consecutive limit-ups; Guangdong Ultrasonic Electronics, Sinovatio and Kaisheng New Energy follow at three.
ROLLING Track the most high-profile presentations and clinical readouts from this year's ASCO annual meeting. Focus on A- and H-share listed companies; explain the drug mechanisms and trial readouts in plain terms for a reader without a biomedical background; assess what the readouts imply for revenue and valuation.	• [+3] concept_clarity [+2] data_quality [+2] analysis_depth [+2] user_need_coverage • [+1] source_authority [+1] data_timeliness [−2] data_distortion	Gold as of 2026-06; "this year" is ASCO 2026, May 29 to June 2. concept_clarity • Rank the endpoints and say what each measures: OS (do patients live longer) above PFS (is the disease held in check) above ORR (do tumours shrink, a short-term signal that may not last). source_authority • Tier 1: ASCO abstracts and plenary sessions, company announcements and IR materials, simultaneous publication in *The Lancet*, NEJM or JCO. Not tier 1: second-hand summaries by pharma bloggers or WeChat public accounts, unattributed "sources say" reports, investor-forum posts.
FIXED Based on analyst reports and financial databases, conduct a retrospective analysis of East Money (300059.SZ) from 2012 to 2020. Cover operating metrics (active users, ARPU), year-on-year revenue and profit growth, operating profit, share price, and PE and PS multiples; plot the series over time and deliver them in a self-contained HTML file.	[+3] data_quality • [+2] data_completeness [+2] analysis_depth • [+2] chart_fidelity [+1] deliverable_compliance [+1] user_need_coverage • [−2] data_distortion [−1] data_inconsistency	Gold: one archived version; no scheduled refresh. data_completeness • Full credit requires every series at monthly resolution in the HTML file; annual values alone earn a lower tier. chart_fidelity • Plot PE on a logarithmic axis, or truncate it with a note: the 2013 and 2014 values (about 2,040× and 204×) would flatten every other year on a linear axis. data_distortion • The brokerage subsidiary was consolidated in December 2015; net profit or revenue reported on an unconsolidated basis after that date counts as a distortion.

Reference-version history is reported in Appendix G.6; execution records and reference versions must be interpreted together.

2.4 Score Aggregation for Investment Research

Simple additive scoring allows strong presentation to offset factual weaknesses that limit a report's usefulness. Our expert-derived criteria motivate separating these dimensions and explicitly penalizing material failures. An error's severity and effect depend on the task and scoring specification; not every inaccuracy makes a report unusable.

The DQR score combines data capability (D), answer quality (Q), and reliability (R), as summarized in Figure 5. These factors specify how judgments affect the score, while the three semantic families group criteria by what they measure. D combines positive data criteria; Q combines the remaining positive criteria; and R is the complement of the weighted penalty. Thus, a severe factual failure belongs to the data and factuality family but contributes to R. Appendix A gives all 27 criteria and their factor assignments.

As in rubric-based frameworks (Rao & Callison-Burch, 2026), let $s_{tj} \in [0, 1]$ be a normalized criterion value and w_{tj} its task-specific weight. Positive criteria use a weighted mean within their respective sets, $\mathcal{D}_t$ and $\mathcal{Q}_t$. Negative weights identify failure criteria (Akyürek et al., 2025); for the penalty set $\mathcal{P}_t = \{j : w_{tj} < 0\}$, $s_{tj} = 1$ denotes a detected failure. Rather than subtracting penalties

from a common sum, we multiply the positive factors by their weighted complement:

$$X_t = \frac{\sum_{j \in \mathcal{X}_t} w_{tj} s_{tj}}{\sum_{j \in \mathcal{X}_t} w_{tj}}, \quad (X \in \{D, Q\}) \tag{1}$$

$$R_t = 1 - \frac{\sum_{j \in \mathcal{P}_t} |w_{tj}| s_{tj}}{\sum_{j \in \mathcal{P}_t} |w_{tj}|}, \qquad S_t = 100\, D_t Q_t R_t \tag{2}$$

Each factor lies in $[0, 1]$, and S_t is reported on a 0–100 scale. Its normalized value $S_t/100$ cannot exceed any factor. The reliability factor discounts scores according to the task's penalty weights; Appendix A.3 illustrates partial and zero-score cases.

The benchmark averages these task scores rather than multiplying the pool means of D, Q and R. These averages reflect the curated task mix and its deliberate adjustments to category coverage. An empty contributing factor takes the neutral value 1. Failed evaluations are rerun and kept distinct from valid non-applicability; scores are reported alongside task coverage. The score summarizes judgments under the specified rubrics; neither the aggregation rule nor a high score guarantees that the judge has identified every factual error.

2.5 Human Calibration and Evaluator Repeatability

Opus 4.8 is the primary judge. Evaluator consistency is distinct from judgment correctness (Zheng et al., 2023). IRAB scores criteria separately to reduce conflation of data accuracy, analytical quality, and presentation (Yamauchi et al., 2025); Appendix B gives the calibration and validation protocols.

Human calibration of grading anchors. During rubric development, we iteratively refined the grading anchors. After the judge scored responses with the revised rubrics, 12 experts reviewed results across 30 tasks, selecting *too high*, *too low*, *reasonable*, or *this dimension should not be assessed*. Across all expert feedback responses, 82% marked the score as *reasonable*, with all four categories in the denominator. This rate summarizes feedback during rubric development, not held-out judgment accuracy (Appendix B.1).

Repeated scoring of fixed reports. Seven configurations span pool scores of 39–73. Within each configuration, the same reports are scored under a fixed judge, rubric version, and reference. Task-level DQR repeatability, measured by Krippendorff's α with interval distance (Krippendorff, 2004), ranges from 0.782 to 0.845 (Table 5). The mean absolute difference between pass-level pool scores, averaged over pass pairs and then configurations, is 1.76 points on the 0–100 scale; this is neither a confidence interval nor a universal threshold for comparing models. Pool-level stability can conceal large task-level changes: about 60% of scoring pairs differ by at most 5 points, while 4–9% differ by more than 50 points, associated with penalty verdicts that flip between passes (Figure 4A). Close leaderboard comparisons therefore require attention to both pool-level variation and the sensitivity of individual task scores to penalty verdicts.

Conditional differences between judge families. We compare Opus 4.8 and GPT-5.5 on archived reports, motivated by evidence of judge self-preference (Pombal et al., 2026). Relative to each judge's fit on neutral-authored batches, own-family residuals in severe-distortion firing rates are -3.5 percentage points for the Anthropic-family judge and -26.6 for the OpenAI-family judge (Table 6). Negative residuals are consistent with own-family leniency, but report characteristics can also affect them; the comparison neither identifies a unique cause nor establishes that the primary judge is neutral (Appendix B.3). Repeated agent executions across dates introduce additional variation in evidence, execution, and judging (Appendix G.5).

2.6 Evaluation Harness and Data Pipeline

The reference harness, IRAB Agent, combines a coding-agent runtime with five financial tools: research-corpus search, structured global market data, structured Chinese market data, public web search, and webpage reading (Appendix C). These support iterative retrieval, quantitative processing, and synthesis. The leaderboard in Section 3 was produced in the hosted production environment and evaluates model configurations together with its tools and orchestration. The reference harness

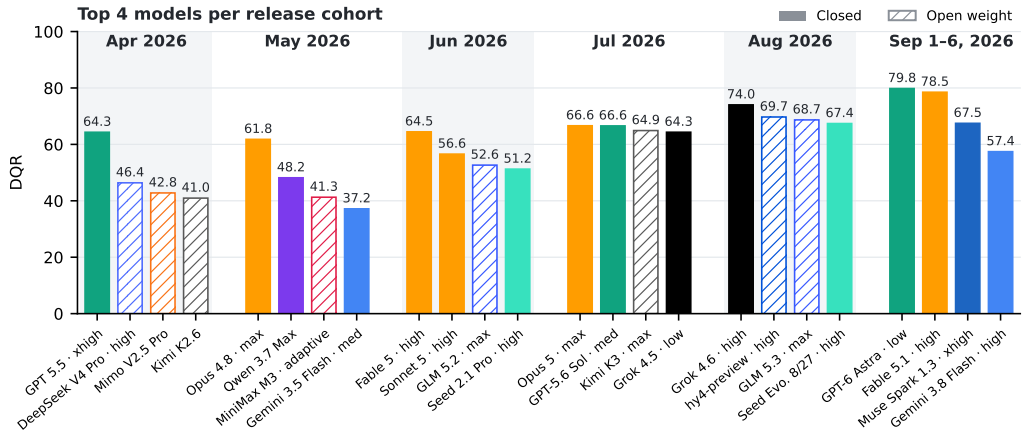

Figure 2: Published DQR scores at the September 6, 2026 cut-off. Each release-month group shows the four highest-scoring base models using their best available configuration. All bars belong to the same leaderboard snapshot; task coverage is reported in Table 8.

specifies tool contracts and artifact handling; equivalent outcomes with the hosted system have not been established (Appendix C.3).

Agents follow a *collect-process-deliver* loop, using the coding workspace for operations such as table joins and financial-ratio calculations, and determine when to finalize their deliverables. Source markers link reported data to retrieved records. The execution archive retains deliverables, source records, tool invocations, token consumption, and wall-clock time for subsequent inspection.

Task coverage, execution date, reference version, and judge configuration remain part of the reported setting. Licensed access and changing content can constrain redistribution and historical replay; tool-access differences also matter when comparing with public-source financial benchmarks (Wang et al., 2026). Appendix H documents the benchmark and reference-agent repositories.

3 How AI Conducts Investment Research

We analyze the September 6, 2026 leaderboard snapshot of 58 configurations from 42 base models evaluated in the hosted production environment (Appendix C.3). Research quality and resource cost vary substantially across configurations. Later-released configurations make more execution-related tool calls and score higher on evidence-related criteria, yet case analyses reveal numerical distortions, internal conflicts, and insufficiently verified claims in the delivered reports.

3.1 Leaderboard and Cost–Performance Trade-offs

Published DQR scores range from 27.2 to 79.8 across the evaluated configurations (Table 8, Appendix D). Astra low has the highest published score on 99 tasks, while Fable 5.1 high leads the configurations evaluated on all 101 tasks at 78.5. The full leaderboard reports task coverage alongside each score; comparisons depend on this coverage and, for close rankings, evaluator variation (Section 2.5).

Even the leading configurations combine high answer-quality scores with unmet data requirements. Across the five highest-scoring configurations, mean data capability is 87.9, answer quality is 93.3 and reliability is 92.1 on their respective 0–100 scales. By exposing these factors separately, IRAB makes weaknesses in data and factual reliability visible alongside the overall quality of the deliverable.

Quality at different resource costs. Similar published scores can come at markedly different costs (Figure 3A). Fable 5.1 high and Astra high both score about 78.5 on the full task set, at $5.11 and $12.62 per task, respectively. The highest-scoring evaluated open-weight configuration, hy4-preview

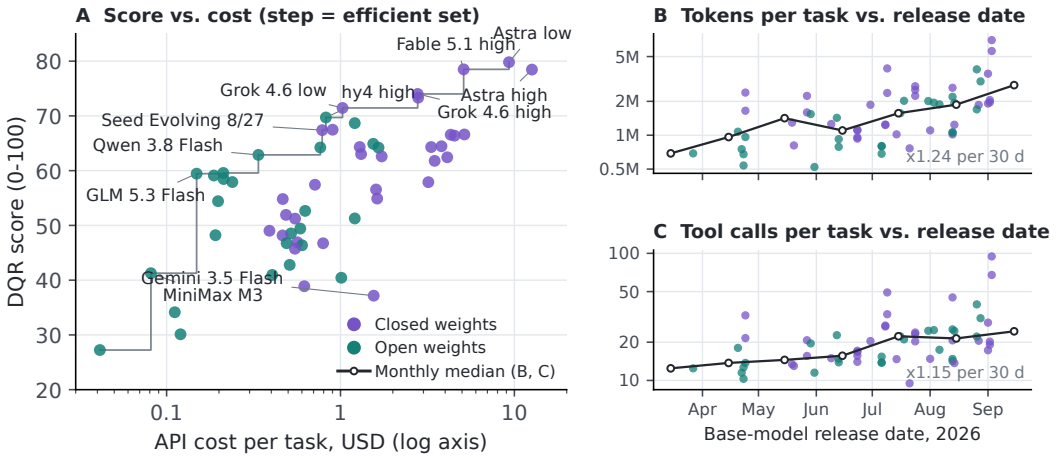

Figure 3: **A:** DQR against mean API cost per task for the 58 configurations; the step line joins the observed efficient set. **B, C:** mean tokens and tool calls per task against the base model's release date; lines join monthly medians. Costs use run-time prices; tokens include cached tokens. Task coverage varies across configurations (Table 8).

high, also covers all 101 tasks, scoring 69.7 at $0.82 per task. The frontier summarizes the observed score–cost trade-offs under each configuration's reported task coverage. Table 8 complements this comparison with execution time, tokens and tool calls.

3.2 Capability Evolution and Evidence Awareness

Across release cohorts, research workflows become more execution-intensive and evidence-related scores improve. Data capability rises by 9.8 points between the earliest and newest release tertiles (Table 9). These comparisons are descriptive: task coverage varies, and some base models contribute multiple configurations (Appendix E.1).

Expanding research execution. Later-released configurations devote more resources to carrying out research workflows. Both tokens and tool calls per task increase (Figure 3B–C). Execution-related tool calls—shell commands, file operations, code writing and plotting—show a fitted increase of 26% per 30 days of model release date, compared with 9% for retrieval-related tool calls (Appendix E.3). This pattern is consistent with advances in long-horizon coding, which emphasize sustained planning, tool use and iterative execution (OpenAI, 2025; Anthropic, 2026). Such capabilities may also help research agents carry a question through retrieval, data processing, calculation and report construction.

Recognizing current evidence. Newer configurations are better at selecting evidence that is current for the task. Timeliness shows the largest improvement among the seven criteria reported on the leaderboard, rising by 16.4 points between the earliest and newest release tertiles (Table 9). An IPO report illustrates the underlying judgment: Astra high identified a prospectus resubmitted in July and distinguished its financial reporting period from that of the older version (Appendix E.4). Selecting the appropriate document version helps avoid treating superseded disclosures as the current basis for analysis.

Distinguishing primary records from retellings. Newer configurations also ground claims in sources closer to the original record. Source authority rises by 6.8 points between the earliest and newest release tertiles (Table 9). This criterion distinguishes first-hand records, such as company announcements and financial statements, from professional secondary analysis and subsequent retellings. In an amino-acid capacity analysis, Astra high preferred Fufeng's annual report over an industry table retaining older figures (Appendix E.4). This choice grounds the reported capacity in the company's own disclosure.

Evidence across repeated runs. A separate repeated-run analysis examines source consistency. In a 14-task subset with identical gold references, Opus 4.8's cited documents have a median Jaccard overlap of 0.12 across runs 12 days apart (Appendix G.3). Different sources may support the same facts; retaining each report's cited evidence avoids relying on reruns to reconstruct it.

3.3 Failure Patterns in Research Deliverables

Better source selection still leaves the challenge of checking and integrating evidence into a final analysis. The following selected cases draw on reports and judge findings across 24 batches to illustrate numerical distortions, internal conflicts, and insufficient verification of retrieved claims. Table 10 provides excerpts, and Appendix F.1 describes case selection; the cases illustrate failure mechanisms without estimating their prevalence.

Order-of-magnitude errors. Numerical scale can be lost even when the company, metric and surrounding explanation remain plausible. Fable 5.1 high reports Luxshare Precision's first-half 2026 net profit attributable to shareholders as RMB 0.784 billion; the company's interim report discloses RMB 7.843 billion, approximately ten times as much (Luxshare Precision Industry Co., Ltd., 2026). Other observed errors involve decimal and percentage scales or currency units. These cases show how fluent discussion can conceal materially distorted quantities.

Conflicts across a long deliverable. Information available within a report may still be used inconsistently across its sections. In a humanoid-robot supply-chain report, Astra low's summary describes a component price of 800 CNY as already reached, while the body gives a current price of 2,000 CNY and treats 800 as a future target. The conflict concerns both value and timing, illustrating a failure to reconcile information already present within the report.

Reliance on ready-made analyses. Retrieved analyses can supply conclusions that the model does not sufficiently examine for itself. In a lithium-price report, Opus 5 max presents a range supported by several agreeing sell-side sources but omits the opposing forecasts identified in the reference. In a company analysis, GPT-5.5 repeats an expert-note repost's claim about an exclusive supply agreement as fact, despite a company clarification recorded in the reference. The first case overlooks competing views; the second leaves a claim's factual status unresolved. Both illustrate the need to independently compare and verify retrieved conclusions.

4 Limitations and Conclusion

Limitations. The task bank reflects one research environment's predominantly Chinese-language demand, with few tasks in several categories. Licensed sources and changing web content limit exact replay of live evaluations. Experts contributed to rubric design and calibration, but we have not established ongoing expert review as models and research outputs evolve. Several criteria have near-ceiling means in the newest release tertile (Appendix E). Close DQR comparisons also depend on task coverage and evaluator variation (Section 2.5). Finally, we have not measured how the generated reports affect practitioners' investment decisions.

Conclusion. We introduced IRAB to evaluate investment research agents against practitioner objectives and expert-derived criteria in a changing information environment. Task-specific rubrics and dated references support comparisons of research quality and resource cost, exposing weaknesses in data use and factual reliability.

Descriptive comparisons across release cohorts show more extensive research execution and higher data timeliness and source authority scores in later-released configurations. Case analyses still reveal numerical scale errors, internal contradictions, and insufficient verification of retrieved analyses. Citation turnover in the separate repeated-run study motivates preserving each report's cited evidence. These findings motivate agents to check quantities, reconcile claims across a report, and verify retrieved conclusions against primary records and competing evidence. IRAB supports assessing these capabilities, with ongoing expert calibration and evaluation of investment-decision impact as next steps.

REFERENCES

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego Mares, Pavit Singh, Michael Liu, Subodh Chawla, Pete Cline, Lucy Ogaz, Ernesto Hernandez, Zihao Wang, Pavi Bhatter, Marcos Ayestaran, Bing Liu, and Yunzhong He. PRBench: Large-scale expert rubrics for evaluating high-stakes professional reasoning, 2025. URL https://arxiv.org/abs/2511.11562v1.

Anthropic. Introducing Claude Opus 4.6, February 2026. URL https://www.anthropic.com/news/claude-opus-4-6.

Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. Finance agent benchmark: Benchmarking LLMs on real-world financial research tasks, 2025. URL https://arxiv.org/abs/2508.00828v1.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents, 2025. URL https://arxiv.org/abs/2506.11763v1.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6258–6292, 2025.

Liang Hu, Jianpeng Jiao, Jiashuo Liu, Dongyuan Mutu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, Fei Hu, Yali Liao, Zaiyuan Wang, Jingkai Liu, Daibin Sun, Ziqing Zeng, Zhiyuan Zeng, Chenghao Yang, Qianyu Yang, Mingren Yin, Xinying Feng, Ge Zhang, Xinyi Zhang, Xiying Zhao, Zhenwei Zhu, Hongseok Namkoong, and Wenhao Huang. FinSearchComp: Towards a realistic, expert-level evaluation of financial search and reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4d42358702dff82e1436550a05ade260-Paper-Conference.pdf.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering, 2023. URL https://arxiv.org/abs/2311.11944v1.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, 2 edition, 2004.

Beidi Luan, Rui Sun, Sinuo Wang, Yan Gu, Chao Li, Zhenliang Xiong, Jing Li, and Zuo Bai. FinResearchBench II: A deep research benchmark with consensus-derived gold rubrics for distinguishing financial report quality, 2026. URL https://arxiv.org/abs/2607.12252v2.

Luxshare Precision Industry Co., Ltd. 2026 interim report, August 2026. URL https://static.cninfo.com.cn/finalpage/2026-08-25/1225500221.PDF. In Chinese. Financial highlights, p. 8.

OpenAI. Introducing GPT-5.2-Codex, December 2025. URL https://openai.com/index/introducing-gpt-5-2-codex/.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. HoH: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6036–6063, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.301. URL https://aclanthology.org/2025.acl-long.301/.

José Pombal, Ricardo Rei, and André F. T. Martins. Self-preference bias in rubric-based evaluation of large language models, 2026.

Delip Rao and Chris Callison-Burch. Autorubric: Unifying rubric-based LLM evaluation, 2026. URL https://arxiv.org/abs/2603.00077v2.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. ResearchRubrics: A benchmark of prompts and rubrics for evaluating deep research agents. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010639.

Bertie Vidgen, Abby Fennelly, Evan Pinnix, Julien Benchek, Daniyal Khan, Zach Richards, Austin Bridges, Calix Huang, Kanishka Sahu, Abhishek Kottamasu, Bo Ma, Ben Hunsberger, Isaac Robinson, Akul Datta, Chirag Mahapatra, Dominic Barton, Cass R. Sunstein, Eric Topol, Brendan Foody, and Osvald Nitski. The AI productivity index (APEX), 2025. URL https://arxiv.org/abs/2509.25721v6. Version 6: APEX-v1-extended.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.813. URL https://aclanthology.org/2024.findings-acl.813/.

Alex Wang, Georg Meinhardt, Jacob Katz, Joseph H. Kim, Pratyush K. Chaudhary, Chase Blagden, and Eric Xu. BigFinanceBench: A workflow-grounded benchmark for financial-research agents, 2026. URL https://arxiv.org/abs/2606.03829v1.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

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

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. An empirical study of LLM-as-a-judge: How design choices impact evaluation reliability, 2025.

Yuhao Zhang, O. Ozan Koyluoglu, Thejas Venkatesh, Richard Diehl Martinez, Vishank Bhatia, Arash Alidoust, and Ashwin Paranjape. Frontierfinance: A challenging benchmark for measuring frontier intelligence of finance agents, 2026. URL https://arxiv.org/abs/2608.11683v1.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Appendix guide.

	Supporting material	Main text
A	Benchmark construction and scoring	§2.1–2.4
B	Evaluator validation	§2.5
C	Reference harness and execution	§2.6
D	Complete evaluation results	§3.1
E	Capability and tool-use analyses	§3.2
F	Failure-mode analysis and cases	§3.3
G	Evidence turnover and repeated evaluation	§3.2
H	Code and data availability	§2.6

A Benchmark Construction and Scoring Details

This appendix gives the task metadata and scoring specification underlying Sections 2.1–2.4.

A.1 Task-bank composition

Table 3: Task-bank metadata. Counts refer to the 101 tasks analyzed in this paper. A static reference has no scheduled refresh; a refreshing reference carries a refresh frequency and a next-refresh date. Both may receive corrections.

Dimension	Counts
Research activity	Company research 21; sector research 18; data lookup 14; peer comparison 7; tooling/monitoring 7; market review 7; event/mechanism 6; valuation/modeling 6; macro/policy 5; backtesting 4; visualization 4; screening 2
Reference	Static 32; refreshing 69 (quarterly 54; monthly 11; weekly 2; daily 1; yearly 1)

A.2 Rubric definitions and elicitation

The rubric criteria were derived from interviews with investment practitioners who routinely use or produce investment research. Each interview focused on a concrete deliverable: the practitioner identified (i) the decision to support, (ii) sufficient evidence, (iii) consequential failures, and (iv) the minimum requirements for a passing answer. We expressed these observations as reusable rubric templates with applicability rules, typed grading scales, anchor examples, and task-specific references.

Table 4 summarizes the capability or failure measured by each criterion.

Table 4: Definitions, scoring roles and task coverage of the 27 criteria used in the benchmark. The three families group criteria by meaning; the Factor column identifies their role in Equations 1–2. The Tasks column gives the number of tasks to which each criterion is bound. The two R criteria are negative-weight penalties.

Criterion	What it measures	Type	Factor	Tasks
Data and factuality				
Data quality	Evaluates the accuracy and coverage of reported facts against the reference.	Ordinal	D	97
Data completeness	Checks whether required data are present, regardless of their correctness.	Ordinal	D	13
Data timeliness	Assesses whether the analysis uses explicitly dated data current at the reference date.	Ordinal	D	73
Source authority	Assesses the authority of sources supporting the report's central claims.	Nominal	D	49

Continued on next page

Table 4 continued

Criterion	What it measures	Type	Factor	Tasks
Answer correctness	Checks whether all questions with definitive answers are answered correctly.	Binary	D	12
Calculation correctness	Checks whether numerical results follow correctly from the stated inputs.	Ordinal	D	27
Formula construction	Assesses whether required spreadsheet calculations use correct, live formulas.	Ordinal	D	4
Data distortion	Detects severe factual errors in confidently asserted claims central to the conclusion.	Binary	R	77
Data inconsistency	Detects material numerical contradictions within the response.	Binary	R	29
Reasoning and insight				
User-need coverage	Measures how fully the response addresses the user's explicit research requests.	Ordinal	Q	92
Analysis depth	Assesses whether the analysis explains causal mechanisms with case-specific detail.	Ordinal	Q	67
Cross-subject comparison	Assesses whether comparisons yield supported judgments about differences or patterns across subjects.	Nominal	Q	31
Key-driver identification	Assesses whether the analysis explains which drivers matter most and why.	Ordinal	Q	15
Actionable signal	Checks whether the response provides an investment recommendation specific enough to act on.	Binary	Q	13
Reasoning-angle coverage	Measures coverage of the explanatory mechanisms specified in the expert reference.	Nominal	Q	10
Uncertainty honesty	Assesses whether the analysis explains the magnitude of case-specific upside and downside uncertainty.	Ordinal	Q	9
Concept clarity	Assesses whether unfamiliar concepts and their investment relevance are explained clearly.	Ordinal	Q	6
Logic-chain integrity	Assesses whether conclusions follow logically from the stated premises.	Ordinal	Q	5
Non-consensus view	Assesses whether the report offers defensible insights beyond the reference consensus.	Ordinal	Q	2
Analytical autonomy	Assesses whether conclusions reflect the model's own synthesis across sources.	Nominal	Q	1
Delivery and expression				
Deliverable compliance	Assesses adherence to the user's explicit requirements for the deliverable's form.	Ordinal	Q	27
Chart fidelity	Assesses whether charts encode the requested information correctly and readably.	Ordinal	Q	7
Template adherence	Assesses whether the output faithfully follows and completes the required template.	Ordinal	Q	5
Skill executability	Assesses whether the specification is implementable and faithful to the user's requirements.	Ordinal	Q	3
Code quality	Assesses code logic and completeness through static inspection.	Nominal	Q	2
Content organization	Assesses how easily readers can follow the analysis and locate key findings.	Ordinal	Q	2
Content discipline	Assesses whether the content serves the task's purpose without unnecessary material.	Ordinal	Q	1

A.3 TYPED NORMALIZATION AND TASK-LEVEL AGGREGATION

The bank uses 18 ordinal, five nominal and four binary criteria. The raw ordinal scales are criterion-specific: data quality, completeness, calculation correctness and formula construction use 0–10; timeliness and concept clarity use 1–3; the remaining ordinal criteria use 1–4. An ordinal value x on $[a, b]$ is normalized to $(x - a)/(b - a)$, with out-of-range values clamped to the endpoints. Nominal verdicts use the criterion's predefined utility mapping. For example, analytical autonomy maps own

synthesis, mixed synthesis and restatement to 1, 0.5 and 0; a non-applicable verdict has no numerical value.

The four binary criteria have different roles. Answer correctness and actionable signal are positive criteria in D and Q, respectively. Data distortion and data inconsistency have negative weights and contribute to R. Weights come from the task-specific rubric bindings.

Aggregation excludes criteria whose value or weight is non-numeric. An empty contributing set gives its factor the neutral value 1, including $R = 1$ when no penalty is bound. Valid non-applicability and provider or parsing failures are recorded separately; failed evaluations are rerun. Task coverage is checked separately, because neutral factor defaults do not establish that a report was successfully evaluated.

Equations 1–2 combine the normalized values within each task and report the result on a 0–100 scale. With $D = 0.8$, $Q = 0.9$ and no detected penalty, the task scores 72. If one of two equally weighted penalties fires, $R = 0.5$ and the score is 36; if the sole bound penalty fires, the score is 0. The benchmark averages these task-level products, rather than multiplying the pool means of D, Q and R. The criterion-weighted mean emitted separately by the judge is distinct from this multiplicative leaderboard score.

A.4 Reference maintenance and date metadata

A temporal anchor specifies the period a task addresses; a refresh policy specifies when its reference is revisited. A fixed historical question can therefore have a scheduled review, and a reference without a schedule can still receive a correction. The task index records the as-of date, refresh frequency, next-refresh date, and archived-version dates separately; Table 3 gives the frequency counts. Reference versions are matched automatically by code for each evaluation. Appendix G.6 summarizes the versions available as of September 17, 2026.

A.5 Illustrative task specifications

Table 2 contrasts a factual lookup, an explanation of clinical evidence, and a historical analysis with a visual deliverable. The requests define the research objective and output, while the task-specific references supply factual targets and grading anchors. The displayed tasks are English translations of the adapted Chinese benchmark tasks, and only selected criteria and reference items are shown.

B Evaluator Validation

This appendix describes human calibration of grading anchors and subsequent analyses of evaluator behavior. Opus 4.8 is the primary judge. Re-scoring holds the agent output fixed; the agent re-runs in Appendix G.5 can change the output, retrieved evidence and reference version.

B.1 Expert review during rubric calibration

We calibrated grading anchors during rubric development (Appendix A.2). After rubric revisions, the judge scored task responses using the updated rubrics. Twelve experts then reviewed scoring results covering 30 tasks. The review asked experts to assess the judge's assigned scores using four feedback categories: *too high*, *too low*, *reasonable*, and *this dimension should not be assessed*. The last category addresses whether the criterion is appropriate for the task.

We define the *expert acceptance rate* as the proportion of individual expert feedback responses labeled *reasonable*, pooling responses across experts and tasks. All four categories contribute to the denominator, including *this dimension should not be assessed*. The final calibration review achieved an expert acceptance rate of 82%. The rate summarizes feedback on judge-assigned scores during rubric development; it does not measure inter-expert scoring agreement or performance on a held-out benchmark evaluation.

B.2 REPEATED SCORING OF FIXED OUTPUTS

The seven selected configurations span pool scores of 39–73. Within a configuration, the same reports are scored repeatedly under a fixed judge, rubric and reference. Before a pass, the rendered report is checked against its archived content hash; a mismatch is rejected so that an attachment extraction failure cannot masquerade as judge variance. Gateway response caching is disabled and vote latencies are inspected to check that repeated passes invoke the model.

Table 5 reports Krippendorff's α with interval distance on task-level DQR. Pool differences are absolute differences between two pass means, averaged over pass pairs. Task differences compare the same task across passes. These quantities keep comparisons meaningful when configurations have different numbers of repeats; a range over all passes would grow with the repeat count.

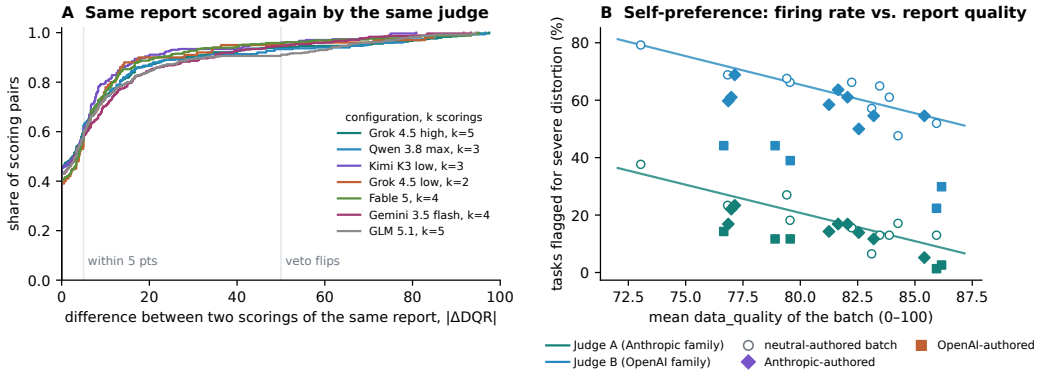

Figure 4: Judge stability. **A:** Repeatability of one judge on seven agent configurations selected to span the leaderboard's DQR range (pool means 39–73). Each configuration's reports were scored k times; for every report and every pair of scores, the curve records the absolute DQR difference, so a point (x, y) reads "a share y of scoring pairs differ by at most x points". The curves coincide: about 60% of pairs differ by at most 5 points and 4–9% by more than 50, the latter reflecting penalty verdicts that flip between passes. **B:** Severe-distortion firing rate against the batch's mean `data_quality` for two judge families; lines are fitted on the neutral-authored batches only. The OpenAI-family judge's points on OpenAI-authored batches (squares, right of the plot) sit far below its own line; the Anthropic-family judge's points on Anthropic-authored batches (diamonds) sit within a few points of its line.

Table 5: Repeated scoring by one judge, on the DQR scale. The seven configurations were selected to span the leaderboard's DQR range (pool means 39–73). Each configuration's reports were re-scored k times under a fixed judge configuration, rubric version and reference; inputs were verified by content hash. α is Krippendorff's alpha (interval) over tasks; the criterion-weighted mean the judge emits directly is shown for comparison. Pool $|\Delta|$ is the mean absolute difference between two passes' pool means; task $|\Delta|$ is summarized by the median and 90th percentile of the per-task pass-to-pass difference.

Configuration	k	n	Pool	α_{DQR}	α_{mean}	Pool $\lvert\Delta\rvert$	Task $\lvert\Delta\rvert$	p90
Grok 4.5 high	5	101	63.6	0.802	0.863	1.59	5.8	28.0
Qwen 3.8 max	3	100	67.3	0.782	0.791	1.30	4.2	32.4
Kimi K3 low	3	101	72.5	0.845	0.862	2.88	4.0	16.7
Grok 4.5 low	2	100	65.2	0.832	0.860	1.44	4.0	17.5
Fable 5	4	96	64.3	0.839	0.892	0.79	5.2	20.2
Gemini 3.5 flash	4	101	39.1	0.824	0.916	2.61	6.6	25.5
GLM 5.1	5	101	45.6	0.812	0.907	1.74	6.4	35.2
Range over 7 configurations			39–73	0.782–0.845	0.791–0.916	1.76 (mean)		

B.3 CROSS-JUDGE COMPARISON AND FAMILY SELF-PREFERENCE

The study compares Opus 4.8 and GPT-5.5 on batches of archived reports. Ten batches were authored by models outside both judge families (*neutral-authored*), nine by Anthropic models, and five by OpenAI models. The outcome is the severe-distortion firing rate: the share of assessed reports on which a judge flags a severe factual error.

Each judge's firing rate is regressed on mean data quality across the ten neutral-authored batches. Its own fit predicts a baseline rate for each family-authored batch; residuals are observed minus predicted rates. Table 6 reports these differences in percentage points and t statistics clustered by authoring model. A negative own-family residual is consistent with leniency, although report characteristics can also affect distortion findings.

Table 6: Conditional differences in severe-distortion firing rates by report-author family. Residuals are observed minus predicted rates, in percentage points, relative to each judge's regression on mean `data_quality` for ten neutral-authored batches. The two judges' mean `data_quality` values agree within 0.02 on these batches. t is a cluster-level statistic with one cluster per authoring model; Models counts distinct authoring models. Negative residuals indicate lower firing rates than predicted by the baseline.

Judge family	Reports authored by	Batches	Models	Residual (pp)	t
OpenAI	Anthropic	9	5	-4.7	-2.99
OpenAI	OpenAI (own)	5	3	-26.6	-19.55
Anthropic	Anthropic (own)	9	5	-3.5	-2.73
Anthropic	OpenAI	5	3	-9.6	-6.64

C Reference Harness and Execution Protocol

C.1 Financial tools and evidence flow

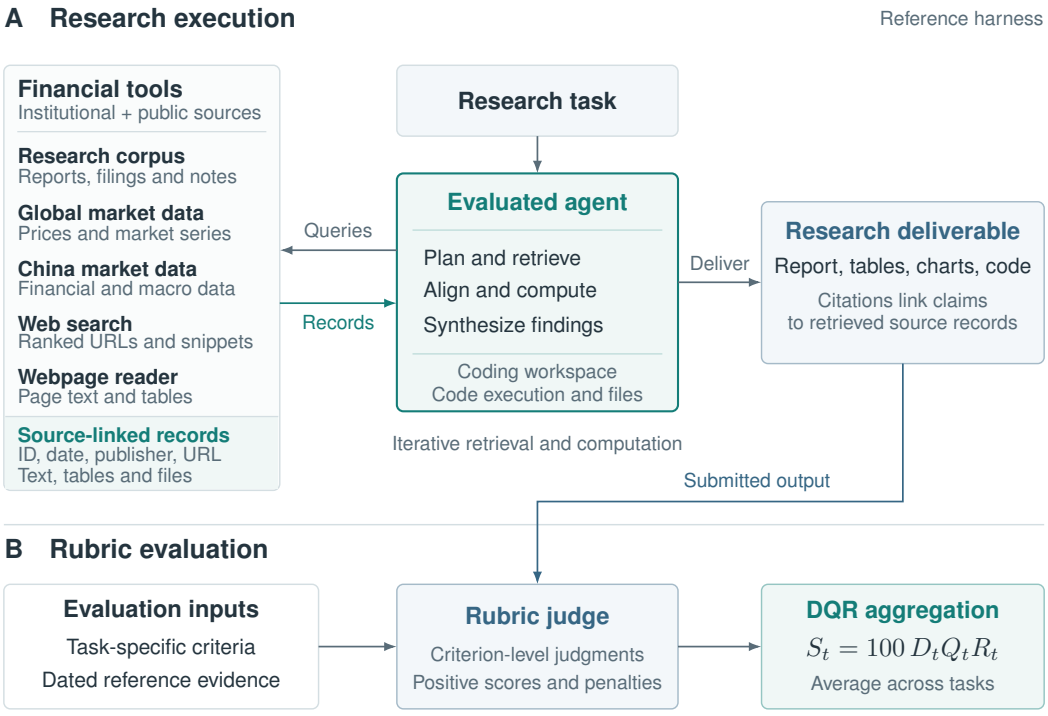

Figure 5: Research execution and evaluation in the reference harness. **A:** The agent queries five financial tools, receives source-linked records, and alternates retrieval and computation to produce a cited deliverable. **B:** The judge evaluates the deliverable against task-specific rubrics and dated reference evidence. Normalized judgments form D_t, Q_t and R_t; the benchmark averages the resulting task scores.

Table 7: The five financial tools, their output fields, and example returned records. The coding workspace supplies code execution and file operations.

Tool	Returned record	Example record
`search_research_corpus`	Report text with title, publisher, date, and source identifier.	Haitong Securities note on CATL (300750.SZ), 2025-03-18, projecting 487 GWh of 2025 battery shipments.
`search_global_market_data`	Daily bars: symbol, date, open/high/low/close, volume, change, and VWAP.	NVDA, 2026-01-30: close 191.13, volume 179.5M, -0.72%.
`search_china_market_data`	Indicator series: observation date, entity–metric name, value, and unit.	Kweichow Moutai (600519.SH) net profit attributable to shareholders, cumulative to 2025-12-31: CNY 82.3 bn.
`search_public_web`	Ranked results: title, domain, publication date, URL, and snippet.	`nvidianews.nvidia.com`, 2026-02-25: record Q4 FY2026 data-center revenue of $62.3 bn.
`read_public_webpage`	Full page text and tables with source metadata for a selected URL.	NVIDIA's Q4 FY2026 release on `sec.gov`, including its revenue-by-platform table.

C.2 Evidence records and an illustrative workflow

Tool responses retain the source identifier, title, publisher, available date, URL, content, an optional structured table, and metadata. Source markers link the delivered report to the retrieved records. These fields preserve provenance; deciding whether periods, entities, units and accounting conventions are comparable remains part of the agent's task.

For a capital-expenditure comparison across companies, corpus search can locate definitions and management guidance, while the two regional data tools retrieve disclosed figures. Web search and webpage reading recover recent issuer releases and inspect disputed definitions. The agent assembles a company–period table, distinguishes realized expenditure from forecasts, and calculates ratios using consistent denominators. Its conclusion must separate directly observed differences from assumptions about AI allocation. The appropriate retrieval sequence depends on the evidence available for each company.

C.3 Execution records and reproduction scope

Leaderboard runs were collected in the hosted production environment. The reference harness specifies tool contracts and artifact handling; equivalent outcomes with the production system have not been established. Comparisons should record the model configuration, reasoning effort, tool access, task set, execution date, reference version, rubric definitions and judge configuration. Licensed services may constrain evidence redistribution and historical replay.

Execution records include deliverables, source records, token consumption, wall-clock time and dispatched work units. These units include retrieval, page reading, parsing, file operations, shell commands, code writing and sub-agent dispatch, so their count extends beyond financial API requests. Appendix E.3 defines the grouping used in the tool-use analysis.

D Complete Evaluation Results

Table 8 reports scores, coverage, resource use and model release dates for all 58 configurations of 42 base models at the 2026-09-06 cut-off.

D.1 Coverage and resource measures

Configurations are ordered by their published DQR scores, with the number of scored tasks reported alongside each score. Task coverage should be considered when comparing scores and interpreting the cost–performance frontier in Figure 3A.

API cost, wall-clock time, token consumption and tool-call count are per-task means. Token counts include cached tokens. Costs use the prices attached to the runs, including base-rate estimates; ranges retain off-peak/peak prices. Model release denotes the base model's first public availability.

D.2 Complete configuration results

Table 8: Complete evaluation results at the 2026-09-06 cut-off (58 configurations). DQR is on a 0–100 scale; n is the number of scored tasks. Cost, time, tokens (millions, including cached tokens) and tool calls are per-task means. Model release denotes the date of first public availability (MM-DD in 2026). Cost ranges retain off-peak/peak prices.

Configuration	DQR	Cost ($)	Time (min)	n	Tokens (M)	Tool calls	Model release
GPT-6 Astra (low)	79.81	9.31	21.5	99	5.602	67.6	09-03
Claude Fable 5.1 (high)	78.50	5.11	14.6	101	3.521	28.5	09-01
GPT-6 Astra (high)	78.46	12.62	30.9	101	7.001	94.5	09-03
Grok 4.6 (high)	74.01	2.78	13.9	101	2.640	45.0	08-13
Claude Fable 5.1 (low)	73.32	2.80	9.4	101	1.913	17.3	09-01

Continued on next page

Table 8 continued

Configuration	DQR	Cost ($)	Time (min)	n	Tokens (M)	Tool calls	Model release
Grok 4.6 (low)	71.45	1.03	5.0	101	1.017	20.7	08-13
hy4-preview (high)	69.71	0.82	15.8	101	3.008	30.9	08-28
GLM 5.3 (max)	68.68	1.21	9.8	101	1.935	24.5	08-14
Muse Spark 1.3 (xhigh)	67.48	0.90	9.0	101	1.960	19.1	09-02
Seed Evolving 8/27 (high)	67.40	0.78	15.8	101	1.859	20.5	08-27
Claude Opus 5 (max)	66.61	5.16	17.8	101	2.529	23.8	07-24
GPT-5.6 Sol (medium)	66.56	4.30	13.3	101	3.922	49.2	07-09
Claude Opus 5 (high)	66.39	4.53	14.0	101	2.238	20.3	07-24
Kimi K3 (max)	64.90	1.54	12.7	101	1.568	22.3	07-16
Claude Fable 5 (high)	64.45	3.81	9.8	96	1.258	15.0	06-09
Grok 4.5 (low)	64.32	1.29	6.9	100	1.244	26.6	07-08
GPT-5.5 (xhigh)	64.31	3.32	10.5	101	2.393	32.5	04-24
DeepSeek V4 Pro (max)	64.23	0.77–1.26	12.3	101	2.194	25.2	08-13
Qwen 3.8 Max (xhigh)	64.20	1.66	23.5	100	1.945	24.9	08-03
Grok 4.5 (high)	63.40	1.31	6.3	101	1.232	27.0	07-08
Qwen 3.8 Flash (max)	62.88	0.34	29.6	101	3.849	39.6	08-26
GPT-5.6 Terra (medium)	62.60	1.73	7.3	101	2.374	33.2	07-09
Claude Opus 5 (no reasoning)	62.44	4.11	11.2	101	2.725	23.1	07-24
Claude Opus 4.8 (max)	61.79	3.48	15.8	101	2.237	20.7	05-27
DeepSeek V4 Flash (max)	59.58	0.21	8.0	101	2.013	24.6	07-31
GLM 5.3 Flash (max)	59.45	0.15	16.3	101	1.697	22.1	08-26
DeepSeek V4 Pro (high)	59.11	0.19–0.35	7.5	101	1.056	14.8	08-13
hy3 (high)	58.47	0.21	8.2	101	0.684	13.8	07-06
DeepSeek V4 Pro (low)	57.92	0.24–0.43	6.5	101	1.061	14.2	08-13
GPT-5.5 (no reasoning)	57.90	3.20	12.7	101	1.651	21.5	04-24
Gemini 3.8 Flash (high)	57.43	0.71	5.2	101	2.054	20.2	09-02
Claude Sonnet 5 (high)	56.56	1.60	8.6	101	1.865	20.4	06-30
Claude Opus 4.8 (no reasoning)	54.90	1.63	5.8	101	1.595	15.6	05-27
Seed Evolving 8/1 (high)	54.82	0.47	8.2	100	1.104	14.8	08-01
hy3 (low)	54.42	0.20	6.0	101	0.801	13.8	07-06
GLM 5.2 (max)	52.65	0.63	10.3	101	0.923	14.8	06-13
Seed Evolving 7/14 (high)	51.92	0.49	6.2	101	1.018	14.7	07-14
Muse Spark 1.2 (xhigh)	51.27	1.21	7.4	101	1.880	17.4	08-06
Seed 2.1 Pro (high)	51.24	0.55	6.9	101	0.957	14.0	06-23
Kimi K2.7 Code (reasoning)	49.43	0.59	6.6	99	1.428	22.7	06-12
Seed 2.1 Turbo (high)	49.03	0.39	7.3	101	0.933	15.6	06-23
GLM 5.2 (no reasoning)	48.54	0.52	4.1	101	0.789	13.9	06-13
hy3 (nothinking)	48.21	0.19	4.8	101	0.797	15.4	07-06
Qwen 3.7 Max (reasoning)	48.17	0.46	7.5	101	0.811	13.0	05-20
Gemini 3.7 Flash (high)	46.92	0.56	4.0	101	1.239	13.7	08-14
GLM 5.1 (reasoning)	46.75	0.49	8.3	101	0.690	12.4	03-27
Seed 2.1 Pro (no reasoning)	46.74	0.79	6.8	101	1.120	16.8	06-23
DeepSeek V4 Pro Preview (high)	46.37	0.60	11.7	101	0.963	13.7	04-24
Seed 2.1 Turbo (no reasoning)	45.74	0.55	6.9	101	1.102	17.1	06-23
Mimo V2.5 Pro (reasoning)	42.79	0.51	7.6	101	0.752	11.5	04-22
MiniMax M3 (adaptive)	41.27	0.08	4.8	101	0.522	11.5	05-31
Kimi K2.6 (reasoning)	40.95	0.41	8.4	101	1.075	18.0	04-20
Inkling (max)	40.44	1.01	8.0	101	2.022	21.0	07-18
Gemini 3.6 Flash (adaptive)	38.89	0.62	3.6	101	0.764	9.5	07-21
Gemini 3.5 Flash (medium)	37.17	1.55	5.9	101	1.293	13.4	05-19
Step 3.7 Flash (high)	34.14	0.11	6.8	101	1.541	19.6	05-29
hy3-preview (high)	30.11	0.12	2.6	101	0.538	10.3	04-23
hy3-preview (low)	27.24	0.04	3.4	101	0.678	12.7	04-23

E Additional Capability and Tool-Use Analyses

E.1 Release-date trends in published scores

The release-date analysis uses the 58 configurations at the September 6 cut-off. Each configuration is assigned its base model's first public release date. The earliest and newest release tertiles contain 22 and 19 configurations, respectively. Configurations of one base model are not independent model releases; these comparisons describe associations across release cohorts.

Table 9 summarizes the leaderboard's seven published criterion means and the three aggregation factors. Penalty criteria are displayed as penalty-free rates so that higher values consistently indicate better performance. Linear slopes are expressed per 30 days.

Table 9: Release-date trends in the published criterion and factor scores for 58 configurations. r is Pearson correlation with release date; slope is the fitted change in score points per 30 days. Earliest and Newest are release-tertile means ($n = 22$ and 19) on the 0–100 scale; Change is Newest minus Earliest in score points, subject to rounding.

Criterion or factor	r	Slope	Earliest	Newest	Change
Data quality	0.68	+2.6	74.7	82.7	+8.0
Distortion (none)	0.60	+4.1	70.8	85.5	+14.7
Inconsistency (none)	0.32	+3.5	60.0	72.0	+12.0
Data timeliness	0.69	+5.5	65.5	81.8	+16.4
Source authority	0.44	+2.1	65.3	72.1	+6.8
Analysis depth	0.59	+2.6	84.4	91.7	+7.3
User-need coverage	0.52	+2.1	80.3	86.7	+6.4
D	0.71	+3.1	73.2	83.0	+9.8
Q	0.54	+1.8	85.9	91.1	+5.3
R	0.55	+4.0	68.5	82.7	+14.2

E.2 Trends across evaluation criteria

Figure 6 covers all 15 criteria bound to at least ten tasks. Its means are computed from per-task judge records matched to each configuration. Some records come from a different judge pass than the published leaderboard snapshot, so their criterion means and cohort differences can differ from those in Table 9.

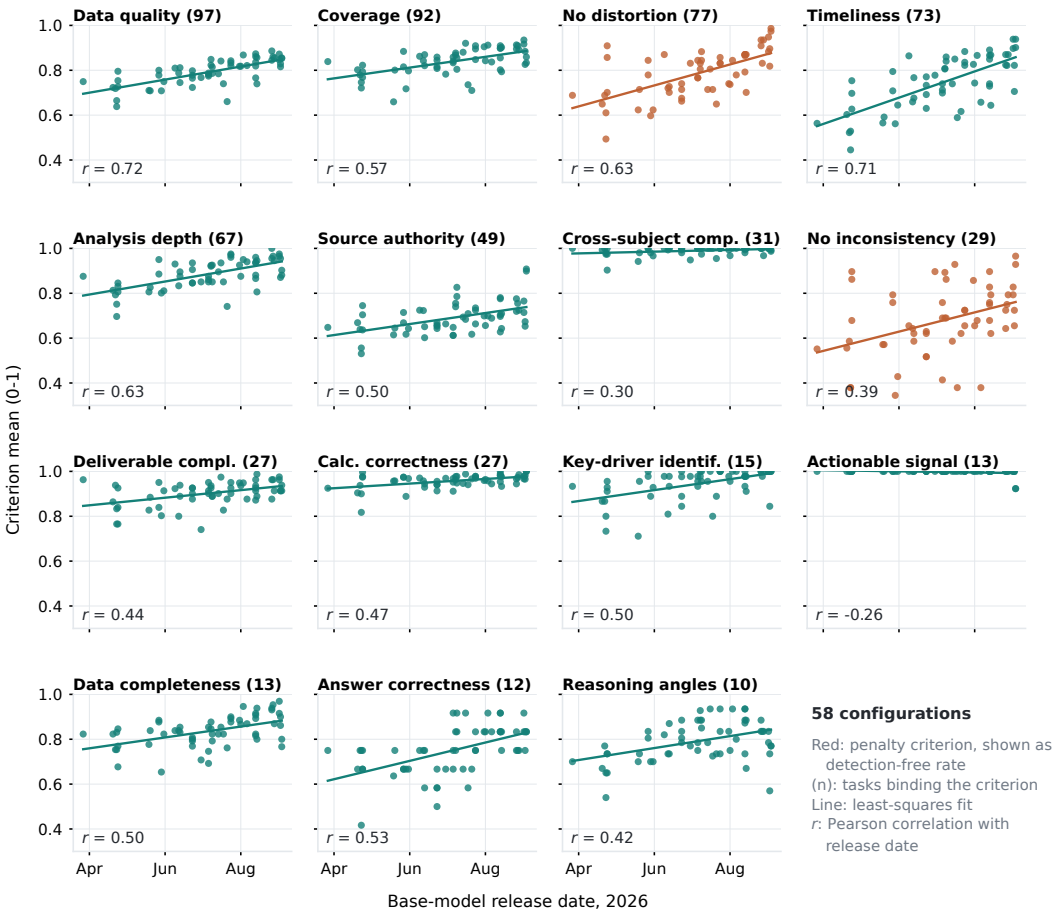

Figure 6: Criterion means against base-model release date for the 58 configurations at the 2026-09-06 cut-off. Each point is a configuration-level mean on a 0–1 scale; lines show least-squares fits, and r denotes Pearson correlation. Penalty criteria are shown as detection-free rates. Parenthetical counts give the number of tasks using each criterion.

Four criteria show limited headroom in the newest release tertile, with means at or above 97 on the 0–100 scale (cross-subject comparison 99.6, actionable signal 99.2, key-driver identification 97.8, calculation correctness 97.3).

The four largest gains from the earliest to the newest tertile are timeliness (+17.2), no distortion (+16.4), answer correctness (+16.4) and no inconsistency (+14.9), all evidence criteria.

E.3 Tool-use definitions and release-date trends

A tool call is one distinct work unit in the execution trace, including sub-agent dispatch. Repeated events for the same unit identifier are counted once. Execution comprises shell commands, file operations, code writing and plotting. Retrieval comprises deep and quick retrieval, document parsing, database queries, web search and browsing. Sub-agent dispatch is counted separately; task management, skill reading and unclassified operations form the remaining category.

For release-date trends, we regress the logarithm of each configuration's mean work-unit count per task on its base model's release date and express the fitted slope as a multiplicative change per 30 days.

E.4 Illustrative judgments about evidence

Two Astra high reports illustrate the evidence-selection behavior discussed in Section 3.2. In the IPO task (q0806), the report identifies a prospectus resubmitted in July and distinguishes its financial period ending April 30, 2026, from the older version's period ending September 30, 2025. It also labels commissioned industry forecasts as projections rather than realized outcomes.

In the amino-acid capacity task (q0831), the report prioritizes Fufeng's annual report over an industry table retaining older values. For Dongxiao, it distinguishes nameplate capacity listed on the company website from a production suspension disclosed in a formal inquiry response, and uses the latter when interpreting effective supply. These examples describe judgments observed in the delivered reports.

F Failure-Mode Analysis and Cases

F.1 Case selection

The case analysis draws on 2,415 task scorings from the 24 batches in the judge-family study: ten neutral-authored, nine Anthropic-authored and five OpenAI-authored batches. Twenty-one configurations appear on the September 6 leaderboard; three are outside that snapshot. The primary findings come from the Opus judge, with the alternate judge used as a cross-check.

For numerical distortions, we inspect the judge's severe-error findings and the corresponding report excerpts; internal conflicts are identified through the inconsistency criterion. Examples of reliance on ready-made analyses are selected by reading the task, report and judge findings together. These cases illustrate the three failure patterns in Section 3.3; they do not estimate their prevalence.

F.2 REPORT EXCERPTS AND JUDGE FINDINGS

Table 10: Report excerpts and judge findings for the three failure patterns of Section 3.3. Excerpts are translated from Chinese with figures and units unchanged; judge findings are quoted or summarized. Labels such as summary, body and table identify the checked locations in the report.

Configuration and task	Report says	Judge finds
Order-of-magnitude errors (`data_distortion` detected)		
GPT-6 Astra high, q0834 silicon-wafer industry	"2026H1 revenue 215,016 million yen (+4.7%), net loss attributable to parent 128,950 million yen"	Gold: net loss 12,895 million yen. "A 10× error on SUMCO net loss."
Claude Fable 5.1 high, q0818 Apple WWDC and supply chain	Luxshare Precision: "Net profit attributable to parent 0.784 bn yuan (+18%, dragged by FX losses)"	Gold: 7.843 bn yuan (+18.04%). "10× understatement."
Grok 4.6 high, q0834 silicon-wafer industry	"Shin-Etsu Chemical about $744.1 bn / 501.8 bn yuan (whole chemicals group, not the wafer segment)"	Should be about $74.4 bn. "A 10× order-of-magnitude currency error on a load-bearing global-leader metric."
Claude Opus 5 no reasoning, q0865 polyester-chain Excel workbook	"FY25 net profit YoY (%): 44.03% | 1709.21% | 10582.31% | 1042.67% | 6913.20% | −741.93%"	Gold: +0.44% for the first company. "All of these are exactly 100× the gold value, a decimal/percentage scaling error."
Conflicts across a long deliverable (`data_inconsistency` detected)		
GPT-6 Astra low, q0827 humanoid-robot supply chain	Summary: "planetary roller screw price has fallen from thousands of yuan to about 800". Body: "2025 unit price about 2,000 yuan, long-term target 800"	"Same-metric mismatch (current screw price): 800 vs 2,000 yuan (2.5×). Material inconsistency."
Claude Fable 5.1 low, q0856 HVDC market sizing	Prose: "global HVDC 2027/28/29/30 about 332 / 615 / 958 / 1,356 hundred-million yuan (2030 baseline about 191 hundred-million USD)". Table: 19.1 hundred-million USD for the same 2030 total	"Prose says 191 and the table 19.1 (hundred-million USD) for the same 2030 global total. A material 10× internal contradiction."
Claude Opus 5 high, q0856 HVDC market sizing	"China 2027–2030: 60 / 84 / 102 / 135 hundred-million yuan" and, in the core-conclusion section, "about 65 / 92 / 120 / 160"	"The same subject, metric and period with two different value sets in the same document."
Reliance on ready-made analyses		
Claude Opus 5 max, q0852 lithium price and battery-sector transmission	"120–180 thousand yuan/t is the profit-optimal band, reasonable centre 150 thousand; below 120 midstream materials lose money, above 200 storage projects are deferred"	`user_need_coverage` 2 of 4: "presents several sell-side views but they mostly agree; does not present the bull (UBS 200, Citi 250) versus bear (Goldman 20% surplus, Morgan Stanley 120) disagreement."
GPT-5.5 (June run), q0875 WF_6 company analysis	"Signed a three-year exclusive agreement with TSMC, with its supply share rising from 10% to 70%" (citing an expert-note repost)	`data_distortion` detected: an unconfirmed claim is stated as fact despite the company's clarification in the June reference that no new long-term or substantial orders had been signed.

G Evidence Turnover under Repeated Evaluation

We examine whether rerunning an investment research task recovers the same cited sources, including when its reference evidence is unchanged. The citation comparisons support the finding in Section 3.2; score comparisons and reference-version history provide context for interpreting repeated evaluations.

G.1 Run pairs and cited-document identity

The citation study contains seven run pairs and 698 task-pair observations: Opus 4.8 at three dates; Fable 5, GLM 5.2 and MiniMax M3 at two dates each; and a same-day Grok 4.6 comparison through two API paths. To match requests across runs, we remove whitespace from query text and compare its first 60 characters. The reports contain 104,697 visible citation markers, from which we extract source type, publisher, title and any recorded date.

A cited document is identified by its publisher and title, with surrounding whitespace removed. Repeated citations to the same document count once in the document set; publishers are likewise counted once in the publisher set. Source types distinguish announcements, broker reports, web pages and other categories recorded by the tools.

G.2 Overlap and retention

For citation sets A and B, overlap is the Jaccard index $|A \cap B|/|A \cup B|$; an empty union has no numerical overlap value. Retention is $|A \cap B|/|A|$, the fraction of the earlier set cited again. We compute document and publisher overlap for each task pair. Retention by source type aggregates retained and earlier-document counts within that type.

The same-day comparison uses a different model from the elapsed-time comparisons, and the three Opus pairs share runs. These comparisons therefore describe source variation without isolating the effect of elapsed time. MiniMax's citation count falls from 29 to 13 per task; we exclude that pair from the citation-overlap ranges summarized below.

G.3 Citation turnover under unchanged references

Within the 12-day Opus 4.8 comparison, 14 tasks have identical reference evidence. Their cited documents have a median Jaccard overlap of 0.12. Thus, even with unchanged reference evidence, rerunning a task can produce a substantially different set of cited sources. This is a subset of the 12-day comparison; Figure 7A summarizes entire run pairs.

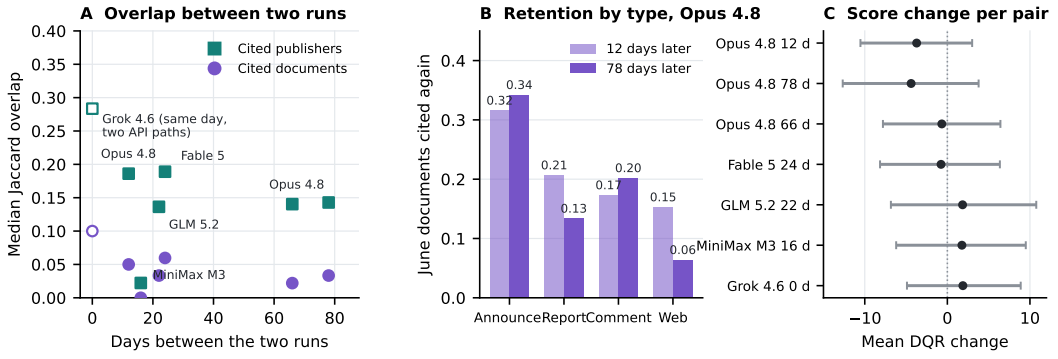

Figure 7: Cited sources and scores across seven run pairs. **A:** Median document and publisher Jaccard overlap against the interval between runs. The zero-day pair compares two API paths for Grok 4.6. MiniMax M3 (16 days) is excluded from the overlap ranges in the text because its citation count fell sharply. **B:** Fraction of Opus 4.8's June documents cited again after 12 and 78 days, by source type. **C:** Mean DQR change with a 95% task-bootstrap interval.

Source turnover also occurs across the broader set of task pairs. The same-day comparison has median document and publisher overlap of 0.10 and 0.28. For runs 12 to 78 days apart, excluding MiniMax, the corresponding medians range from 0.02 to 0.06 and from 0.14 to 0.19. When restricted to tasks with at least five cited documents at each date, both the static-reference and refreshing-reference groups have median document overlap of 0.05.

Retention differs by source type (Figure 7B). Opus 4.8 retains a larger fraction of company announcements than of broker reports or web pages. Announcement retention is 0.32 after 12 days and 0.34 after 78 days; the corresponding values are 0.21 and 0.13 for broker reports, and 0.15 and 0.06 for web pages.

G.4 Citation turnover and score changes

Across the seven run pairs, mean DQR changes range from -4.4 to $+1.9$ points. The 95% intervals, computed from 10,000 bootstrap resamples of paired tasks, all include zero (Figure 7C). These intervals do not establish equivalence between runs. Individual tasks can nevertheless change considerably: median absolute DQR change is 13 points across the elapsed-time pairs and 5 points in the same-day comparison.

Implication for evidence audit. The observed turnover makes retaining the original cited sources necessary for auditing a particular report: a later run may produce a different evidence set even under unchanged references. Document identity and factual support are distinct properties, since different documents may support the same claim. This motivates evaluating whether citations support the associated claims separately from whether a repeated run recovers the same sources.

G.5 Scores across evaluation dates

The longitudinal comparison contains 23 runs across six configurations between June 11 and September 12, 2026. Tasks are matched by query text because identifiers changed between the June and later batches. Inclusion requires the static-reference and refreshing-reference task counts to sum to the overall count, and their count-weighted mean to agree with the overall mean within 10^{-4} points. One GLM 5.2 run on July 14 fails these checks and is excluded.

Within each configuration, the difference between the highest and lowest mean DQR across retained dates ranges from 0.7 to 4.4 points. Each mean covers the tasks answered in that run. Execution routes, judging and task coverage vary across runs, so these differences describe repeated evaluation in a live environment rather than a controlled change in model capability.

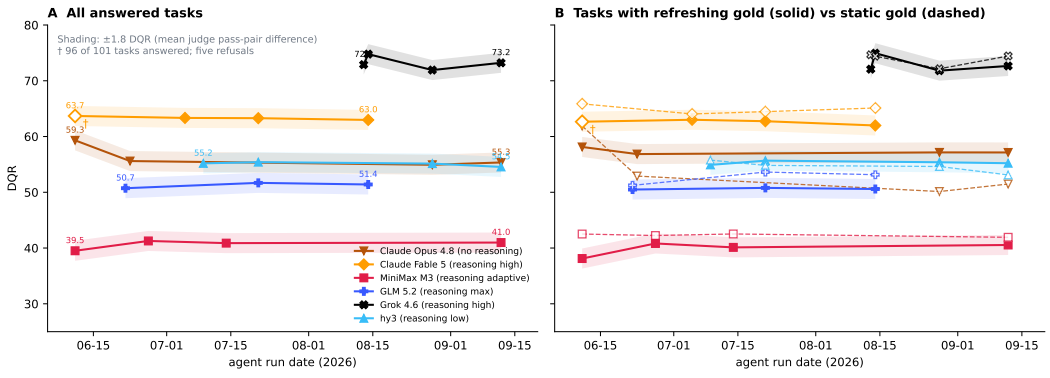

Figure 8: DQR across evaluation dates. **A:** Mean score over answered tasks. Shading shows ± 1.8 points, the rounded mean judge pass-pair difference in Table 5; it is not a confidence interval. **B:** The same runs split into refreshing references (solid) and static references (dashed). The open marker (†) denotes partial task coverage, detailed in Table 11.

Table 11: Run dates, coverage and scores for Figure 8. Runs use the reference current on the stated date. Mean DQR covers the n answered tasks, overall and by reference-refresh policy. Range is the difference between the highest and lowest retained scores. †: 96 of 101 tasks were answered, with five refusals on investment research requests about innovative drugs.

Configuration	Run date	n	All	Refreshing gold	Static gold
Claude Opus 4.8 (no reasoning)	2026-06-11	101	59.3	58.1	61.9
	2026-06-23	101	55.6	56.9	52.9
	2026-08-28	101	54.9	57.2	50.1
	2026-09-12	101	55.3	57.1	51.5
range			4.4	1.3	11.7
Claude Fable 5 (reasoning high)	2026-06-11	96†	63.7	62.7	65.9
	2026-07-05	101	63.3	63.0	64.0
	2026-07-21	101	63.3	62.7	64.5
	2026-08-14	101	63.0	62.0	65.1
range			0.7	1.0	1.8
MiniMax M3 (reasoning adaptive)	2026-06-11	101	39.5	38.1	42.5
	2026-06-27	101	41.3	40.8	42.2
	2026-07-14	101	40.9	40.1	42.5
	2026-09-12	101	41.0	40.5	41.9
range			1.8	2.7	0.6
GLM 5.2 (reasoning max)	2026-06-22	101	50.7	50.5	51.2
	2026-07-21	101	51.7	50.8	53.6
	2026-08-14	101	51.4	50.6	53.2
range			1.0	0.3	2.4
Grok 4.6 (reasoning high)	2026-08-13	101	72.9	72.1	74.7
	2026-08-14	101	74.8	74.9	74.4
	2026-08-28	101	71.9	71.8	72.2
	2026-09-12	101	73.2	72.7	74.4
range			2.9	3.1	2.5
hy3 (reasoning low)	2026-07-09	101	55.2	54.9	55.8
	2026-07-21	101	55.4	55.7	54.8
	2026-08-28	101	55.1	55.4	54.6
	2026-09-12	101	54.5	55.2	53.1
range			0.9	0.8	2.7

G.6 Reference-version inventory

The September 17, 2026 inventory records 34 tasks with scheduled refreshes and at least one archived refresh, comprising 181 archived versions. One task without a scheduled refresh, q0883, has an archived correction. Each recorded transition changed reference content. These counts describe the available version history; the refresh policies and date fields are defined in Appendix A.4.

H Code and Data Availability

Benchmark repository:
https://anonymous.4open.science/r/investment-research-agent-benchmark-20E1/

Agent harness repository:
https://anonymous.4open.science/r/investment-research-agent-harness-4C14/