arXiv:2611.10926v1 [cs.CL] 7 Nov 2026

DeepFinance: Can AI Diligence a $100MM Deal?

Drew Hahn* **Tyler Shibata*** **Jackson Grove*** **Rahul Sonwalkar***

Julius Research, San Francisco

November 5, 2026

Abstract

Existing evaluations of AI models on high-finance work focus on retrieving answers from financial documents and producing numerically correct outputs. Real private equity (PE) and investment banking (IB) work also depends on choosing between conflicting sources, making sound assumptions, and keeping every step auditable. We introduce DeepFinance, an ultra long-horizon benchmark of 21 end-to-end PE and IB tasks that take a human professional an average of 27.1 hours. Each task pairs a request with a data room averaging 19 files and is graded against an average of 84 expert-written rubric criteria covering output, workflow, and judgment. We evaluate 22 model configurations, each in its native agent harness. The best configuration, GPT-6 Astra with max reasoning, satisfies 57.6% of rubric criteria, and the top four configurations are not statistically distinguishable. Frontier models often calculate correctly from the wrong inputs and find decision-relevant evidence without acting on it.

1 Introduction

DeepFinance is an ultra long-horizon benchmark for measuring models on real-world private equity (PE) and investment banking (IB) workflows that would take a human professional between 20 and 30 hours to complete. DeepFinance was composed and reviewed by PE and IB professionals from firms such as J.P. Morgan, UBS, and Evercore, and institutions such as Harvard Business School, Wharton, and Columbia Business School. The benchmark has the following characteristics:

- Task count: 21
- Average human time: 27.1 hours per task
- Average rubric criteria: 84 criteria per task
- Average environment size: 19 files per data room
- Average prompt length: 240 words, approximately 29% shorter than GDPval [1] prompts

We evaluate 22 model configurations from 12 labs. The highest mean rubric score is 57.6%, and the four leading configurations are not statistically distinguishable (Section 5). The leaderboard and an interactive sample task are available at https://www.juliusresearch.ai/blog/deep-finance.

* *All contributors contributed equally.* Correspondence to: research@julius.ai.

2 Motivation

There is a fundamental gap in how AI models are evaluated on real-world high-finance work. We believe this gap stems from a deeply flawed understanding of how analysts in high finance perform real-world tasks.

Existing evaluations measure models in two ways: their ability to retrieve answers from a set of financial documents, and their ability to produce numerically correct final outputs. Real-world PE and IB workflows are meaningfully more complicated.

In software benchmarks such as SWE-bench [2] and DeepSWE [3], models are graded on whether their output passes tests. If the tests pass and the fix works, the path (or trajectory) the agent took matters less. Finance has no similar test suite. A high-finance valuation is like a pull request that has to pass a line-by-line code review, where every number cites its source and every assumption needs to hold up in front of the managing director, the investment committee (IC), and a client.

In PE and IB, two analysts could produce identical numbers for the same task while one output rests on completely incorrect assumptions. Change a single number upstream, and the entire workflow no longer makes sense. The trajectory of an ultra long-horizon high-finance task is full of assumptions and judgment calls. Without being able to audit the analyst's steps and judgment, a professional in high finance cannot blindly trust the work presented to them.

3 Related Work

Existing high-finance benchmarks have two main limitations in how they evaluate models: (1) they overemphasize question answering over financial documents, and (2) their grading is limited to the numerical correctness of outputs.

Existing benchmarks also often fail to consider the working environment that high-finance professionals face daily. A set of financial documents and revisions in varied formats, long email chains, sparse meeting notes, and other data-room objects almost invariably contains conflicting information that must be reconciled. A model's ability to judge and choose between overlapping sources is rarely measured.

- **FinanceBench** [4]: questions are simple and too easy, and are not realistic or representative of the deal work analysts are hired for.
- **BizBench** [5]: checks one final number against an answer key; DeepFinance grades end-to-end workflows against up to 140 criteria.
- **FinanceQA** [6]: supplies the relevant 10-K sections with each question, leaving no room to evaluate the model's trajectory.
- **BigFinanceBench** [7]: uses solid trajectory reasoning but limits the scope of its questions and stops well short of end-to-end workflows.

Benchmarks such as OfficeQA [8] track a company taxonomy well, but rarely go beyond grading numerical correctness. For ultra long-horizon high-finance tasks, the trajectory is crucial for understanding and auditing the method and assumptions behind the work. ELO-based preference ranking, as used by GDPval [1], compares only final outputs, misses the nuance of complex long-horizon workflows, and does not grade the atomic trajectory.

Some benchmarks, such as Halluminate's Westworld Finance Diligence Bench [9], have attempted to measure semi-long-horizon work. They study the agent's trajectory but leave it out of

the score: every point on their leaderboard comes from checks on the final output. This may be an effective way to check final answers, but it is not how finance work is judged in the real world. Correct numbers without a source or auditability will be scrutinized by an investment committee. DeepFinance grades whether every number is sourced and every judgment call is auditable and traceable.

DeepFinance was built on the premise that the path of a task carries its assumptions, and that grading must therefore account for the trajectory. Auditability is paramount to internal and external stakeholders such as lenders, clients, and brokers. Understanding how a model (or an analyst) arrived at an assumption is key to understanding what could go right, wrong, or sideways in a multi-million-dollar deal. Table 1 compares rubric granularity across benchmarks.

Table 1: Average number of rubric criteria per task.

Benchmark	Criteria per task
BigFinanceBench	16.9
PRBench Finance	17.1
BlueFin	35.1
DeepFinance	**84**

4 Methodology

4.1 Task Construction

DeepFinance covers four primary pillars of high-finance work:

- **Valuation:** estimate what a business is worth and justify the assumptions.
- **Due diligence:** investigate the records, identify important risks, and explain their effect on the deal.
- **Underwriting:** assess the deal's risks and economics, then determine acceptable financing or investment terms.
- **Financial modeling:** build a model whose assumptions, calculations, and outputs remain consistent.

Each ultra long-horizon task pairs a request with a data room: the collection of records needed to complete the assignment. Tasks use real data that has been scrubbed and anonymized. The request states the objective, constraints, and expected deliverables; from there, the agent must work out how to complete the analysis. Problem statements average 240 words, which we find realistic for an end-to-end workflow. Other evaluations use much shorter prompts, but cover only simple arithmetic or quantitative tasks.

Each data room contains on average 19 files and comprises the same material analysts work with day to day. File types include spreadsheets (.xlsx), Word documents (.docx), emails (.eml), PDFs, and raw CSV extracts. Contents include emails, chats, and deal-team notes; models and projections; CIMs, SPAs, term sheets, commitment letters, and transaction data; diligence reports; IC minutes; and market data and comparables.

Each task is representative of end-to-end work at top PE and IB firms such as Evercore, J.P. Morgan, and UBS. The tasks simulate analyst work on real deals, not simple question answering over a set of documents or the search for one-off answers. There is ambiguity, and there are conflicting sources.

4.2 Sample Task

The prompt for an investment due diligence task is reproduced below.

> The IC is meeting this afternoon on Corliss.
>
> Declan came back last night. He still wants $50.50 at completion and another $3.00 paid later if both Quarrie parts go into prod. What would we have to pay at completion and what would it cost us? What price would you put to IC? Show the split at that price and how you get to your numbers.
>
> If IC agrees, signing is probably happening end of Jan. Thx, G

The task comes with a data room of 18 files: a close ledger, supporting registers, a management pack, executed instruments, a pricing framework, a financial reporting policy, debt schedules for two peer companies, the share rights and option plan, a securities register, the sale mandate, an IC price check template, an earlier working, four emails, and an investment team chat export. To answer, the agent must reconcile FY2025 revenue and EBITDA, decide which customer programs count toward sustainable EBITDA under the pricing framework, value the company against its peers, and run the shareholder waterfall for both Declan's structure and an all-cash offer. The reference solution recommends offering 50.803 (USD millions) in cash with no Quarrie payment, with completion nominated for April 1, 2026. The task's rubric has 136 criteria. Appendix A lists the data room files, the reference solution's key figures, and a sample of rubric criteria.

4.3 Scoring

Graders evaluate the finished deliverable together with evidence of the trajectory taken. Sources, tool results, calculations, and revisions can show whether the agent completed the required work. Agents are graded against rubrics whose criteria are labeled with one of three categories:

- **Output:** Are the calculations, conclusions, and requested deliverables correct?
- **Workflow:** Did the agent find the relevant evidence and use it in the analysis?
- **Judgment:** Does the evidence support the agent's assumptions and decisions?

Each rubric criterion was written and verified by experts from top PE and IB firms. By tracking the steps taken with respect to these three categories, reviewers can audit both workflow and judgment, irrespective of the final output. For ultra long-horizon tasks that can take an analyst 20 to 30 or more hours, the trajectory is invaluable.

DeepFinance tasks test whether an agent can carry decisions through a long assignment and follow best practices along the way. A source chosen early may affect a forecast, the amount a company can borrow, its valuation, and the final recommendation. When an input changes, every calculation that depends on it must change too. An agent should not be penalized only for upstream decisions that go wrong; the quality of the steps and judgments made along the way should also

be evaluated, irrespective of output. This is how DeepFinance captures the trajectory, and graders give partial credit when a model takes the right steps.

The rubric score is the percentage of rubric criteria satisfied, averaged over attempts and then over tasks. Attempts that errored were re-run, so every configuration is scored on 10 successful attempts per task.

4.4 Quality Assurance

Every task was reviewed by multiple layers of agentic verifiers and by human expert peers with backgrounds at top firms such as Evercore and institutions such as Harvard Business School and Wharton. Expert review and verification established four conditions:

1. The task is clear, and the source files contain enough information to complete it.
2. The reference solution is supported by the evidence.
3. The rubric checks the requested work and accepts valid solutions.
4. A failed criterion reflects the agent's work, rather than a faulty reference, grader, or environment issue.

Reviewers traced disputed grades and expert commentary back to the source files and agent traces. This distinguishes model mistakes from evaluation mistakes, so that errors in the evaluation process do not influence conclusions about model capability.

4.5 Harness

Unlike many benchmarks, DeepFinance evaluates each model in the harness it is most likely to be used in (Table 2). Each harness has nearly every native tool enabled, including web search. The one exception is the ask-user tool, which is disabled so that every run can finish on its own. Many evaluations cut off internet access so that agents cannot look up answer keys; DeepFinance does not need to. Every run has its own sandbox, and the answer key and rubric sit in separate, isolated grading environments.

The goal is to test agents in environments representative of those PE and IB professionals use day to day. This yields a more realistic evaluation than a single universal harness that imposes arbitrary or artificial constraints on the models.

Table 2: Agent harness used for each model family. All harnesses run on Node.js v22.23.4.

Models	Harness	Version
Claude	claude-code	2.1.318
GPT	codex	0.166.0
Gemini, Muse, DeepSeek, GLM, Nemotron, Inkling	opencode-ai	1.19.3
Kimi	kimi-code	2.4.1
Qwen	qwen-code	0.26.6
Grok	x.ai	1.0.70
Mistral	mistral-vibe	2.24.3

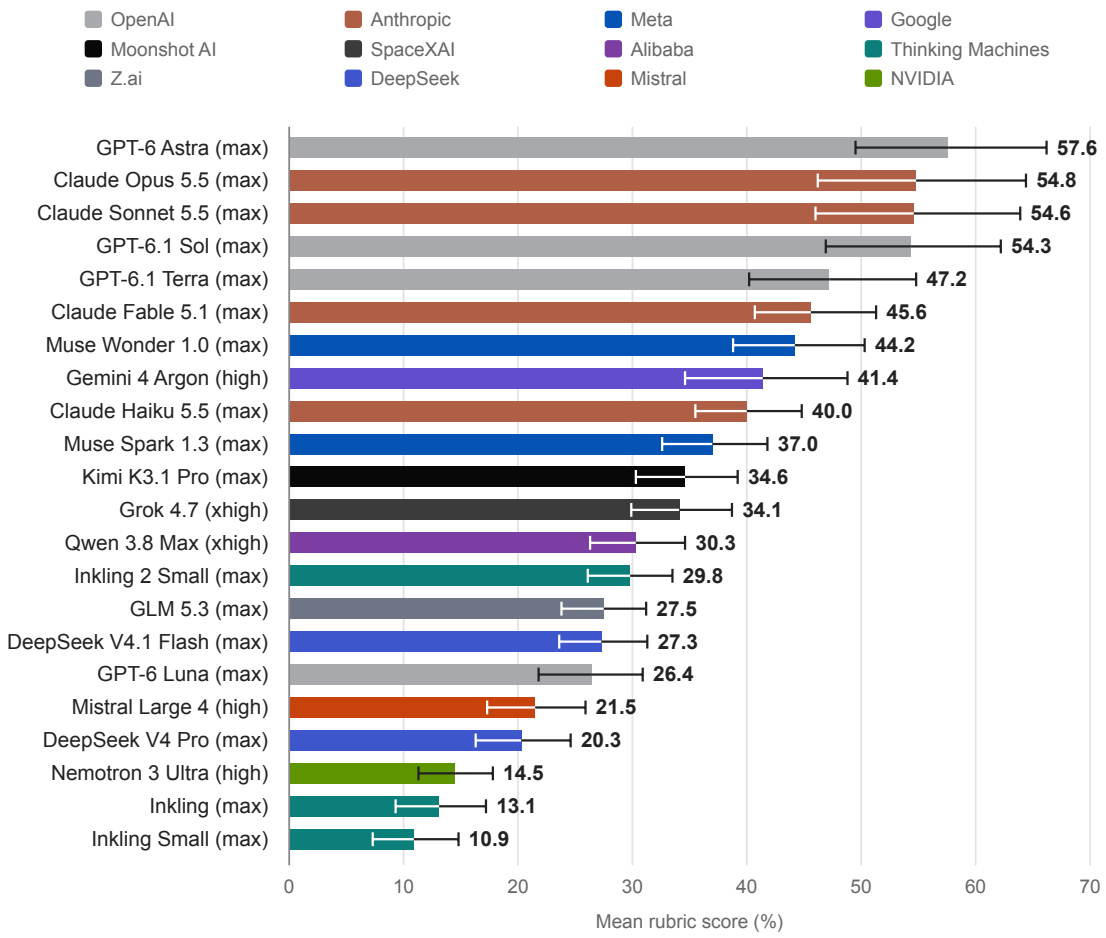


Figure 1: Mean rubric score by model configuration on DeepFinance (21 tasks, 10 attempts per task), colored by lab. Rubric score is the percentage of rubric criteria satisfied. Error bars denote 95% bootstrap confidence intervals (100,000 resamples over tasks and attempts).

5 Results

Figure 1 shows the mean rubric score of each configuration, and Table 3 lists every score with its confidence interval, cost, and token usage.

5.1 Quantitative Analysis

The gap is widening on the frontier. Anthropic and OpenAI models take the six highest scores. The top four configurations, GPT-6 Astra, Claude Opus 5.5, Claude Sonnet 5.5, and GPT-6.1 Sol (57.6, 54.8, 54.6, and 54.3), are not statistically distinguishable (paired bootstrap, 95%). The best model from another provider, Muse Wonder 1.0, scores 44.2, and the bottom group (Nemotron 3 Ultra, Inkling, and Inkling Small, at 10.9 to 14.5) sits roughly 45 points below the top.

GPT-6.1 class models dominate the Pareto frontier. GPT-6.1 Sol with max reasoning averaged $1.17 per completed task with a score of 54.3%. The cheapest model scoring above 30%

Table 3: Results by model configuration, best first. Cost is mean API cost per completed task; tokens are mean tokens per task.

#	Model (reasoning)	Lab	Score (%)	95% CI	Cost (USD)	Tokens
1	GPT-6 Astra (max)	OpenAI	57.6	49.5–66.2	6.73	2.41M
2	Claude Opus 5.5 (max)	Anthropic	54.8	46.2–64.4	8.63	10.49M
3	Claude Sonnet 5.5 (max)	Anthropic	54.6	46.0–63.9	6.94	22.55M
4	GPT-6.1 Sol (max)	OpenAI	54.3	46.9–62.2	1.17	2.52M
5	GPT-6.1 Terra (max)	OpenAI	47.2	40.2–54.8	0.69	4.21M
6	Claude Fable 5.1 (max)	Anthropic	45.6	40.7–51.3	12.02	3.15M
7	Muse Wonder 1.0 (max)	Meta	44.2	38.8–50.3	9.41	8.18M
8	Gemini 4 Argon (high)	Google	41.4	34.6–48.8	2.84	6.67M
9	Claude Haiku 5.5 (max)	Anthropic	40.0	35.5–44.8	1.73	16.47M
10	Muse Spark 1.3 (max)	Meta	37.0	32.6–41.8	4.87	10.87M
11	Kimi K3.1 Pro (max)	Moonshot AI	34.6	30.3–39.2	2.67	6.08M
12	Grok 4.7 (xhigh)	SpaceXAI	34.1	29.9–38.7	4.71	5.90M
13	Qwen 3.8 Max (xhigh)	Alibaba	30.3	26.3–34.6	4.99	12.62M
14	Inkling 2 Small (max)	Thinking Machines	29.8	26.1–33.5	0.50	6.38M
15	GLM 5.3 (max)	Z.ai	27.5	23.8–31.2	1.11	5.61M
16	DeepSeek V4.1 Flash (max)	DeepSeek	27.3	23.6–31.3	0.13	3.93M
17	GPT-6 Luna (max)	OpenAI	26.4	21.8–30.9	0.12	3.63M
18	Mistral Large 4 (high)	Mistral	21.5	17.3–25.9	3.69	5.04M
19	DeepSeek V4 Pro (max)	DeepSeek	20.3	16.3–24.6	1.49	2.61M
20	Nemotron 3 Ultra (high)	NVIDIA	14.5	11.3–17.8	0.64	2.83M
21	Inkling (max)	Thinking Machines	13.1	9.3–17.2	0.64	2.40M
22	Inkling Small (max)	Thinking Machines	10.9	7.3–14.8	0.15	1.09M

was GPT-6.1 Terra, at $0.69 per completed task with a score of 47.2%. GPT-6 models make up 4 of the 6 configurations on the cost Pareto frontier across the 22 tested (Figures 2 and 3).

5.2 Qualitative Analysis

Fable for finance: not worth the cost. Although Claude Fable 5.1 is regarded for its intelligence, this did not carry over to high finance: it scored below GPT-6 Astra, Claude Opus 5.5, Claude Sonnet 5.5, GPT-6.1 Sol, and GPT-6.1 Terra. It was also the most expensive model tested, while scoring under 50% on end-to-end tasks.

Even frontier models are weak at turning evidence into sound judgment. In most DeepFinance tasks, agents find the relevant information without recognizing its implications. They also have difficulty challenging a supplied assumption or evaluating better opportunities unless explicitly told to, even where a human analyst would do so instinctively. When a CFO says one thing and an analyst says another, agents make little effort to reconcile the two perspectives.

Upstream mistakes cause cascading downstream failures. Incorrect earnings, source selection, or completion timing can affect debt capacity, fees, cash flows, valuations, and investor returns. Atomic output checks reveal those consequences, but the number of failed checks should not be interpreted as the number of independent reasoning failures.

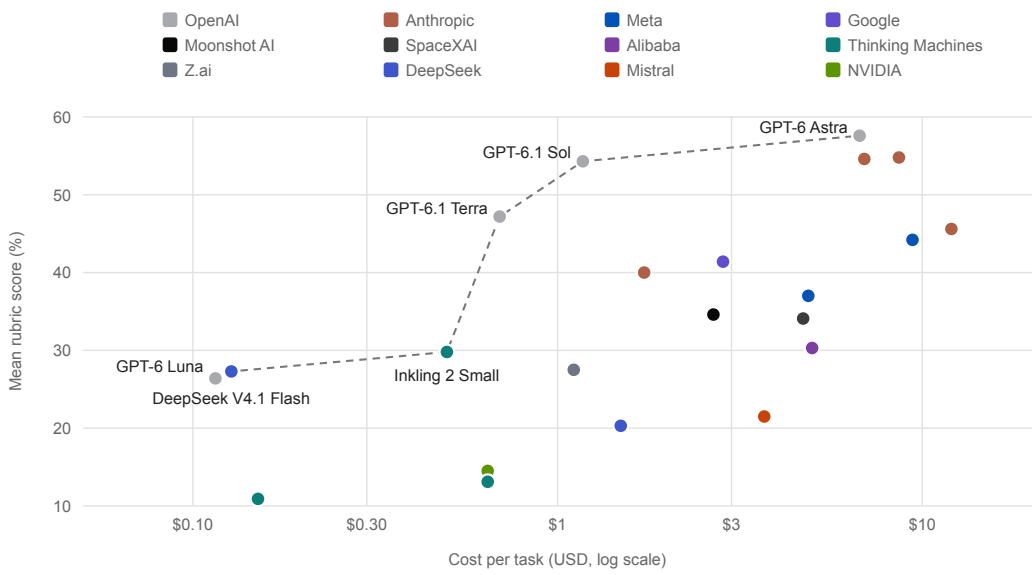


Figure 2: Mean rubric score against mean API cost per task. The dashed line marks the Pareto frontier: configurations not dominated on both score and cost. Frontier configurations are labeled.

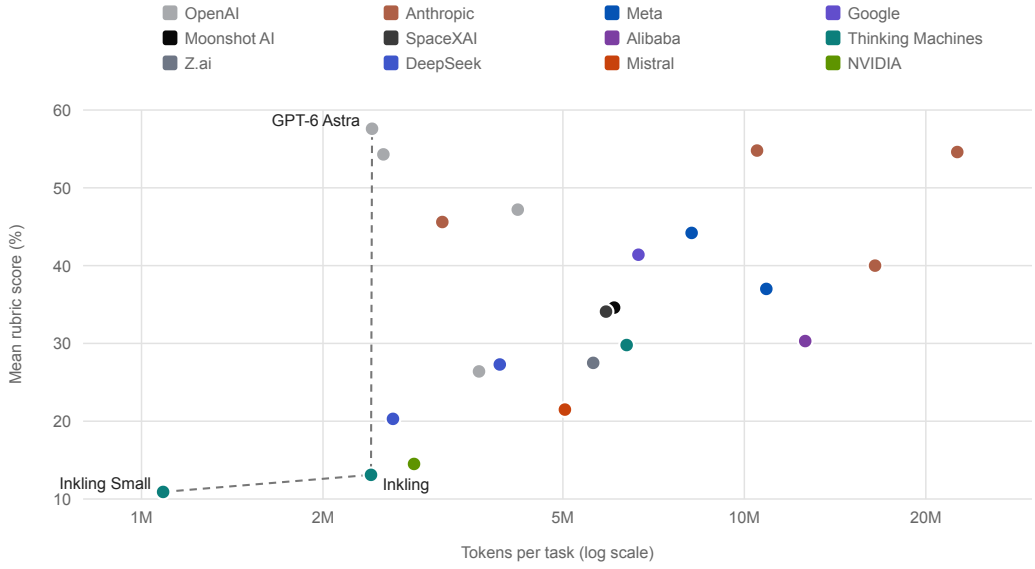


Figure 3: Mean rubric score against mean tokens per task. The dashed line marks the Pareto frontier: configurations not dominated on both score and tokens. Frontier configurations are labeled.

Models often execute calculations correctly after selecting the wrong inputs. In a lending task, agents correctly calculated debt capacity using an incorrect earnings subtotal. In a pharmaceutical royalty task, they correctly applied payment caps and milestone rules to a forecast that omitted an entire sales channel. The agents knew how to perform the calculations but failed to verify the numbers feeding them.

Agents find decision-relevant evidence without acting on it. In one acquisition task, moving the closing date changed shareholder payouts and lowered the required offer by $1.25 million. Every attempt by frontier models missed this saving, even though several had identified the buyer's right to choose the date.

6 Conclusion

For AI to be useful in dealmaking, its work needs to hold up when someone checks it. Across high-finance assignments that take professionals 20 to 30 hours, agents must choose sound assumptions, use them consistently, and explain the decisions that matter.

DeepFinance evaluates both the result and the work behind it. Auditability gives professionals a way to check the analysis and gives researchers a way to find the mistakes that need fixing. This brings us closer to measuring whether agents can take an assignment from a data room to a decision people can rely on.

Acknowledgments

We thank the Julius Research team and the private equity and investment banking professionals who composed and reviewed the DeepFinance tasks.

References

[1] T. Patwardhan et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025. https://arxiv.org/abs/2510.04374.

[2] C. E. Jimenez et al. SWE-bench: Can language models resolve real-world GitHub issues? *arXiv preprint arXiv:2310.06770*, 2023. https://arxiv.org/abs/2310.06770.

[3] W. Huang, C. Lee, L. Tng, and S. Ge. DeepSWE: Measuring frontier coding agents on original, long-horizon engineering tasks. Datacurve, May 2026. https://deepswe.datacurve.ai/blog/deepswe.

[4] P. Islam et al. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023. https://arxiv.org/abs/2311.11944.

[5] R. Koncel-Kedziorski et al. BizBench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*, 2023. https://arxiv.org/abs/2311.06602.

[6] S. Mateega et al. FinanceQA: A benchmark for evaluating financial analysis capabilities of large language models. *arXiv preprint arXiv:2501.18062*, 2025. https://arxiv.org/abs/2501.18062.

[7] A. Wang et al. BigFinanceBench: A workflow-grounded benchmark for financial-research agents. *arXiv preprint arXiv:2606.03829*, 2026. https://arxiv.org/abs/2606.03829.

[8] Databricks AI Research. Introducing OfficeQA: A benchmark for end-to-end grounded reasoning. Databricks, December 2025. https://www.databricks.com/blog/introducing-officeqa-benchmark-end-to-end-grounded-reasoning.

[9] Halluminate. Westworld Finance Diligence Bench. Halluminate, August 2026. https://www.halluminate.ai/blog/due-diligence-bench.

A Sample Task Details

This appendix expands the investment due diligence task in Section 4.2. Amounts are in USD millions unless stated.

A.1 Data Room

Table 4 lists the 18 files in the task's data room.

Table 4: Files in the sample task's data room.

File	Type	Size (KB)
01_Corliss_Close_Ledger.docx	Word document	37
02_Corliss_Supporting_Registers.docx	Word document	46
03_Corliss_Management_Pack.docx	Word document	37
04_Corliss_Executed_Instruments.docx	Word document	39
05_Corliss_Pricing_Framework.docx	Word document	37
06_Corliss_Financial_Reporting_Policy.docx	Word document	36
07_Ellery_Motion_Borrowings_20260109.docx	Word document	37
08_Corrib_Industrial_Debt_Schedule_20260109.docx	Word document	36
09_Corliss_Share_Rights_and_Option_Plan.docx	Word document	37
10_Corliss_Securities_Register.docx	Word document	37
11_Corliss_Sale_Mandate_and_Costs.docx	Word document	37
12_Corliss_IC_Price_Check_Template.xlsx	Spreadsheet	6
13_Corliss_Working_20260106.xlsx	Spreadsheet	6
14_Email_Re_warranty_costs.eml	Email	1
15_Email_Whitlow_invoice.eml	Email	1
16_Email_Q4_register_D-284_dates.eml	Email	1
17_Investment_Team_Chat_Export.txt	Text file	1
18_Email_Quarrie.eml	Email	1

A.2 Reference Solution

The reference solution is a workbook, `Corliss_IC_Price_Check.xlsx`, with nine sheets: Price check, Allocation, Workings, Completion date, Structures, Assumptions, Waterfall, Inputs, and Audit checks. Table 5 summarizes the Price check sheet.

Table 5: Key figures from the reference solution's Price check sheet (USD millions).

Item	Amount
Reconciled FY2025 revenue	42.880
Reconciled FY2025 EBITDA	7.180
Sustainable EBITDA	7.520
Peer median EV/EBITDA	8.735x
Enterprise value	65.684
Net debt	11.290
Working capital adjustment	(0.517)
Equity value	53.877
Price to put to IC (all cash, completion April 1, 2026)	50.803
Hale group cash at completion at that price	18.500
Declan's structure: price needed at completion (April 1)	50.603
Declan's structure: expected cost, with the 3.00 Quarrie payment at 76.95%	52.911

Two judgments drive the answer. First, sustainable EBITDA includes nothing for Quarrie: both parts need the same Supplier Quality Board approval (95%) and then their own separate field trials (90% each), so the chance both enter production is $95\% \times 90\% \times 90\% = 76.95\%$, below the pricing framework's 80% threshold. Second, the sale mandate lets the buyer nominate the completion date. Nominating April 1, 2026, after the third anniversary of the Series A issue, lowers the price needed to meet the Hale condition (18.500 in cash at completion for the Hale group) from 52.058 to 50.803. The reference solution's recommendation reads: "Offer 50.803 in cash with no Quarrie payment, nominating completion on April 1, 2026 (any business day in April up to the long-stop). Declan's structure costs more on the framework's probability-weighted basis on either date."

A.3 Rubric

Table 6 shows 9 of the task's 136 rubric criteria, three from each category.

Table 6: Sample rubric criteria for the sample task (9 of 136 shown).

Category	Criterion
Output	The workbook reports reconciled FY2025 revenue of 42.880.
Output	The workbook reports sustainable EBITDA of 7.51 to 7.53.
Output	The workbook puts a price of 50.80 to 50.81 to IC. Justifiable alternative: 50.803 or 50.804; the exact price is 50.8034 and 50.804 is the three-decimal price that clears 18.500.
Workflow	The workbook shows a line-by-line bridge from booked revenue to reconciled revenue, with the record behind each line.
Workflow	The workbook calendarizes Ellery's earnings: the year to March 2025, less the nine months to December 2024, plus the nine months to December 2025.
Workflow	The workbook counts the shared EC-22 approval once, times 90% for each part's own field trial.
Judgment	The workbook treats the equity value, not Owen's 56, as the most Pinecrest may pay.
Judgment	The workbook recommends all cash with no Quarrie payment.
Judgment	The warranty row leaves the IC price whole, with no deduction.