22 July 2026

Frontier Financial Judgement: Can agents tell what might move a stock?

Sonto[*]

Abstract

We introduce Frontier Financial Judgement, a challenging new benchmark developed in collaboration with professional equity analysts to assess agents' ability to replicate expert human judgements. Rapidly identifying new information, evaluating its implications and determining its valuation impact is one of the most time-consuming and challenging aspects of real-world equity coverage. This is becoming ever more difficult and important as AI rapidly increases the quantity of new information to process. The strongest agent we evaluate on Frontier Financial Judgement matches all expert labels in only 52.4% of cases. We also find significant divergence in estimated false-positive rates among frontier agents, ranging from ~1% for GPT-5.6 Sol to ~32% for Claude Sonnet 4.6. To construct the benchmark and make it representative of real-world settings, we combine human-designed and labelled synthetic articles with live news articles and historical documents, creating 656 items for assessment. The resulting task requires agents to distinguish genuinely new, valuation-relevant financial information from stale, immaterial or misleading news under realistic conditions. We find substantial trade-offs among agent accuracy, cost, false positives and reliability that continue to hinder the reliable deployment of news-flow filtering in practice.

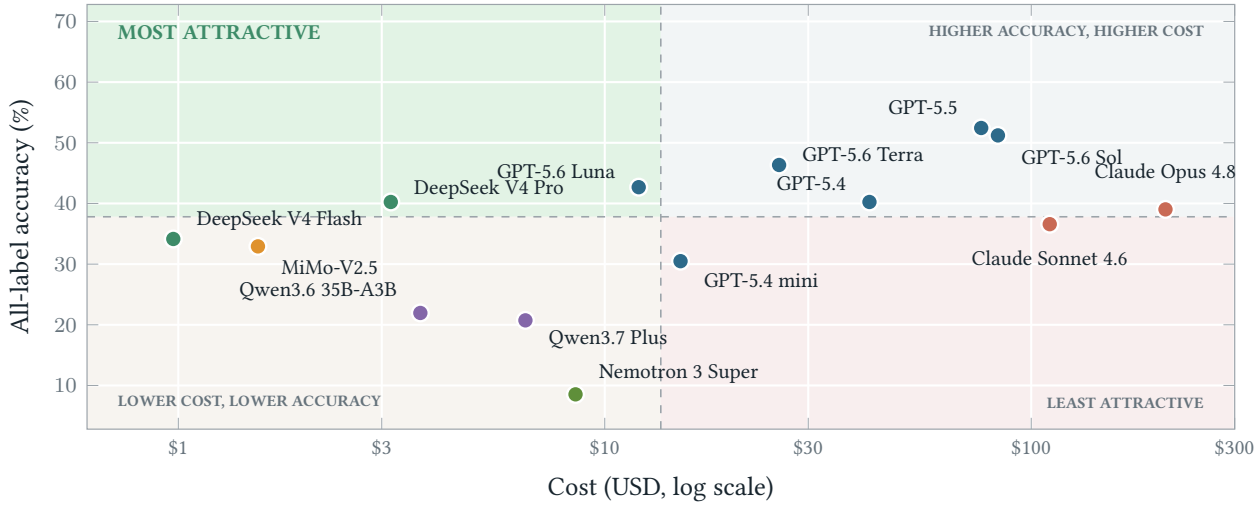

Figure 1. Cost and strict all-label accuracy on Frontier Financial Judgement.

[*]Corresponding author: joshua.harris@sontoresearch.ai

1 Introduction

One of the most time-consuming and important parts of real-world buy-side equity analysis is rapidly assessing new events and news flow for potential market-moving information [5, 6]. This task is challenging for a number of reasons: (1) it requires deep knowledge of what historical information has already been released [29], (2) it often requires complex financial analysis to disentangle the underlying change [1, 30], (3) it is highly time-sensitive [27], (4) it can cover any aspect of a company's business or related markets, (5) news flow occurs in large volumes across every asset [3, 14, 18], and (6) a tiny fraction of overall news flow should be flagged [5, 18]. Most importantly, it requires something akin to financial "taste" where experts can rapidly identify subtle changes that imply significant shifts in company performance or sentiment [6, 25, 30]. This task is also becoming increasingly difficult as AI-assisted analysis and reporting drive growth in the volume of new information. These features make it both a highly challenging test for AI agents and a potentially crucial use case.

In addition to the core task of identifying significant news flow, assessing the impact of new events also forms part of many other important financial tasks. For example, in order to update a model, an analyst will often need to review the latest developments to inform changes to model assumptions and projections [1, 20]. In order to value a stock or assess a potential acquisition, analysts will often also consider whether there are any recent relevant market, economic, or company events that should inform expectations [9, 20].

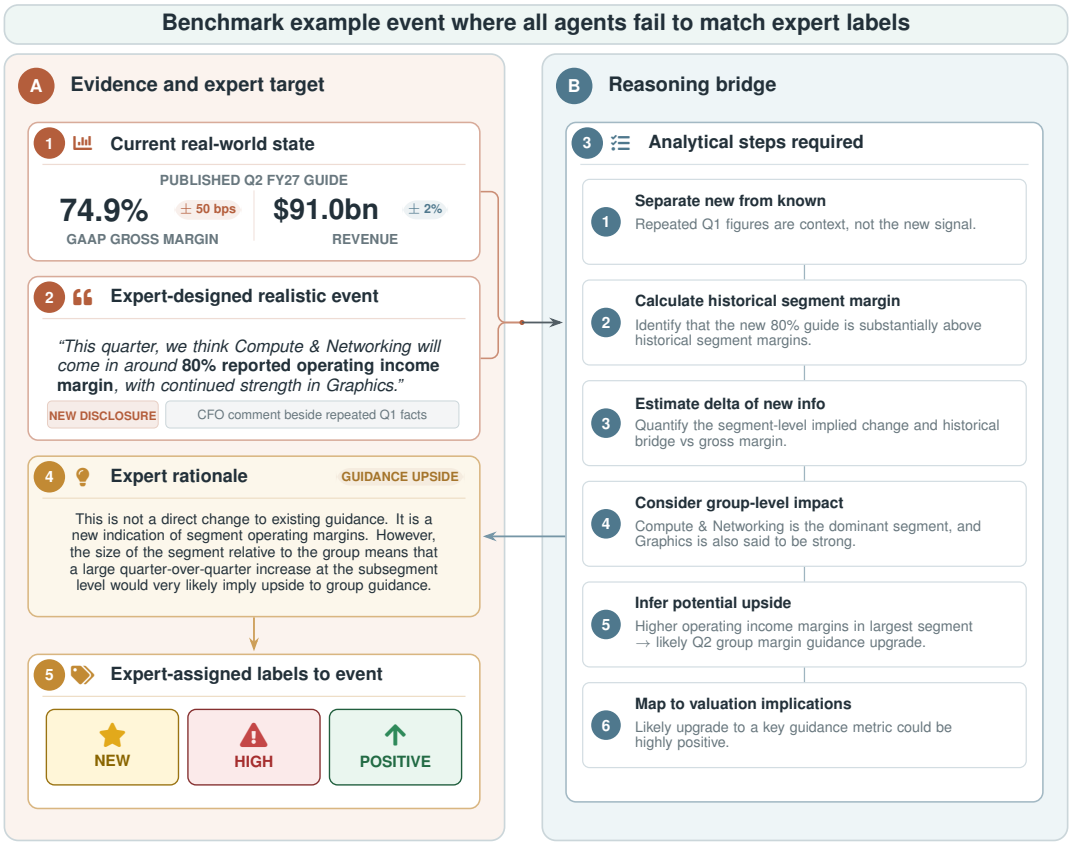

Figure 2. Example of a Frontier Financial Judgement benchmark case that requires a segment-level profitability disclosure to be translated into likely consolidated guidance upside.

Therefore, assessing the impact of new company news flow and events captures many of the core skills an expert equity analyst requires. However, evaluating AI agents on this task is also problematic. Valuation-

relevant real-world news flow is sparse, highly clustered, and suffers from the look-ahead problem [5, 12, 14]. In this work, we introduce a novel approach for robustly evaluating AI agents' capabilities on this task.

To build Frontier Financial Judgement, we collaborate with expert equity analysts to design realistic synthetic news flow grounded in current real-world company reporting. This allows us to control the difficulty, type, and novelty of the event while ensuring the event is a plausible development that could actually occur. We then mix synthetic news flow with live articles and documents to replicate the conditions an AI agent would often face in real-world usage. By combining purely synthetic events with actual articles, we allow agents to access normal web search tools while still ensuring that there is minimal risk the correct answer will leak from online sources [17, 32].

2 Related work

2.1 Financial LLM and agent evaluations

Historical finance LLM evaluations often used question answering to assess the financial knowledge and reasoning capabilities of LLMs. Earlier work primarily focused on static document collections; for example, FinanceBench evaluates open-book question answering over public company filings [16].

More recently, a number of benchmarks have evaluated LLM-based agents on broader financial research tasks with access to search engines, SEC filings and specialist tools [2, 15, 31]. Related work has also measured document and passage retrieval, or used structured rubrics and logic trees to assess intermediate research steps [8, 28]. Most recently, DiligenceBench evaluates agents, defined by their model and harness, on open-ended equity-research questions, with answers scored against dense task-specific rubrics [26].

2.2 Financial news and point-in-time evaluations

Financial news evaluations have most commonly focused on sentiment classification, event extraction, or subsequent stock movement [21, 35]. For example, Chen et al. [7] introduce an event-level financial sentiment task which jointly extracts the relevant company, event and sentiment from news articles. More recent Retrieval-Augmented Generation (RAG) work has extended this to temporally constrained financial analysis using news, filings, prices and other data [34, 36]. Point-in-time financial RAG is particularly relevant as it evaluates news-triggered event-impact prediction using only contemporaneous evidence. However, it uses subsequently realised market returns as labels, which may conflate event interpretation with wider market conditions.

The closest prior work to our label structure is by Bluteau et al. [4], which uses LLMs to annotate ESG news for novelty, relevance, materiality and severity. Separately, ExAnte evaluates whether LLMs comply with fixed information cutoffs and measures temporal leakage [22]. These works motivate explicit novelty and materiality labels and a reproducible evidence cutoff, but do not jointly evaluate prior-information retrieval, valuation impact and directional reasoning using web-enabled agents.

2.3 Financial news filtering

Our work with AI agents builds on the established finance literature investigating whether human investors distinguish genuinely new information from repeated or recombined news. In particular, Tetlock [29] defines the staleness of company news using textual similarity to prior articles and finds that investors continue to react to stale information, with subsequent return reversals. More recently, Fedyk and Hodson [11] show that even finance professionals can struggle to identify old information when it is recombined from multiple

sources, and that these recombinations generate larger price responses than direct reprints. These studies establish that filtering previously disclosed information is an economically meaningful component of financial analysis, rather than a purely linguistic novelty task.

2.4 Financial misinformation and verification

Finally, this work also intersects with previous research investigating financial claim verification and misinformation detection. FinDVer evaluates evidence-grounded claim verification over long financial documents, whilst other work introduces classification and explanation benchmarks for financial misinformation [23, 33]. More recently, RFC-Bench uses minimally perturbed real financial news to test counterfactual changes, and AuditFraudBench evaluates misleading disclosure narratives using company filings and regulatory evidence [19, 24].

These evaluations are relevant because financial articles can remain plausible whilst being misleading through accounting, scope, causal, or temporal framing. However, determining whether an article is false is distinct from determining whether it contains new information that is material to a company's valuation. An article may be factually correct whilst recycling prior disclosures, reporting economically irrelevant details, or presenting indirect industry information with limited read-through. Therefore, there remains an important gap for an agent benchmark that jointly evaluates novelty, materiality and directional impact at a fixed evidence cutoff, including under realistic information overload.

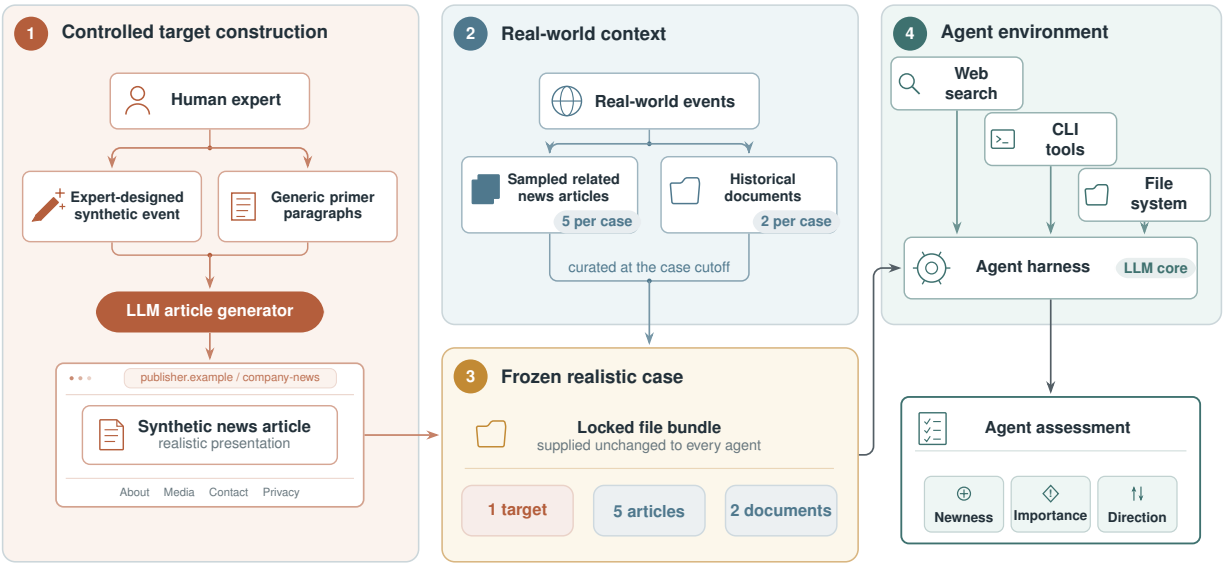

Figure 3. Overview of the benchmark setup.

3 Methods

Assessing real-world financial news can require a wide spectrum of financial analysis, company context, and in-depth research. Online news articles and press releases are also often adversarial, repeating old disclosures, containing subtle wording or factual changes, overstating significance, or drawing tenuous conclusions. This makes it a hard and very general task for financial AI agents.

However, evaluating agents on this task is particularly challenging because real valuation-relevant news is sparse, very hard to isolate from other similar stories, and often ambiguous, making a live real-world

benchmark intractable. To address these issues, in this work we use a combination of live news and synthetic expert-designed articles. This allows us to replicate very closely what an agent would face in the real world, but also to control the type, quantity, and difficulty of the news flow on which agents are evaluated.

3.1 Human expert event generation

We ask professional hedge fund and equity analysts to design plausible company-specific events. For each event, the expert provides a description of the underlying business or market development, supporting evidence, gold labels, and a label rationale.

Experts assign three labels. *Information new* indicates whether the article contains a factual or substantive qualitative development that was not previously public. *Expected importance* is the expected impact of the information on company valuation: none, low, medium, or high. *Direction* records whether the expected effect is positive, negative, neutral, or unclear. For genuine boundary cases, experts can specify more than one accepted importance or direction label. The expert-designed events are created to represent common but challenging forms of financial news; see Table 1 for examples of synthetic events relating to Company A.

To construct the final articles, we combine the expert event and evidence with generic primer paragraphs. An LLM renders these inputs in different source styles and article formats, including breaking news, market commentary, and analysis notes. Finally, we apply web-page chrome templates, which add realistic surrounding content, such as navigation, reader controls, advertisements, and footers, so that the inputs closely resemble articles collected from the web.

3.2 Benchmark setup

In order to closely replicate noisy real-world conditions, we place each synthetic article within a larger company-specific information bundle and use a short, realistic human instruction. Each case contains the labelled synthetic article, five recently collected real articles, and two historical company documents, all of which the agent must assess. This reflects how agents will often have to filter many pieces of news at once, without highly detailed instructions, and will often do so as part of a longer task (e.g. updating model assumptions).

3.3 Dataset

Our v1 benchmark dataset includes 82 synthetic events and a total of 656 items to assess. We focus initially on semiconductor supply-chain companies, including ASML, NVIDIA, Ciena, and Infineon. To avoid any risk of confusion with real events, we have removed company names from the examples in this paper; however, the real company names are provided to the agents during benchmark runs.

This dataset is designed to cover the full range of common news flow and events; representative examples are shown in Table 2. For this version of the benchmark, all cases and live data were frozen as of 17 July 2026.

Table 1. Illustrative Company A events used in Frontier Financial Judgement. The examples span recycled disclosures, roadmap changes, interpretation of opportunity size, and financially material read-through from detailed guidance.

Synthetic event	Details	Labels	Label rationale
Margin variability restated	A new CFO interview says memory pricing could make second-quarter gross margin fluctuate by up to half a percentage point.	**New:** No **Imp.:** None **Dir.:** Neutral	The same gross-margin variability, up to half a percentage point, was already contained in first-quarter guidance and prior CFO commentary; fresh article framing adds no new valuation-relevant information.
Rubin Ultra timing slips	The CEO places Rubin Ultra availability in early 2028, versus the previously communicated second-half 2027 window.	**New:** Yes **Imp.:** High **Dir.:** Negative	The new date implies a possible one-to-two-quarter delay, weakening the annual product-cadence thesis that supports Company A's competitive position against custom accelerators.
Vera folded into Rubin opportunity	Investor relations says the standalone Vera CPU opportunity should be viewed within the broader Vera Rubin outlook, rather than stacked as a separate incremental pool.	**New:** Yes **Imp.:** Medium[a] **Dir.:** Negative	It was previously stated by Company A management that the Vera CPU opportunity was standalone and in addition to Rubin estimates. The change therefore reduces overall stated opportunity size, although it could partly reflect presentation rather than final economics.
Subsegment margin signals guide upside	The CFO says Compute & Networking operating margin could reach approximately 80% in the second quarter, alongside continued strength in Graphics.	**New:** Yes **Imp.:** High **Dir.:** Positive	An 80% operating margin in Company A's largest reporting group would likely imply consolidated second-quarter performance above the company's existing top-level guide.
Positive commentary recycled	An article presents familiar claims about frontier-model share, Vera CPU demand, LPX, and AI-factory expansion as fresh upside.	**New:** No **Imp.:** None **Dir.:** Neutral	Each theme had already been discussed publicly. There is no new numerical guidance, customer commitment, product-schedule change, or financial update.

Notes: Labels report information newness, expected importance, and expected direction; [a] high importance is also accepted for this event because the disclosure may be interpreted as a reduction in total opportunity rather than only a reporting change.

Table 2. Representative synthetic financial-news categories in Frontier Financial Judgement, with abbreviated examples from expert-designed events.

Type of synthetic news	Example	Challenge
Guidance restatements	Company B calls 43% full-year gross margin consistent with its original forecast, despite its latest 44.5–45% guide.	Track the guidance chronology: this is an implicit 150–200 basis-point cut despite continuity framing.
Economic news	Taiwan electronics exports fall 6%, while AI-relevant integrated-circuit exports rise 5%.	Decompose the aggregate and map company exposure; the headline points in the wrong direction.
Restatement of previous disclosure	An article presents Company A's previously discussed frontier-AI share, Vera CPU, LPX and AI-factory opportunities as fresh upside.	Recognise old disclosures repackaged as news or combined with unsubstantiated opinion; new framing is not new information.
Short-seller claims	A short seller argues Company A's disclosed inventory build shows channel stuffing and delayed hyperscaler orders.	Separate public facts from unsupported causal claims and require new evidence before assigning valuation relevance.
Company updates	Company A repeats its second-quarter outlook but changes the non-GAAP margin bridge via fiscal 2025 add-backs.	Find the incremental detail in recycled guidance and translate an accounting bridge into earnings.
Product timeline shifts	Company B moves WaveLogic 7 customer sampling from 2027 to 2028.	Compare the new window with the prior roadmap and competing DSP timelines, while distinguishing sampling from commercial shipments.
Industry reports	An industry research house reports quantities, prices and discounts for ten Company A Vera CPU customer contracts.	Apply the discounts, aggregate net value to $28.9 billion and compare it with the prior $20 billion standalone-CPU expectation.
Customer contract changes	A customer replaces €2.1 billion of Company D High-NA orders with €1.25 billion of Low-NA systems and €150 million of upgrades.	Net removed and retained content; supplier continuity masks a €700 million reduction.
Read-across from competitors	Competitor A targets volume MI450 rack shipments in the second half of 2026 with major customer commitments.	Translate competitor news into Company A share and estimate implications without confusing market growth with company benefit.
Changes in outlook wording	Company C changes its Segment Result Margin outlook from "around 20%" to "up to 20%".	Retrieve the prior wording: one operator turns an approximate target into a ceiling amid an otherwise unchanged outlook.

3.4 Scoring

We report accuracy separately for newness, expected importance, and direction. We additionally report *atomic accuracy*, the proportion of these individual classifications that are correct, and *all-label accuracy*, which requires all three classifications for an event to be correct. A prediction is counted as correct when it matches either the primary expert label or an accepted secondary label for a genuine boundary case. Missing or malformed outputs remain in the benchmark denominator and are counted as incorrect.

The live articles and historical documents are curated distractors rather than individually expert-labelled negative examples. We therefore report an approximate false-positive proxy: the proportion of parseable distractor decisions that an agent classifies as both new and more important than none. We report this overall and separately for live articles and historical documents.

Accuracy alone is insufficient for assessing a practical news-filtering agent. Financial news often needs to be processed at high volume and speed for the output to remain useful. We therefore also report valid-output counts, token use, cost, and web-search volume. These secondary metrics distinguish agents that are accurate but too slow or expensive for current large-scale use and help identify how much external research is required to reach an answer.

3.5 Agent setup and harness

We treat the complete agent configuration, comprising the model, harness, reasoning configuration, prompt, and tool policy, as the evaluated unit. Agents can use web search to investigate company background and prior public information. One large advantage of using synthetic events is that there is minimal risk of related stories surfacing through web search and revealing the answer.

We use Harbor [13] to run each case in an isolated container and capture the answer, execution logs, and agent trajectory. For every article or document, the agent returns the three labels, a rationale, and supporting evidence in a schema-validated JSON file. We use medium reasoning across all models to balance cost and speed.

4 Results

We evaluate 14 agents on the same 82 realistic cases. We find that no agent can reliably replicate overall expert human judgement: the highest all-label accuracy is 52.4%. Treating each label as a separate classification, we see higher scores but still under 75% across all agents.

Given the importance of false positives to any real-world agent deployment, we also find that accurately classifying the synthetic news does not necessarily lead to a lower false-positive rate. For example, GPT-5.4 and Claude Opus 4.8 perform similarly on the target events, but Claude Opus has approximately five times the false-positive rate on the surrounding live articles (35% versus 7%).

4.1 Model performance

As shown in Table 3, GPT-5.5 is the highest-scoring agent, achieving 71.1% atomic accuracy and 52.4% all-label accuracy. GPT-5.6 Sol follows closely at 69.1% and 51.2%, respectively. Performance then falls away: the next-best agent reaches 63.8% atomic and 46.3% all-label accuracy. Qwen3.6 35B-A3B reaches 43.9% atomic and 22.0% all-label accuracy with 78 valid answers, while Nemotron 3 Super reaches 15.9% and 8.5% with only 28 valid answers. Nemotron's result therefore reflects substantial output and execution unreliability as well as classification errors. Even the strongest agent is fully correct on only around half of the events,

Table 3. Headline performance and operational results on the 82-case realistic benchmark. Agents are ordered by all-label accuracy within access group.

Agent	All labels (%)	Atomic (%)	Valid	Cost ($)	Input tok.	Output tok.	Searches
Closed-weight agents							
GPT-5.5	**52.4**	**71.1**	82/82	76.34	32.66M	645k	945
GPT-5.6 Sol	51.2	69.1	82/82	83.55	62.36M	644k	1,037
GPT-5.6 Terra	46.3	63.8	82/82	25.63	35.33M	433k	563
GPT-5.6 Luna	42.7	63.4	82/82	12.01	40.71M	538k	657
GPT-5.4	40.2	59.3	82/82	41.75	29.83M	771k	2,041
Claude Opus 4.8	39.0	59.8	82/82	206.46	50.29M	1.07M	206
Claude Sonnet 4.6	36.6	63.0	82/82	110.52	54.55M	1.47M	433
GPT-5.4 mini	30.5	54.1	81/82	15.06	32.52M	1.21M	2,115
Open-weight agents							
DeepSeek V4 Pro	**40.2**	57.7	82/82	3.15	42.41M	555k	578
DeepSeek V4 Flash	34.1	**58.1**	81/82	0.97	40.58M	482k	570
MiMo-V2.5	32.9	56.9	81/82	1.54	68.57M	486k	1,120
Qwen3.6 35B-A3B	22.0	43.9	78/82	3.70	37.42M	644k	737
Qwen3.7 Plus	20.7	49.2	77/82	6.52	44.30M	528k	624
Nemotron 3 Super	8.5	15.9	28/82	8.54	95.11M	428k	594

Notes: Atomic accuracy is the proportion of the 246 newness, importance, and direction classifications that are correct; all-label accuracy requires all three labels for a case to be correct. Malformed or missing outputs remain in the 82-case denominator. Cost, token, and search totals include the single attempt represented by each score: valid completed outputs and agent-completed invalid-format outputs scored as incorrect. Failed and superseded attempts are excluded. API-billed agents use reconciled request costs; subscription-access agents use harness-reported usage-equivalent costs. Search-service charges and local compute are excluded.

despite getting more than two-thirds of the individual labels right. Figure 4 shows one example of an all-label failure shared by every agent.

Shared failure: one operator changes the guidance

Previous guidance	New article info	What changed
Segment Result Margin **around 20%**	Segment Result Margin **up to 20%**	An approximate target becomes a ceiling
Prop. Correct	**Majority response**	**Expert assessment**
0/14 (0%)	**11/14: Not new[a], none[b], neutral[c]**	**New[a], high[b], negative[c]**

Common failure mode. Eleven agents matched the repeated company, period, metric and number to the earlier disclosure, then treated the article as recycled. The remaining outputs also miss the negative direction; one identifies the change and assigns an accepted importance label but predicts positive direction.

[a] *Information newness* [b] *Level of importance* [c] *Direction*

Figure 4. A shared all-label failure on Company C's margin guidance. None of the 14 agents matches the complete expert assessment; 11 return the majority not-new, none and neutral response.

Recorded cost is not monotonic with accuracy. DeepSeek V4 Flash comes within 1.7 percentage points of

Claude Opus 4.8 on atomic accuracy (58.1% versus 59.8%) at recorded costs of $0.97 and $206.46, respectively. DeepSeek V4 Pro reaches the same 40.2% all-label accuracy as GPT-5.4 at $3.15 rather than $41.75, although its atomic accuracy is 1.6 points lower. Conversely, Qwen3.7 Plus costs more than MiMo-V2.5 while scoring 7.7 points lower atomically. Qwen3.6 costs only $3.70 but reaches 43.9% atomic accuracy.

Research behaviour also differs substantially without following the accuracy ranking. GPT-5.4 mini issues 2,115 searches, more than twice GPT-5.5's 945, but scores 17.0 points lower atomically. At almost the same accuracy, Claude Opus 4.8 uses only 206 searches compared with GPT-5.4's 2,041. Token use shows the same pattern: Nemotron consumes the most input tokens at 95.11 million across only 59 score-bearing executions, yet produces just 28 valid answers. GPT-5.5 leads the benchmark using 32.66 million input tokens, while Claude Sonnet 4.6 produces the most output tokens at 1.47 million without reaching the leading accuracy tier. The observed totals therefore provide no evidence that longer trajectories or more extensive search alone improve event classification.

4.2 Label-level performance

Table 4. Accuracy by target label and strict all-label accuracy, reported as percentages and split by the expert newness label.

Agent	Newness	Importance	Direction	All: new	All: not new
Closed-weight agents					
GPT-5.5	**80.5**	**62.2**	**70.7**	**50.0**	61.1
GPT-5.6 Sol	76.8	61.0	69.5	43.8	77.8
GPT-5.6 Terra	69.5	58.5	63.4	35.9	**83.3**
GPT-5.6 Luna	75.6	58.5	56.1	34.4	72.2
Claude Sonnet 4.6	76.8	50.0	62.2	32.8	50.0
Claude Opus 4.8	68.3	51.2	59.8	29.7	72.2
GPT-5.4	70.7	51.2	56.1	28.1	**83.3**
GPT-5.4 mini	62.2	47.6	52.4	21.9	61.1
Open-weight agents					
DeepSeek V4 Flash	65.9	51.2	**57.3**	28.1	**55.6**
DeepSeek V4 Pro	**68.3**	**53.7**	51.2	**39.1**	44.4
MiMo-V2.5	**68.3**	50.0	52.4	35.9	22.2
Qwen3.7 Plus	64.6	36.6	46.3	15.6	38.9
Qwen3.6 35B-A3B	47.6	40.2	43.9	12.5	**55.6**
Nemotron 3 Super	18.3	14.6	14.6	1.6	33.3

Notes: Values are percentages. Each individual label uses all 82 cases; invalid outputs score zero. Predictions matching an expert-accepted secondary importance or direction label are correct. The strict split uses 64 gold-new and 18 gold-not-new events.

Newness is the strongest individual label across the evaluated agents, with 749/1,148 decisions correct (65.2%), compared with 620/1,148 for direction (54.0%) and 563/1,148 for expected importance (49.0%). Every agent has lower importance accuracy than newness accuracy (Table 4). The central difficulty is therefore not only recognising that an article contains a new fact, but calibrating how much that fact should change valuation-relevant beliefs.

This difficulty is more apparent under the joint measure. Across all agent-event decisions, all-label accuracy is 262/896 (29.2%) on the 64 genuinely new events, compared with 146/252 (57.9%) on the 18 not-new events. Most not-new events resolve to the simpler joint output of not new, no importance and neutral direction. A new event additionally requires the agent to translate the factual change into an importance and

direction judgement. GPT-5.5 has the highest all-label accuracy on new events at 50.0%, while no other agent exceeds 43.8%. Figure 5 illustrates how failure to get one of the expert labels often highlights a fundamental misunderstanding of the important change.

Direction trap: the headline uses the wrong comparison

Positive headline	Required comparison	Hidden reversal
Q1 FY27 revenue expected to rise 10% year on year	Q1 FY27 $\approx$ €4.028bn; implied Q4 FY26 $\geq$ €4.426bn	$\geq$9% sequential decline, versus prior guidance that Q1 should be better than Q4
Agent newness	**Agent direction**	**Expert assessment**
14/14 identify it as new[a]	**14/14 predict positive[c] direction**	**New[a], high[b], negative[c]**

Common failure mode. The agents followed the article's positive year-on-year framing. They did not change the comparison base to the preceding quarter, where the figures imply a decline of at least 9% and reverse prior guidance.

[a] *Information newness* [b] *Level of importance* [c] *Direction*

Figure 5. The Company C fiscal Q1 2027 case separates novelty detection from directional reasoning. Every agent recognises the new forecast, but the positive year-on-year framing masks a negative sequential revision relative to the prior outlook.

4.3 False-positive trade-off

Approximate false-positive behaviour varies substantially and is not determined by target-event accuracy (Table 5). GPT-5.6 Sol combines near-leading target performance with the lowest overall rate, at 1.0%, compared with 6.4% for GPT-5.5. GPT-5.6 Terra and Luna have almost identical atomic accuracy (63.8% and 63.4%) but false-positive rates of 5.1% and 14.6%. Similarly, Claude Opus 4.8 and GPT-5.4 differ by only one correct atomic classification, but their false-positive rates are 24.9% and 5.7%.

The false positives are concentrated almost entirely in live articles. Across all agents, 1,132/5,410 article decisions (20.9%) are escalated as both new and more important than none, compared with 13/2,164 historical-document decisions (0.6%). Thus, 1,132 of the 1,145 estimated false positives arise from articles. Agents can usually recognise that an old filing or company report is not current news; the harder problem here, and in the real world, is distinguishing genuinely incremental information from recently published derivative coverage, repeated announcements, and immaterial read-across.

Table 5. Approximate false-positive rates on curated noise items, overall and by item type.

Agent	Overall	Live articles	Historical documents
Closed-weight agents			
GPT-5.5	37/574 (6.4%)	37/410 (9.0%)	0/164 (0.0%)
GPT-5.6 Sol	**6/574 (1.0%)**	5/410 (1.2%)	1/164 (0.6%)
GPT-5.6 Terra	29/574 (5.1%)	29/410 (7.1%)	0/164 (0.0%)
GPT-5.6 Luna	84/574 (14.6%)	84/410 (20.5%)	0/164 (0.0%)
Claude Sonnet 4.6	185/574 (32.2%)	184/410 (44.9%)	1/164 (0.6%)
Claude Opus 4.8	143/574 (24.9%)	143/410 (34.9%)	0/164 (0.0%)
GPT-5.4	33/574 (5.7%)	29/410 (7.1%)	4/164 (2.4%)
GPT-5.4 mini	55/567 (9.7%)	55/405 (13.6%)	0/162 (0.0%)
Open-weight agents			
DeepSeek V4 Flash	108/567 (19.0%)	108/405 (26.7%)	0/162 (0.0%)
DeepSeek V4 Pro	99/574 (17.2%)	99/410 (24.1%)	0/164 (0.0%)
MiMo-V2.5	117/567 (20.6%)	113/405 (27.9%)	4/162 (2.5%)
Qwen3.7 Plus	134/539 (24.9%)	134/385 (34.8%)	0/154 (0.0%)
Qwen3.6 35B-A3B	**84/546 (15.4%)**	84/390 (21.5%)	0/156 (0.0%)
Nemotron 3 Super	31/196 (15.8%)	28/140 (20.0%)	3/56 (5.4%)

Notes: A noise decision is counted when an agent labels the item as both new and more important than none. This is an approximate escalation proxy rather than a conventional gold-labelled false-positive rate. Denominators include only parseable decisions from valid answers; repeated pool items are counted each time they occur in a case.

5 Discussion

Our results suggest that current frontier agents still cannot assess equity news flow to the standard of professional equity analysts. The strongest agent, GPT-5.5, matches the expert assessment on all three labels in only 52.4% of cases. GPT-5.6 Sol reaches 51.2% all-label accuracy and Claude Opus 4.8 reaches 39.0%. Their higher atomic scores show that they often identify part of the correct answer, but this is insufficient for a task where novelty, importance and direction jointly determine whether and how an event should be escalated. Even the strongest agents therefore fail to reproduce the complete expert judgement on almost half of the events.

The shared failures indicate that the remaining gap is not simply one of retrieving more information. Agents overlook subtle changes within otherwise repeated disclosures, as in the shift from a margin target of *around 20%* to a ceiling of *up to 20%* in Figure 4. They also accept favourable article framing without selecting the comparison that is financially relevant. In Figure 5, every agent recognises that the forecast is new, but every agent follows its positive year-on-year headline rather than calculating the negative sequential revision implied by the figures and prior guidance. Together with the consistently weaker importance results, these cases suggest that agents often fail to integrate small factual changes, historical expectations and their implied financial impact. These are failures of contextual financial judgement rather than simple document retrieval or sentiment classification.

This limitation matters both for direct news monitoring and for broader financial workflows. In a news-filtering application, missing a small but consequential change creates a false negative, while an incorrect importance or direction judgement can cause analysts to prioritise the wrong event. The same assessment is also an upstream input to updating financial models, revising valuations, assessing acquisitions and conducting wider equity research. Frontier Financial Judgement does not evaluate these longer workflows end to end, so our results do not establish their overall performance. They do, however, identify a consequential subtask on

which the quality of downstream analysis may depend.

For real-world applications, accuracy is not the only factor. Latency and cost are often important barriers to practical deployment at scale. Although recorded benchmark cost and token use are imperfect proxies for actual wall-clock time, they reflect the fact that substantial agent trajectories must be completed within a realistic trading window after an article is published. Practical deployment must therefore trade off accuracy against both cost and timeliness, with Figure 1 showing a clear cost-accuracy frontier. Therefore, even if frontier models approach human expert judgement in the future, there will likely still be a substantial need for research into harness and model optimisation in order to meet real-world latency and cost requirements.

Finally, target accuracy and false-positive behaviour form separate components of real-world usability. GPT-5.6 Sol combines near-leading target accuracy with a 1.0% approximate false-positive rate, whereas Claude Opus 4.8 reaches 24.9%. The comparison between Claude Opus and GPT-5.4 is particularly instructive: they differ by only one correct atomic classification, yet GPT-5.4's approximate false-positive rate is 5.7%. In a low-base-rate news environment, these differences can translate into substantially different volumes of irrelevant material being escalated for review. GPT-5.6 Sol provides the strongest combined result in this experiment, but no single accuracy measure captures the operational trade-off. Practical news agents should therefore be evaluated jointly on judgement accuracy, restraint, output reliability and cost. As the noise items are curated distractors rather than an individually expert-labelled negative set, these rates remain controlled estimates of noise escalation rather than population-wide false-positive rates for financial news.

6 Limitations

The primary limitation of the synthetic-event approach is that it does not reproduce the noise created by competing interpretations of the same information across different articles. Real-world news-flow filtering may therefore be even more challenging than the benchmark suggests. This first version of Frontier Financial Judgement focuses on semiconductor supply-chain companies, and its results may not generalise to other sectors. Finally, some expert judgements are inherently subjective, and other financial professionals may reasonably have assigned different labels.

7 Conclusion

This paper introduces Frontier Financial Judgement, a benchmark for evaluating whether research agents can reproduce professional equity analyst assessments of news newness, importance and direction. Across 14 agents, the strongest result reaches only 52.4% all-label accuracy. Even frontier agents frequently miss subtle changes, accept misleading framing or fail to integrate an event's implied financial impact. Their performance also varies substantially in cost, output reliability and restraint when presented with irrelevant material.

These findings suggest that current agents can support equity-news monitoring, but cannot yet replace professional judgement in consequential financial workflows. Practical deployment requires evaluation of complete decisions rather than isolated labels, alongside false-positive behaviour, reliability and cost. Frontier Financial Judgement provides a reproducible foundation for measuring progress on these dimensions, while future work should extend its coverage across sectors and more closely reproduce the competing interpretations found in real-world news flow.

References

[1] Paul Asquith, Michael B. Mikhail, and Andrea S. Au. Information content of equity analyst reports. *Journal of Financial Economics*, 75(2):245–282, February 2005. doi: 10.1016/j.jfineco.2004.01.002. URL https://doi.org/10.1016/j.jfineco.2004.01.002.

[2] 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. doi: 10.48550/arXiv.2508.00828. URL https://arxiv.org/abs/2508.00828.

[3] Elizabeth Blankespoor, Ed deHaan, and Iván Marinovic. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics*, 70(2–3):101344, November 2020. doi: 10.1016/j.jacceco.2020.101344. URL https://doi.org/10.1016/j.jacceco.2020.101344.

[4] Keven Bluteau, Frank Coggins, and Gilles Boevi Koumou. Enhancing ESG news annotation: Leveraging GPT for the analysis of ESG news and events. *SSRN Electronic Journal*, February 2025. doi: 10.2139/ssrn.5128896. URL https://ssrn.com/abstract=5128896.

[5] Jacob Boudoukh, Ronen Feldman, Shimon Kogan, and Matthew Richardson. Information, trading, and volatility: Evidence from firm-specific news. *The Review of Financial Studies*, 32(3):992–1033, March 2019. doi: 10.1093/rfs/hhy083. URL https://doi.org/10.1093/rfs/hhy083.

[6] Lawrence D. Brown, Andrew C. Call, Michael B. Clement, and Nathan Y. Sharp. The activities of buy-side analysts and the determinants of their stock recommendations. *Journal of Accounting and Economics*, 62(1): 139–156, August 2016. doi: 10.1016/j.jacceco.2016.06.002. URL https://doi.org/10.1016/j.jacceco.2016.06.002.

[7] Tianyu Chen, Yiming Zhang, Guoxin Yu, Dapeng Zhang, Li Zeng, Qing He, and Xiang Ao. EFSA: Towards event-level financial sentiment analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7455–7467, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.402. URL https://aclanthology.org/2024.acl-long.402/.

[8] Chanyeol Choi, Jihoon Kwon, Alejandro Lopez-Lira, Chaewoon Kim, Minjae Kim, Juneha Hwang, Jaeseon Ha, Hojun Choi, Suyeol Yun, Yongjin Kim, and Yongjae Lee. FinAgentBench: A benchmark dataset for agentic retrieval in financial question answering. *arXiv preprint arXiv:2508.14052*, 2025. doi: 10.48550/arXiv.2508.14052. URL https://arxiv.org/abs/2508.14052.

[9] Efthimios G. Demirakos, Norman C. Strong, and Martin Walker. What valuation models do analysts use? *Accounting Horizons*, 18(4):221–240, December 2004. doi: 10.2308/acch.2004.18.4.221. URL https://doi.org/10.2308/acch.2004.18.4.221.

[10] Exa. Exa Search API, 2026. URL https://exa.ai/docs/reference/search. API documentation, accessed 19 July 2026.

[11] Anastassia Fedyk and James Hodson. When can the market identify old news? *Journal of Financial Economics*, 149(1):92–113, July 2023. doi: 10.1016/j.jfineco.2023.04.008. URL https://doi.org/10.1016/j.jfineco.2023.04.008.

[12] Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *The Journal of Financial Data Science*, 6(1):25–42, 2024. doi: 10.3905/jfds.2023.1.143. URL https://doi.org/10.3905/jfds.2023.1.143.

[13] Harbor Framework Team. Harbor: A framework for evaluating and optimizing agents and models in container environments, January 2026. URL https://github.com/harbor-framework/harbor. Software, version 0.15.0.

[14] David Hirshleifer, Sonya Seongyeon Lim, and Siew Hong Teoh. Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance*, 64(5):2289–2325, October 2009. doi: 10.1111/j.1540-6261.2009.01501.x. URL https://doi.org/10.1111/j.1540-6261.2009.01501.x.

[15] Liang Hu, Jianpeng Jiao, Jiashuo Liu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, Fei Hu, Yali Liao, Zaiyuan Wang, Chenghao Yang, Qianyu Yang, Mingren Yin, Zhiyuan Zeng, Ge Zhang, Xinyi Zhang, Xiying Zhao, Zhenwei Zhu, Hongseok Namkoong, Wenhao Huang, and Yuwen Tang.

FinSearchComp: Towards a realistic, expert-level evaluation of financial search and reasoning. *arXiv preprint arXiv:2509.13160*, 2025. doi: 10.48550/arXiv.2509.13160. URL https://arxiv.org/abs/2509.13160.

[16] 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. doi: 10.48550/arXiv.2311.11944. URL https://arxiv.org/abs/2311.11944.

[17] Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308/.

[18] Yoontae Jeon, Thomas H. McCurdy, and Xiaofei Zhao. News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies. *Journal of Financial Economics*, 145(2):1–17, August 2022. doi: 10.1016/j.jfineco.2021.08.002. URL https://doi.org/10.1016/j.jfineco.2021.08.002.

[19] Yuechen Jiang, Zhiwei Liu, Yupeng Cao, Yueru He, Ziyang Xu, Chen Xu, Zhiyang Deng, Prayag Tiwari, Xi Chen, Alejandro Lopez-Lira, Jimin Huang, Junichi Tsujii, and Sophia Ananiadou. All that glisters is not gold: A benchmark for reference-free counterfactual financial misinformation detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10737–10776, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.492. URL https://aclanthology.org/2026.acl-long.492/.

[20] S. P. Kothari, Eric So, and Rodrigo Verdi. Analysts' forecasts and asset pricing: A survey. *Annual Review of Financial Economics*, 8:197–219, October 2016. doi: 10.1146/annurev-financial-121415-032930. URL https://doi.org/10.1146/annurev-financial-121415-032930.

[21] Baptiste Lefort, Eric Benhamou, Beatrice Guez, Jean-Jacques Ohana, Ethan Setrouk, and Alban Etienne. FinMarBa: A market-informed dataset for financial sentiment classification. *arXiv preprint arXiv:2507.22932*, 2025. doi: 10.48550/arXiv.2507.22932. URL https://arxiv.org/abs/2507.22932.

[22] Yachuan Liu, Xiaochun Wei, Lin Shi, Xinnuo Li, Bohan Zhang, Paramveer Dhillon, and Qiaozhu Mei. ExAnte: A benchmark for ex-ante inference in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1551–1571, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.72. URL https://aclanthology.org/2026.eacl-long.72/.

[23] Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. FMDLlama: Financial misinformation detection based on large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1153–1157. Association for Computing Machinery, 2025. doi: 10.1145/3701716.3715599. URL https://doi.org/10.1145/3701716.3715599.

[24] Zhiwei Liu, Yueru He, Qing Ou, Tianlei Zhu, Xiaorui Guo, Xueqing Peng, and Sophia Ananiadou. AuditFraudBench: Benchmarking audit judgment in detecting fraudulent misstatements. *arXiv preprint arXiv:2606.08345*, 2026. doi: 10.48550/arXiv.2606.08345. URL https://arxiv.org/abs/2606.08345.

[25] Michael B. Mikhail, Beverly R. Walther, and Richard H. Willis. The effect of experience on security analyst underreaction. *Journal of Accounting and Economics*, 35(1):101–116, April 2003. doi: 10.1016/S0165-4101(02)00099-X. URL https://doi.org/10.1016/S0165-4101(02)00099-X.

[26] Malthe Have Musaeus, Faisal Sayed, Mersad Abbasi, Daanish Khazi, and Karina Nguyen. DiligenceBench: An equity-research agent evaluation, July 2026. URL https://www.paperinstruments.com/blog/diligence-bench. Paper Instruments and Thoughtful Lab.

[27] James M. Patell and Mark A. Wolfson. The intraday speed of adjustment of stock prices to earnings and dividend announcements. *Journal of Financial Economics*, 13(2):223–252, June 1984. doi: 10.1016/0304-405X(84)90024-2. URL https://doi.org/10.1016/0304-405X(84)90024-2.

[28] 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. *arXiv preprint arXiv:2507.16248*, 2025. doi: 10.48550/arXiv.2507.16248. URL https://arxiv.org/abs/2507.16248.

[29] Paul C. Tetlock. All the news that's fit to reprint: Do investors react to stale information? *The Review of Financial Studies*, 24(5):1481–1512, May 2011. doi: 10.1093/rfs/hhq141. URL https://doi.org/10.1093/rfs/hhq141.

[30] Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, June 2008. doi: 10.1111/j.1540-6261.2008.01362.x. URL https://doi.org/10.1111/j.1540-6261.2008.01362.x.

[31] 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. *arXiv preprint arXiv:2606.03829*, 2026. doi: 10.48550/arXiv.2606.03829. URL https://arxiv.org/abs/2606.03829.

[32] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e4a46394ba5378b3f9a186a5b4c650d1-Abstract-Conference.html. ICLR 2025 Spotlight.

[33] Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. FinDVer: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.818. URL https://aclanthology.org/2024.emnlp-main.818/.

[34] Zijie Zhao and Roy E. Welsch. Point-in-time financial RAG with frozen LLMs and market-feedback adaptive retrieval. *arXiv preprint arXiv:2605.31201*, 2026. doi: 10.48550/arXiv.2605.31201. URL https://arxiv.org/abs/2605.31201.

[35] Zhihan Zhou, Liqian Ma, and Han Liu. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.186. URL https://aclanthology.org/2021.findings-acl.186/.

[36] Fengbin Zhu, Junfeng Li, Liangming Pan, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. Towards temporal-aware multi-modal retrieval augmented generation in finance. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6289–6297. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3755723. URL https://doi.org/10.1145/3746027.3755723.

A Appendix

A.1 Prompts

Figure 6 reproduces the complete instruction used for the realistic benchmark. The agent receives this instruction alongside a task manifest containing the company, ticker, assessment cutoff and paths to the eight supplied items.

Realistic news-flow assessment prompt

I am an analyst and have been given the following articles and documents to review so I can update the equity sales team on any new developments that could be valuation relevant for the company. They are obviously busy and know the stock well so you should only flag genuinely new valuation relevant information that could actually influence company valuation.

The company and ticker to assess the articles / documents are given in `/app/task.json` as `task.company` and `task.ticker`.

Please provide me with three classifications for each article / document so I can only mention / pass onto the team genuinely new valuation relevant info.

New information classification

- `information_new = false` when the important content is equivalent to information already public before `as_of_utc`, independently of the supplied item itself, or the item contains no substantive factual development.
- `information_new = true` only when the item contains a factual or substantive qualitative development that was not otherwise public before `as_of_utc`.

Expected importance classification

- `none`: no novel factual delta, or the novel information is not expected to change valuation-relevant beliefs substantively.
- `low`: new information with a plausible but limited or uncertain connection to valuation-relevant expectations.
- `medium`: new information likely to matter to valuation-relevant expectations, but not clearly major on its own.
- `high`: new information highly likely to change valuation-relevant expectations meaningfully.

Direction classification

- `positive`: the novel valuation-relevant information is likely to support or raise expectations around valuation drivers.
- `negative`: it is likely to weigh on or lower those expectations.
- `neutral`: there is no meaningful novel valuation-relevant delta, or the new information has no expected valuation impact.
- `unclear`: there is novel potentially valuation-relevant information, but direction cannot be determined or has materially offsetting implications.

Notes

Read `/app/task.json` first and review `task.items[*].path`.

Assess every item as of `as_of_utc`. Do not use information or stock price movements after `as_of_utc`.

Some supplied material may come from third-party or internal feeds and may not appear verbatim on the open web. Use web search for background and prior public information, but do not treat the absence of an exact online match as evidence that an item is false, already known, or unimportant.

Use the supplied items and allowed search tools to ground each assessment. Cite key evidence in the evidence field.

Evidence from a supplied article or document uses:

- `type = feed_item`
- `itemId` = the relevant supplied `itemId`

Evidence from search results uses:

- `type = web_search_result`
- include `title`, `url`, and `published_at_utc`; do not put native-tool internal citation handles such as `turn0search4` in `result_id`

Before writing the final answer, verify that `answer.assessments[*].itemId` exactly matches the `itemIds` in `/app/task.json`, with no missing, duplicate, or extra `itemIds`.

Write exactly one JSON object matching `/app/answer.schema.json` to `/logs/artifacts/answer.json`.

Figure 6. The complete realistic-setting instruction presented to each agent. Markdown headings and lists are reformatted typographically, but the wording is unchanged.

A.2 Example primer paragraphs

Synthetic articles combine event-specific evidence with generic company background. The background paragraphs are selected from a reviewed company-specific primer library and do not determine the expert labels. Figure 7 shows three examples from the NVIDIA library.

NVIDIA's shift from graphics chips to accelerated computing platforms

NVIDIA began as a graphics processor company, but investors now tend to analyze it as a platform supplier for accelerated computing. Its business spans data-center accelerators, networking, systems, software libraries, gaming GPUs, professional visualization, automotive computing, and robotics. The company's central position in AI infrastructure comes from combining chips with interconnects, systems design, developer software, and an ecosystem of hardware and cloud partners. That platform breadth means a single product or demand update can affect expectations for multiple parts of the business, not just one chip SKU.

AI factories and data-center infrastructure constraints

Large AI clusters require more than accelerator supply. Operators need power availability, cooling capacity, high-bandwidth networking, physical space, systems integration, and software to schedule workloads across many chips. NVIDIA describes this buildout as AI factory infrastructure because the facilities convert electricity and data into model training and inference output. For investors, that makes deployment pace sensitive to construction schedules, utility connections, rack readiness, and the ability of customers to absorb complete systems.

Gross-margin sensitivity to product mix and supply chain

NVIDIA's margins can be affected by product mix, system content, memory costs, packaging capacity, and the balance between chips, full systems, and software. Data-center products have often supported strong profitability, but more complex rack-scale systems can also bring different cost structures. Investors therefore pay close attention to whether new information points to changes in pricing power, supply costs, or the mix of products being shipped.

Figure 7. Example generic background primers used when constructing NVIDIA synthetic articles.

A.3 Agent setup and harness

All agents run within Harbor using the same frozen cases, answer schema, artifact capture and verification procedure. The evaluated model-specific configurations are:

- **GPT-5.4:** Harbor with the single-agent Codex adapter and Codex CLI 0.142.3; native Codex web search; OpenAI through Codex subscription access.
- **GPT-5.4 mini:** Harbor with the single-agent Codex adapter and Codex CLI 0.142.3; native Codex web search; OpenAI through Codex subscription access.
- **GPT-5.5:** Harbor with the single-agent Codex adapter and Codex CLI 0.142.3; native Codex web search; OpenAI through Codex subscription access.
- **GPT-5.6 Sol:** Harbor with the single-agent Codex adapter and Codex CLI 0.144.1; native Codex web search; OpenAI through Codex subscription access.
- **GPT-5.6 Terra:** Harbor with the single-agent Codex adapter and Codex CLI 0.144.1; native Codex web search; OpenAI through Codex subscription access.
- **GPT-5.6 Luna:** Harbor with the single-agent Codex adapter and Codex CLI 0.144.1; native Codex web search; OpenAI through Codex subscription access.
- **Claude Opus 4.8:** Harbor with the single-agent Claude Code adapter and Claude Code 2.1.206; native Claude web search; Anthropic through Claude subscription access.
- **Claude Sonnet 4.6:** Harbor with the single-agent Claude Code adapter and Claude Code 2.1.206; native Claude web search; Anthropic through Claude subscription access.
- **DeepSeek V4 Flash:** Harbor with the single-agent OpenCode adapter and OpenCode 1.18.2; the shared Exa web-search tool; OpenRouter routing pinned to DeepSeek with provider fallback disabled.

- **DeepSeek V4 Pro:** Harbor with the single-agent OpenCode adapter and OpenCode 1.18.2; the shared Exa web-search tool; OpenRouter routing pinned to DeepSeek with provider fallback disabled.
- **MiMo-V2.5:** Harbor with the single-agent OpenCode adapter and OpenCode 1.18.2; the shared Exa web-search tool; OpenRouter routing pinned to Xiaomi's FP8 endpoint with provider fallback disabled.
- **Qwen3.7 Plus:** Harbor with the single-agent OpenCode adapter and OpenCode 1.18.2; the shared Exa web-search tool; OpenRouter routing pinned to Alibaba with provider fallback disabled.
- **Qwen3.6 35B-A3B:** Harbor with the authenticated-Exa OpenCode adapter and OpenCode 1.18.2; the shared Exa web-search tool; OpenRouter routing pinned to Parasail's FP8 endpoint with provider fallback disabled.
- **Nemotron 3 Super:** Harbor with the authenticated-Exa OpenCode adapter and OpenCode 1.18.2; the shared Exa web-search tool; OpenRouter routing pinned to DeepInfra's BF16 endpoint with provider fallback disabled.

A.4 Live news articles

The live article pool is collected immediately before benchmark cases are frozen. For each company, the Exa Search API [10] receives the configured query `Latest news on {company} {ticker}`. The request uses Exa's automatic search mode to retrieve up to 50 results published during the preceding three days, including the full extracted article text.

Candidates are normalised, deduplicated and ordered newest first. Each is then reviewed for company relevance. Direct company news and customer, supplier, competitor, regulatory, sector or market news with a credible company read-through are retained. Incidental mentions, generic ticker lists, articles too thin to classify, malformed pages and stories about unrelated entities without a plausible read-through are rejected. Review stops once the candidate batch has yielded 25 relevant articles or all candidates have been reviewed. The final 25-article pool is fixed before case construction, after which five articles are selected deterministically for each realistic benchmark case.