

DEALTRACE: Localizing and Intervening on Failures in Agentic Financial Reasoning

Alina Hyk

Halluminate

alina@halluminate.ai

Abstract

Analyzing a financial deal is a long, multi-step task: an analyst reads a narrative memorandum against the Excel model behind it, rebuilds the forecast under the valuation, and writes a recommendation. LLM agents increasingly perform such work, but financial benchmarks measure isolated skills or final answers and do not show where a long analysis fails. We present DEALTRACE, a benchmark built on real deal packages, each pairing a confidential information memorandum with its financial model. A five-stage pipeline—extract, reconcile, project a masked forecast, assess market context, recommend—follows the analyst workflow; every clause of the recommendation cites the stage that produced it, and knockout and repair interventions attribute the outcome to individual stages. Across six frontier models, three scaffolding conditions, and four model judges, extraction, reconciliation, and grounding are reliable, and forecasting is the decisive stage: it is the strongest correlate of judged quality, the costliest stage to remove, and the only repair that improves every model, after which the models reach nearly the same score. Models flag discrepancies accurately but rarely use them in the verdict, a failure that outcome-level evaluation cannot detect. Scaffolding improves stage-level performance far more than the final score, and a critique pass adds cost without reliable gain. DEALTRACE evaluates the analysis process itself, not only its final answer.

1 Introduction

Large language models are increasingly deployed as agents doing the multi-step work of a domain expert: gathering evidence from messy sources, calling tools, and holding a long reasoning chain together toward an output someone acts on (Guo et al., 2024; Chen et al., 2025b; Xu et al., 2024). Finance is a sharp case: a real deal analysis ends in an investment recommendation, yet existing

benchmarks test isolated skills, numerical question answering (Chen et al., 2021; Zhu et al., 2021), long-context filing QA (Islam et al., 2023; Reddy et al., 2024), program synthesis (Krumdick et al., 2024), mathematical reasoning (Zhao et al., 2024a), and trading decisions (Li et al., 2025a), and score final answers, which records whether a recommendation is right but not where the reasoning broke. Process-level evaluation localizes failure but stays tied to short, clean chains in mathematics and adjacent settings (Lightman et al., 2024; Wang et al., 2024; Zheng et al., 2025a; Zeng et al., 2024; Song et al., 2025), not the long, document-grounded reasoning a deal demands. Models rarely locate their own errors (Tyen et al., 2024), and the chain of thought often fails to reflect the computation behind the answer (Turpin et al., 2023; Chen et al., 2025a; Lanham et al., 2023), so an evaluation must check the reasoning externally rather than trust the model’s account of it.

DEALTRACE closes that gap. Built on real deal packages pairing a memorandum-style document with its Excel model, it scores a five-stage pipeline (extract, reconcile, project a masked forecast, assess market context, recommend) in which every recommendation clause cites the stage that produced it. Provenance localizes failure; a knockout that removes a stage and a repair that substitutes its ground truth then test whether the stage is consequential, turning the score into a diagnosis. We contribute: **1)** DEALTRACE, evaluating the complete deal-analysis workflow on live transaction materials rather than isolated skills on synthetic ones; **2)** a harness whose typed, citable stage outputs are graded deterministically wherever possible and by judges only where necessary; **3)** knockout and repair as a general recipe for attributing an agent’s output to individual stages; **4)** deployment guidance: effort belongs in forecasting, flagged risks need a carry-through mechanism, structure

outperforms critique.

2 The DEALTRACE Benchmark

The deal packages. DEALTRACE is built on ten complete private-equity and merger-and-acquisition deal packages: a confidential information memorandum or management presentation, the deal team’s Excel model, and often supporting disclosures, the bundle an analyst receives on a live deal. Packages vary along the three axes an analyst must absorb: the target’s industry, the model’s structure, and the memorandum’s drafting conventions; normalizing that variation away would substitute an easier task for the real one. A memorandum is a selling document whose headline claims do not always tie to the workbook, and the number that settles a claim sits among thousands of cells that do not matter, so surfacing gaps and locating evidence are the task; synthetic benchmarks do not reproduce this (Ravnik et al., 2026). The unit of evaluation is the graded item, not the deal: each package yields roughly a hundred checkable citations, several thousand graded decisions.

Because DEALTRACE is built on live deal materials rather than reconstructions, the corpus carries the confidentiality obligations of the deals it came from, and the packages cannot be redistributed. We therefore release the staged harness, the stage schema, the deterministic metric code, full aggregate results, and sample de-identified, synthetic-mirror corpus to allow for faithful replication of our work.

A five-stage pipeline. We split the task into five stages that follow an analyst’s workflow, each emitting a typed, citable artifact: extract the facts that matter from the documents and the model; reconcile the narrative against the model and flag mismatches; project the masked forecast; reason about market context; and recommend. The decomposition is deliberate with each stage being a distinct analyst competency and a distinct failure mode a reviewer checks for, with boundaries at the checkable intermediates an analyst produces (an extracted figure, a flagged discrepancy, a forecast cell). Decomposing a long task into typed sub-tasks is known to improve reasoning and make it auditable (Zhou et al., 2023a; Lightman et al., 2024); it also enables the benchmark’s core purpose, since a failure can be attributed to a stage only if the stages are separated. For instance, a

wrong recommendation built on a correct forecast is a different failure from a confident recommendation built on a forecast no one checked.

Stage artifacts and provenance. Stages communicate through a single typed envelope rather than free text. Stage k reads the artifacts of the earlier stages and writes its own, and every item carries an identifier and a source pointer. This design is deliberate. Free-text handoffs propagate errors in pipelined LLM systems (Cemri et al., 2025), and structured inter-role artifacts are the standard remedy (Hong et al., 2024; Wu et al., 2023a). Typing also makes the evaluation computable: only atomic, addressable claims can be graded, cited, knocked out, or repaired (Min et al., 2023; Rashkin et al., 2023; Gao et al., 2023). Because format restrictions can degrade reasoning (Tam et al., 2024), the envelope constrains only the interface between stages, and within a stage the model reasons freely. Because every recommendation sentence cites the items that produced it, any conclusion, and any failure, is attributable to a specific stage by construction rather than by reconstruction. The trace is checked evidence, not narrated reasoning: unlike a chain of thought, which often fails to reflect the model’s actual computation (Turpin et al., 2023; Chen et al., 2025a), every typed assertion is verified against its cited source.

Masked forecasting. In the projection stage the agent reconstructs the deal team’s forecast: the revenue, margin, and cash-flow lines over the deal’s horizon from which the valuation is computed. These cells are removed from the workbook before the agent sees it, so their values must be derived from the historical financials and the stated assumptions rather than read. Each sheet’s forecast region is masked as a block, so a masked total cannot be recovered by summing visible components. Scoring is restricted to a fixed set of canonical cells, the same financial fields over the same horizon in every deal, making forecasts comparable across deals and models.

Agent access to the package. The agent has full read access to the package through a tool interface over the two sources: ten operations for listing, reading, and searching the workbook, four for the documents, and a calculator over literals (Appendix A). Cell reads return the cached computed value and, where present, the formula, which distinguishes hard-coded inputs from derived figures; masked cells are never returned. Documents that restate the forecast in prose are withheld at ev-

ery stage, not only during projection, and an automated check confirms zero leakage of masked values in the tool output at every stage, further verified by a human annotator. Every assertion must include an exact reference, a cell or a page and span, and these references are what the evaluation checks.

3 Evaluation

Ground truth. Each deal carries a curated ground-truth annotation, verified by a human annotator. Across the ten deals analysed in this paper, the annotation pins 599 targets: the true values of the 362 forecast cells masked for prediction (Section 2), 30 expected reconciliation verdicts and 207 critical facts (113 workbook, 94 memorandum; Appendix J), with the workbook itself, confirmed through the spreadsheet tool (Section 2), as the extraction reference.

Deterministic metrics. Each metric compares a stage’s cited output to the ground-truth targets above. Extraction and reconciliation are scored by precision: the share of cited items matching the source. Forecasts are scored by accuracy at five and ten percent relative error, mean absolute percent error, and coverage of the scored grid, so accuracy cannot be inflated by answering only easy cells. Grounding is the share of market quotes and recommendation citations that verify against their sources. Risk propagation is the share of flagged reconciliation findings that reach the recommendation. Overall accuracy pools every gradable item. Exact definitions and recall against the curated targets are in Appendices B and J.

A panel of four model judges. The market and recommendation stages produce prose that deterministic checks only partly cover, so four frontier models grade them as judges (Zheng et al., 2023): Claude Opus 4.7, Claude Fable 5, GPT-5.6 Sol, and Gemini 3.6 Flash. Each judge grades every run under an identical rubric and prompt, and results are reported as the four-judge panel mean, with per-judge scores in Table 1 and the appendix. The judges agree on ordering and differ mainly in leniency: mean pairwise $r = 0.89$, $\text{ICC}(2, 4) = 0.95$ for the panel mean, and no own-family favoritism (Appendix E). Per claim, the judge grades whether each market or recommendation claim follows from its own cited evidence; per run, it returns a *holistic score* on a hundred-point scale rating the analysis as a whole, together

with a reliability label and a committee-readiness ruling from a fixed rubric (Appendix C; prompts in Appendix D). The *panel score* reported throughout is the mean of the four judges’ holistic scores.

The three experimental conditions. To separate what a model does from what the harness does for it, each model is run under three harness conditions with various levels of scaffolding; the full details are available in Appendix D.

Unguided. The baseline supplies the task and the tools and nothing else. It measures the model’s ability to organize its reasoning process independently, without specialized scaffolding.

Outline. Adds the full scaffold: the canonical field list, naming the facts and forecast cells each stage must produce; the staged method, fixing the five stages and their order; and per-stage guidance, describing what a complete artifact at each stage contains.

Outline + critique. Appends a single critic pass: the critic reads the staged output and its citations, flags unsupported claims, missing reconciliations, and blank forecast cells, and the model revises once. The critic never states what the correct value would be, so the condition tests self-correction under realistic information, a setting in which it is known to be difficult (Tyen et al., 2024). Appendix G examines critique separately.

4 Results

We evaluate six frontier models: Claude Opus 4.7, GPT-5.5, and Gemini 3.1 Pro, together with Claude Fable 5, GPT-5.6 Sol, and Gemini 3.6 Flash. All six run on the full benchmark under all three conditions. In the critique condition the critic is a separate model from the analyst: for Fable, Sol, and GF 3.6 the analyst is the smaller family variant and the critic is the flagship (Appendix G). Every run is graded by all four judges of Section 3; judge-scored tables report each judge’s score and the panel mean, and Appendix E analyses their agreement ($r = 0.89$, $\text{ICC}(2, 4) = 0.95$, no own-family favoritism).

4.1 Overall performance

Table 1 reports each judge’s holistic score, the four-judge panel mean, overall deterministic accuracy, measured cost, and efficiency per model and condition. Claude Fable 5 attains the highest pooled panel score and GPT-5.5 the second highest. The between-model analysis of variance

		Opus	GPT-5.5	G. Pro	Fable	Sol	GF 3.6
Ung.	by Opus	46.3	51.6	42.9	61.4	56.1	53.0
	by Fable	47.7	50.4	43.4	56.4	49.8	49.5
	by GF 3.6	50.5	58.7	49.2	65.3	59.6	54.0
	by Sol	46.9	57.3	46.7	57.9	57.6	50.2
	Panel	47.9	54.5	45.6	60.2 [*]	55.8	51.7
	Acc. (%)	74.3	83.9	67.0	84.5 [*]	75.1	67.0
	Cost (\$)	63.6	46.7	23.7	146.4	20.5	17.2
	Effic.	0.56	0.98	1.29	0.35	2.05[*]	2.02
	by Opus	53.0	57.0	47.9	62.9	61.5	48.7
	by Fable	59.1	63.1	52.2	65.7	65.5	54.0
Out.	by GF 3.6	60.0	62.5	51.5	69.8	71.8	54.3
	by Sol	53.2	62.9	54.3	63.7	66.7	50.8
	Panel	56.3	61.4	51.5	65.5[*]	66.4[*]	52.0
	Acc. (%)	81.3	85.2	67.4	89.4[*]	87.3	66.8
	Cost (\$)	95.0	55.5	26.5	207.8	28.8	22.0
	Effic.	0.48	0.94	1.31	0.28	2.01 [*]	1.58
	by Opus	47.9	62.6	48.9	52.8	43.3	43.4
	by Fable	51.8	64.2	47.9	58.1	51.2	41.2
	by GF 3.6	48.0	67.4	52.3	62.4	53.2	43.3
	by Sol	46.2	64.6	48.3	57.5	53.7	45.4
O+C [†]	Panel	48.5	64.7[*]	49.3	57.7	50.4	43.3
	Acc. (%)	77.8	87.0[*]	63.3	77.2	73.0	71.5
	Cost (\$)	183.6	161.3	137.0	724.0	19.2	16.6
	Effic.	0.21	0.35	0.23	0.06	1.91 [*]	1.86

Table 1: Per-judge scores, their panel mean, accuracy, cost, and efficiency by condition, over the ten deals. Columns are the models run; the “by” rows are the *judges* scoring them and Panel is their mean. Names, used in all tables: Opus is Claude Opus 4.7, Fable is Claude Fable 5, Sol is GPT-5.6 Sol, GF 3.6 is Gemini 3.6 Flash, G. Pro is Gemini 3.1 Pro; Ung., Out., and O+C are the unguided, outline, and outline + critique conditions. Efficiency is the panel score weighted by accuracy per dollar of ten-deal cost, taken from per-call usage ledgers (Appendix M). In the Panel, accuracy, and efficiency rows, **bold** marks the best condition per model and ^{*} the best model in that condition. O+C is solo self-critique for Opus, GPT-5.5, and G. Pro. [†]In the O+C block, Fable, Sol, and GF 3.6 are hybrid runs critiquing Claude Sonnet 5, GPT-5.6 Luna, and Gemini 3.1 Flash-Lite respectively, scored on the analyst and costed as analyst plus critic (Appendix G).

pooled over conditions is significant on the panel score ($F = 7.64$, $p < 0.001$) and on accuracy ($F = 7.00$, $p < 0.001$); in Holm-corrected pairwise tests, Claude Fable 5 and GPT-5.5 each exceed Claude Opus 4.7 and both Gemini models on the panel score, and both Gemini models on accuracy. GPT-5.6 Sol posts the highest outline panel score (66.4), although its pooled mean carries the daggered hybrid cell and significantly exceeds only Gemini 3.6 Flash. Cost, meanwhile, varies by an order of magnitude at comparable quality: that outline run cost \$28.8 over the ten

deals against \$207.8 for Claude Fable 5’s 65.5, one point apart at seven times the cost and not significantly different, and GPT-5.6 Sol is the most efficient model in every condition while Claude Fable 5 is the least.

Within models, outline improves on unguided for five of six on the panel (+5.3 to +10.6 points), although no single paired test reaches significance at ten deals ($p \geq 0.085$). Pooled across models, the condition analysis of variance is significant on the panel score ($F = 5.69$, $p = 0.008$ after Holm across the two summary metrics) and not on accuracy ($F = 1.27$, $p = 0.28$). In Holm-corrected pairwise Welch t -tests, the outline mean of 58.8 exceeds both the unguided mean of 52.8 (+6.1, $p_{\text{Holm}} = 0.015$) and the outline + critique mean of 52.2 (+6.7, $p_{\text{Holm}} = 0.015$), which do not differ from each other ($p = 0.77$).

Critique is the only condition change with significant within-model movement, and the effect is negative. Under self-critique, Claude Opus 4.7 drops 7.9 points ($p = 0.019$, paired by deal), while GPT-5.5 (+3.3) and Gemini 3.1 Pro do not change significantly. The daggered O+C cells are hybrid runs pairing a smaller analyst with its family’s flagship as critic (Claude Sonnet 5 with Claude Fable 5, GPT-5.6 Luna with GPT-5.6 Sol, Gemini 3.1 Flash-Lite with Gemini 3.1 Pro). A flagship critic benefits only the strongest analyst: Claude Sonnet 5 gains +5.9 panel points, positive under every judge but not significant ($p = 0.09$ uncorrected), GPT-5.6 Luna is unchanged (−0.2), and the Gemini 3.1 Flash-Lite pair cannot be evaluated because its analyst produces a median of five forecast cells per deal against forty to fifty for the others. In every family the critiqued analyst remains well below the flagship performing the task alone (cell differences −7.8, −16.0, and −8.1; pooled paired −11.2, $p < 0.001$ after Holm), a gap attributable to the smaller analyst rather than to critique itself, and at higher cost (\$724 for the Claude Sonnet 5 pair against \$208 for Claude Fable 5 alone). A critic does not recover quality lost elsewhere in the chain, and scaffolding’s effects are concentrated at the stage level rather than in the holistic assessment (Section 4.2). The full hybrid design, per-judge scores, and costs are in Appendix G.

4.2 Stage-level effects of scaffolding

Although panel scores change little across conditions (Section 4.1), stage-level performance

Model	Targeted acc. (%)			Coverage (%)			Forecast quality		
	Ung.	Out.	O+C	Ung.	Out.	O+C	Ung.	Out.	O+C
Opus	5.9	34.8	23.6	34.9	89.6	94.7	0.09	0.36	0.24
GPT-5.5	8.2	40.5	45.4	22.3	88.5	90.0	0.19	0.42	0.47
G. Pro	3.2	37.7	18.8	14.7	77.6	70.6	0.11	0.40	0.20
Fable	13.4	36.3	39.9 [†]	36.5	85.7	93.6 [†]	0.19	0.38	0.41 [†]
Sol	8.6	34.0	18.0 [†]	31.7	85.5	92.4 [†]	0.14	0.35	0.20 [†]
GF 3.6	10.6	25.1	0.6 [†]	38.5	81.9	9.5 [†]	0.16	0.26	0.03 [†]

Table 2: Forecasting by condition, mean across ten deals. Targeted accuracy is the share of standardized cells within five percent; coverage is the share of the grid predicted at all; forecast quality is $0.8 \times$ targeted accuracy + $0.2 \times$ the within-5% rate over all predicted cells with a workbook value. [†]For Fable, Sol, and GF 3.6 the O+C forecasts are produced by the hybrid run’s small analyst (Sonnet 5, GPT-5.6 Luna, Gemini 3.1 Flash-Lite) under that model’s critique (Appendix G), so these cells measure the analyst’s model forecasting.

changes substantially. Under outline, forecast coverage rises from 14–39 to 78–90 percent and targeted accuracy from below 14 to 25–41 percent in every model (Table 2), with error increasing over the forecast horizon (Figure 1). Sign agreement remains at 96–100 percent, so residual error is one of magnitude rather than direction. Reconciliation is gradable only under scaffolding, reaching 93.4 percent precision under outline and 88.1 under critique, the difference attributable to the hybrids’ smaller analysts (Appendix I). Scaffolded recommendations draw on a broader set of upstream stages (Figure 8) and stay on the prescribed grid, with 1192 off-grid predictions under outline against 3332 unguided and higher accuracy on those that remain (Appendix K). All three forecast metrics differ by condition after Holm correction, with outline above unguided on each ($p_{\text{Holm}} < 0.001$).

Critique yields mixed stage-level effects: only GPT-5.5 improves both coverage and targeted accuracy, Claude Opus 4.7 gains coverage at the cost of accuracy, and Gemini 3.1 Pro declines on both. It is also the expensive part of the scaffold, raising cost by a factor of 1.9 to 5.2 (Appendix M), and the outline runs dominate on quality per dollar (Table 1).

4.3 Reconciliation and risk propagation

Models reconcile the narrative against the financial model with near 90 percent precision, and their recommendation citations almost always resolve to real upstream items (Table 3). Precision

		Opus	GPT-5.5	G. Pro	Fable	Sol	GF 3.6
		Opus	GPT-5.5	G. Pro	Fable	Sol	GF 3.6
S4 grounding	Out.	93.2	97.4	90.0	93.3	94.0	99.0
	O+C	77.2	93.9	85.8	73.0 [†]	89.9 [†]	86.8 [†]
Risk prop.	Out.	22.2	28.4	28.9	37.3	1.6	6.1
	O+C	17.7	7.7	13.2	43.8 [†]	6.8 [†]	23.8 [†]

Table 3: Stage-4 grounding and risk propagation, mean across deals, as percentages. Both are blank under the unguided baseline, which emits no gradable market or reconciliation citations. [†]O+C is the hybrid run in which that model critiques a smaller same-family analyst: Fable with Sonnet 5, Sol with GPT-5.6 Luna, and GF 3.6’s column with G. Pro critiquing Gemini 3.1 Flash-Lite (Appendix G).

is computed over every finding the model asserts, a median of 14 to 53 per deal, and extraction recall against the curated targets is 28 to 71 percent (Appendix J). Flagged risks, however, reach the recommendation only 2 to 37 percent of the time, and strong grounding does not imply strong propagation: GPT-5.6 Sol and Gemini 3.6 Flash ground their market claims at 94 and 99 percent yet propagate risk least, at 1.6 and 6.1 percent. An ANOVA across models, one observation per (model, deal) run pooled over outline and critique, finds models differ in risk propagation ($F = 3.80$, $p = 0.007$ after Holm over the two metrics), with only the 36-point Claude Fable 5 to GPT-5.6 Sol gap surviving correction, but not in grounding ($F = 1.83$, $p = 0.11$). Paired by deal, critique reduces grounding (-10.0 , $p_{\text{Holm}} < 0.001$) and leaves propagation unchanged (-3.2 , $p = 0.58$). Discrepancies are identified in stage two and discarded by stage five, consistently across all six models and invisibly to outcome-level evaluation.

5 Ablations: Intervening on the Reasoning Chain

The provenance trace localizes failure, and the ablations test whether a localized stage is consequential and empirically supported. The knock-out removes a single stage and re-runs the remaining chain, and the repair replaces a stage with its ground truth before re-running. Together the two interventions quantify how strongly the final recommendation depends on each stage, which final-answer evaluation cannot measure. The correlational analysis includes all six models, and the interventions include Opus, Fable, Sol, and GF 3.6. Models and judging are identical to Section 4.1.

Which stages drive the panel score. Be-

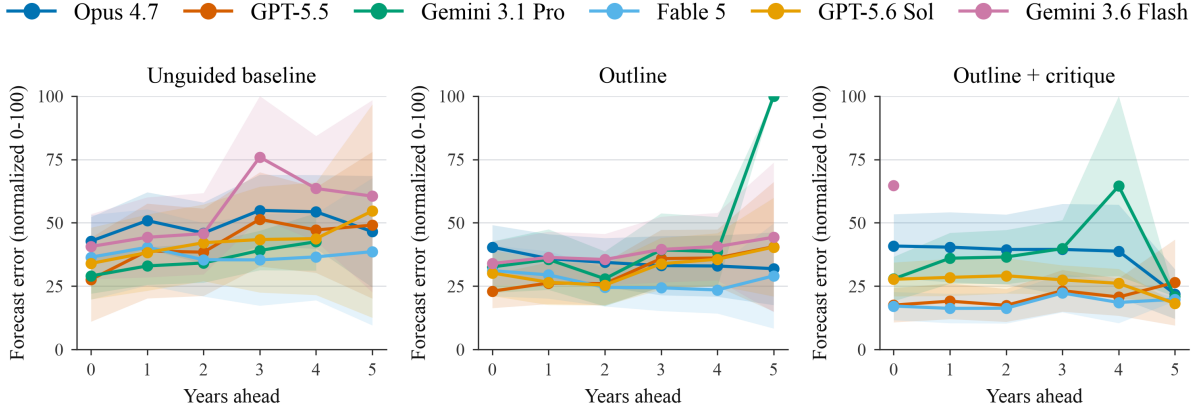


Figure 1: Forecast error by predicted year, one panel per condition and one line per model. Shading is the standard error across deals; error is symmetric MAPE on magnitudes, so sign-convention differences are not scored as errors, and each panel is normalized to 0–100 within itself. In the outline + critique panel the Fable, Sol, and GF 3.6 lines are their hybrid runs and the single point is the Flash-Lite analyst (Appendix G).

fore intervening, we ask which stage metrics the panel score tracks. With each (model, condition, deal) run as one observation, pooling the main-grid, hybrid, knockout, and repair runs and excluding runs without gradable stage output, we regress the panel score on the five stage metrics with standardized predictors ($n = 313$). Forecast quality dominates at $\beta = 0.77$ and univariate $r = 0.85$, far above extraction precision (0.21 and 0.42), grounding (0.18 and 0.31), and reconciliation precision (0.13 and 0.26); every stage is significant at $p < 0.001$ except risk propagation ($\beta = 0.01$, $p = 0.38$; Table 11 in Appendix F). The panel score is, to first order, a function of forecasting, and the same correlation holds per model ($r = 0.68$ to 0.96).

Knockout: which stages are necessary. The knockout removes a single stage and re-runs the rest of the chain; a positive value is the drop in panel score caused by that removal. Removing the forecast costs more than removing reconciliation for all four models and under every judge (Table 4, Appendix F). Relative to each outline base, forecast removal costs 4.0 points for Claude Opus 4.7, 10.2 for Claude Fable 5, 11.3 for GPT-5.6 Sol, and -0.3 for Gemini 3.6 Flash, against -2.4 , 5.5 , 5.3 , and -5.3 for reconciliation removal, which leaves Claude Opus 4.7 and Gemini 3.6 Flash slightly higher than their base. The within-model ANOVA over the base and the three knockouts is significant for Claude Fable 5 ($F = 4.06$, $p = 0.014$) and GPT-5.6 Sol ($F = 4.54$, $p = 0.008$), and not significant for Claude Opus 4.7 ($p = 0.080$) or Gemini 3.6 Flash ($p = 0.15$). Removing both stages

Condition	Opus	Fable	Sol	GF 3.6
Outline base	56.3	65.5	66.4	52.0
–reconcile	58.7	60.1	61.0	57.3
–forecast	52.3	55.4	55.1	52.2
–both	49.0	51.7	52.7	48.2

Table 4: Knockout, four-judge panel score, mean across deals, for the four ablated models.

also narrows the per-deal spread (Claude Fable 5 from SD 12.6 to 5.4, GPT-5.6 Sol from 11.7 to 4.7), so reconciliation and forecasting carry most of the deal-to-deal variation.

Repair: the effect of supplying correct upstream outputs. The repair intervention supplies the correct output of an upstream stage and re-runs the rest of the chain. Levels are cumulative, so each level’s change in the final score measures how much of the outcome that stage accounts for. Supplying the true forecast raises the four-judge panel score by 20.4 to 22.6 points in all four models, significant in a paired test by deal for each (all $p < 0.005$) and holding under every judge, with lifts of 17 to 27 points across the sixteen judge-model cells (Figure 2, Appendix F). With every stage repaired the models converge to 83–86 from starting points between 52.0 and 66.4, so the remaining gap between them is concentrated in forecasting. Upstream repairs move Claude Fable 5 and GPT-5.6 Sol by under three points, while Claude Opus 4.7 and Gemini 3.6 Flash gain 5.8 and 10.8 points from a true extraction.

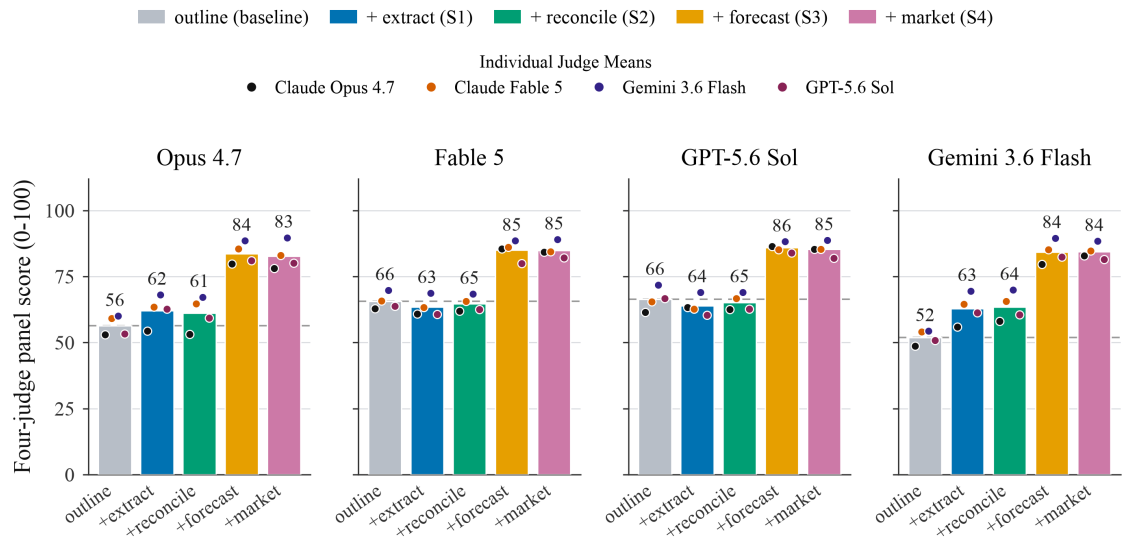


Figure 2: Repair chains for the four ablated models on the four-judge panel. Each level supplies the true output of one more upstream stage and re-runs the rest; the dashed line is the outline base. Coloured dots are the individual judge means in fixed order (Opus, Fable, GF 3.6, Sol), spanning 8.8 points on average and 14.1 at most.

6 Related Work

Finance reasoning and agent benchmarks. Financial benchmarks span single-document numerical QA (Chen et al., 2021; Zhu et al., 2021; Chen et al., 2022; Zhao et al., 2022), long-context filing QA (Islam et al., 2023; Reddy et al., 2024; Lai et al., 2024; Chen et al., 2024a), program synthesis (Krumdick et al., 2024), mathematical reasoning (Zhao et al., 2024a,b), and meta-benchmarks and domain models (Xie et al., 2024; Wu et al., 2023b; Liu et al., 2024b; Xue et al., 2024; Choi et al., 2025; Lu et al., 2025). The closest work reaches professional settings but evaluates the narrative or the financial model, not both: FinanceQA collapses once an answer requires an assumption and leaves spreadsheets untested (Mateega et al., 2025); the Finance Agent Benchmark names spreadsheet reasoning as open (Bigeard et al., 2025); spreadsheet suites probe manipulation and modeling (Yen et al., 2026; Ma et al., 2024; Li et al., 2023; Dong et al., 2024; Chen et al., 2024b; Thorne, 2025), and FinSheet-Bench argues for stage-separated, auditable evaluation (Ravnik et al., 2026); and trading benchmarks score returns rather than reasoning (Li et al., 2025a). None operates at live deal-package scale or attributes failure to the stage that produced it; DEALTRACE adopts the stage-separated, auditable design it identifies as necessary. Deal packages are also a long-context setting, where performance degrades when evidence lies mid-input (Liu et al., 2024a;

Bai et al., 2024; Hsieh et al., 2024).

Forecasting and numerical extrapolation.

The projection stage targets a contested weakness. Early work suggested zero-shot numeric extrapolation rivaling purpose-built forecasters (Gruver et al., 2023), but ablations find no convincing time-series gain and unchanged performance after re-initializing pretrained weights (Tan et al., 2024). The brittleness recurs in finance: numerical reasoning sits well below expert accuracy (Zhao et al., 2024a), and collapses once an answer requires an assumption under incomplete information (Mateega et al., 2025) or multi-step aggregation rather than lookup (Ravnik et al., 2026). Masking the forecast region makes this a measurable, leakage-resistant subtask isolated from extraction and reconciliation, which is what lets us show that the recommendation depends on it most.

Process-level and verifiable evaluation.

Outcome-only evaluation is increasingly treated as insufficient. Process supervision scores each step, from step-level annotation (Lightman et al., 2024) or label-free construction (Wang et al., 2024); a second line evaluates the supervisor itself (Zheng et al., 2025a; Song et al., 2025; Zeng et al., 2024), extended by error typing and reasoning structure (Pala et al., 2025; Li et al., 2025b; Khalifa et al., 2025) and by work in multimodal reasoning (Xu et al., 2025), finance (Zhu et al., 2025), and surveys (Zheng et al., 2025b). Localization matters because models rarely find their own errors yet can fix one whose

location is given (Tyen et al., 2024), and the chain of thought often fails to reflect the computation behind the answer (Turpin et al., 2023; Chen et al., 2025a; Lanham et al., 2023). The machinery for making such evaluation verifiable exists but has not reached structured financial reasoning: attribution frameworks (Rashkin et al., 2023; Bohnet et al., 2022), citation benchmarks (Gao et al., 2023), atomic-fact precision (Min et al., 2023), and grounding against a source (Jacovi et al., 2025), though audits find only about half of generated sentences fully supported (Liu et al., 2023a; Huang and Chang, 2024). DEALTRACE carries localization to a long, document-grounded task, replaces the chain of thought with checked provenance, and moves from detection to intervention. The masked projection supplies the held-out content contamination demands (Sainz et al., 2023; Zhou et al., 2023b); weak holdouts breed overfit agents (Kapoor et al., 2025b,a); and grading by four judges with symmetric agreement analysis addresses the biases of model-based judging (Zheng et al., 2023; Liu et al., 2023b) (Appendix E).

7 Discussion and Conclusion

DEALTRACE evaluates agents on real deal packages. The harness traces every clause of the recommendation to the stage that produced it and intervenes on the chain to attribute the outcome to individual stages. Three findings hold across all six models and all four judges. First, forecasting determines recommendation quality. It is the strongest correlate of the panel score (the four judges’ mean holistic rating of the analysis on a hundred-point scale) in every model, the costliest stage to remove, and the only repair that lifts every model, by roughly twenty-one points, so the quality gap between frontier models on this task is primarily a forecasting gap. Second, models reconcile the narrative against the financial model precisely and then rarely carry the flagged risks into the verdict. The failure is consistent across every model tested and invisible to outcome-level evaluation. Third, scaffolding (the outline of canonical fields, staged method, and per-stage guidance) changes stage-level behavior far more than the panel score, raising forecast coverage several-fold and bringing reconciliation and market evidence into the memo while the score moves by a few points, and critique raises cost by a factor of two

to five without reliable gain. The design transfers to any task that reads a narrative source against a structured one, such as a clinical note against a chart. The evaluation is inexpensive to re-run and difficult to game: the deterministic stages require no model in the loop, the masked forecast resists contamination, and every conclusion survives a change of judge. DEALTRACE thus shows which abilities can be trusted, not just whether the answer is right.

Limitations

Scale. The corpus of ten deal packages used in the presented evaluations limits deal-level power: within-model condition differences and per-model knockouts are directional rather than individually significant. The evaluation itself has rich coverage, with 599 human-verified targets and a stage regression over $n = 313$ runs; a larger deal corpus would nonetheless strengthen the findings.

Critique coverage. Claude Fable 5, GPT-5.6 Sol, and Gemini 3.6 Flash were not run under solo outline + critique; the hybrid design of Appendix G answers the critique question for these families, leaving only the Gemini pair inconclusive.

Recall and propagation bounds. Reconciliation recall is reported against curated targets only (Appendix J), since no exhaustive discrepancy inventory exists for a live package. Risk propagation credits cited carry-through. The stage regression is unchanged with deal and model fixed effects absorbed (Appendix F).

Forecasting tails. Forecast misses are heavy-tailed, so mean absolute percent error is reported alongside within-five-percent accuracy rather than alone.

Data Statement

DEALTRACE uses live, business-confidential private-equity and M&A deal materials that cannot be redistributed. The staged harness, the deterministic metric code and a synthetic-mirror corpus for replication are released at <https://github.com/Halluminate/DealTrace>. For additional, aggregate results see Appendix.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao

- Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. [LongBench: A bilingual, multitask benchmark for long context understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. 2025. [Finance agent benchmark: Benchmarking LLMs on real-world financial research tasks](#). *arXiv preprint arXiv:2508.00828*.
- Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roei Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, and 1 others. 2022. [Attributed question answering: Evaluation and modeling for attributed large language models](#). *arXiv preprint arXiv:2212.08037*.
- Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*.
- Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024a. [FinTextQA: A dataset for long-form financial question answering](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, and 1 others. 2025a. [Reasoning models don't always say what they think](#). *arXiv preprint arXiv:2505.05410*.
- Yibin Chen, Yifu Yuan, Zeyu Zhang, Yan Zheng, Jinyi Liu, Fei Ni, Jianye Hao, Hangyu Mao, and Fuzheng Zhang. 2024b. [SheetAgent: Towards a generalist agent for spreadsheet reasoning and manipulation via large language models](#). *arXiv preprint arXiv:2403.03636*.
- Zhiyu Chen, Wenhua Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. [FinQA: A dataset of numerical reasoning over financial data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. [ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, and 1 others. 2025b. [ScienceAgent-Bench: Toward rigorous assessment of language agents for data-driven scientific discovery](#). In *International Conference on Learning Representations (ICLR)*.
- Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. 2025. [FinDER: Financial dataset for question answering and evaluating retrieval-augmented generation](#). *arXiv preprint arXiv:2504.15800*.
- Haoyu Dong, Jianbo Zhao, Yuzhang Tian, Junyu Xiong, Shiyu Xia, Mengyu Zhou, Yun Lin, José Cambonero, Yeye He, Shi Han, and Dongmei Zhang. 2024. [SpreadsheetLLM: Encoding spreadsheets for large language models](#). *arXiv preprint arXiv:2407.09025*.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling large language models to generate text with citations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. [Large language models are zero-shot time series forecasters](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. [Large language model based multi-agents: A survey of progress and challenges](#). In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*.
- Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiaowu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. [MetaGPT: Meta programming for a multi-agent collaborative framework](#). In *International Conference on Learning Representations (ICLR)*.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Krizan, Shantanu Acharya, Dima Rekesht, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. [RULER: What's the real context size of your long-context language models?](#) In *Conference on Language Modeling (COLM)*.
- Jie Huang and Kevin Chen-Chuan Chang. 2024. [Citation: A key to building responsible and accountable large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. [FinanceBench: A new benchmark for financial question answering](#). *arXiv preprint arXiv:2311.11944*.
- Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas,

- Michelle Liu, Nate Keating, Adam Bloniarz, and 1 others. 2025. [The FACTS grounding leaderboard: Benchmarking LLMs’ ability to ground responses to long-form input](#). *arXiv preprint arXiv:2501.03200*.
- Sayash Kapoor, Benedikt Stroebel, Peter Kirgis, Nitya Nadgir, Zachary S. Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, and 1 others. 2025a. [Holistic agent leaderboard: The missing infrastructure for AI agent evaluation](#). *arXiv preprint arXiv:2510.11977*.
- Sayash Kapoor, Benedikt Stroebel, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. 2025b. [AI agents that matter](#). *Transactions on Machine Learning Research (TMLR)*.
- Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2025. [Process reward models that think](#). *arXiv preprint arXiv:2504.16828*.
- Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. [BizBench: A quantitative reasoning benchmark for business and finance](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Viet Dac Lai, Michael Krumdick, Charles Lovering, Varshini Reddy, Craig Schmidt, and Chris Tanner. 2024. [SEC-QA: A systematic evaluation corpus for financial QA](#). *arXiv preprint arXiv:2406.14394*.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. [Measuring faithfulness in chain-of-thought reasoning](#). *arXiv preprint arXiv:2307.13702*.
- Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K. P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. 2025a. [INVESTOR-BENCH: A benchmark for financial decision-making tasks with LLM-based agent](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. 2023. [SheetCopilot: Bringing software productivity to the next level through large language models](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Xiang Li, Haiyang Yu, Xinghua Zhang, Ziyang Huang, Shizhu He, Kang Liu, Jun Zhao, Fei Huang, and Yongbin Li. 2025b. [Socratic-PRMBench: Benchmarking process reward models with systematic reasoning patterns](#). *arXiv preprint arXiv:2505.23474*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *International Conference on Learning Representations (ICLR)*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paran-jape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics (TACL)*.
- Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023a. [Evaluating verifiability in generative search engines](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Shu Liu, Shangqing Zhao, Chenghao Jia, Xinlin Zhuang, Zhaoguang Long, Jie Zhou, Aimin Zhou, Man Lan, Qingquan Wu, and Chong Yang. 2024b. [FinDABench: Benchmarking financial data analysis ability of large language models](#). *arXiv preprint arXiv:2401.02982*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Guilong Lu, Xuntao Guo, Rongjunchen Zhang, Wenqiao Zhu, and Ji Liu. 2025. [BizFin-Bench: A business-driven real-world financial benchmark for evaluating LLMs](#). *arXiv preprint arXiv:2505.19457*.
- Zeyao Ma, Bohan Zhang, Jing Zhang, Jifan Yu, Xiaokang Zhang, Xiaohan Zhang, Sijia Luo, Xi Wang, and Jie Tang. 2024. [SpreadsheetBench: Towards challenging real world spreadsheet manipulation](#). In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Spencer Mateega, Carlos Georgescu, and Danny Tang. 2025. [FinanceQA: A benchmark for evaluating financial analysis capabilities of large language models](#). *arXiv preprint arXiv:2501.18062*.
- Sewon Min, Kalpesh Krishna, Xixi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Tej Deep Pala, Panshul Sharma, Amir Zadeh, Chuan Li, and Soujanya Poria. 2025. [Error typing for smarter rewards: Improving process reward models with error-aware hierarchical supervision](#). *arXiv preprint arXiv:2505.19706*.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. [Measuring attribution in natural language generation models](#). *Computational Linguistics*.

- Jan Ravník, Matjaž Ličen, Felix Bühmann, Bithiah Yuan, Felix Stinson, and Tanvi Singh. 2026. [FinSheet-Bench: From simple lookups to complex reasoning, where LLMs break on financial spreadsheets](#). *arXiv preprint arXiv:2603.07316*.
- Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krundick, Charles Lovering, and Chris Tanner. 2024. [DocFinQA: A long-context financial reasoning dataset](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)*.
- Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. [NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. [PRMBench: A fine-grained and challenging benchmark for process-level reward models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*.
- Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. 2024. [Are language models actually useful for time series forecasting?](#) In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Simon Thorne. 2025. [Large language models for spreadsheets: Benchmarking progress and evaluating performance with FLARE](#). *arXiv preprint arXiv:2506.17330*.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. 2024. [LLMs cannot find reasoning errors, but can correct them given the error location](#). In *Findings of the Association for Computational Linguistics: ACL 2024*.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. [Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023a. [AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework](#). *arXiv preprint arXiv:2308.08155*.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. 2023b. [BloombergGPT: A large language model for finance](#). *arXiv preprint arXiv:2303.17564*.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. [FinBen: A holistic financial benchmark for large language models](#). In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, and 1 others. 2024. [TheAgentCompany: Benchmarking LLM agents on consequential real world tasks](#). *arXiv preprint arXiv:2412.14161*.
- Zhaopan Xu, Pengfei Zhou, Jiaxin Ai, Wangbo Zhao, Kai Wang, Xiaojiang Peng, Wenqi Shao, Hongxun Yao, and Kaipeng Zhang. 2025. [MPBench: A comprehensive multimodal reasoning benchmark for process errors identification](#). *arXiv preprint arXiv:2503.12505*.
- Siqiao Xue, Xiaojing Li, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2024. [FAMMA: A benchmark for financial domain multilingual multimodal question answering](#). *arXiv preprint arXiv:2410.04526*.
- Thomson Yen, Julian Poeltl, Harshith Srinivas Gear, Yilin Meng, Joshua Fan, Adam Shen, Yili Liu, Ali Bauyrzhan, Patrick Shea, Siri Du, Haoyang Liu, Daniel Guetta, and Hongseok Namkoong. 2026. [MBABench: Evaluating LLM agents on end-to-end spreadsheet tasks in finance](#). *arXiv preprint arXiv:2605.22664*.
- Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, and 1 others. 2024. [MR-Ben: A meta-reasoning benchmark for evaluating system-2 thinking in LLMs](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. [MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

- Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024a. [Finance-Math: Knowledge-intensive math reasoning in finance domains](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024b. [DocMath-Eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025a. [ProcessBench: Identifying process errors in mathematical reasoning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Congmin Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and Weinan Zhang. 2025b. [A survey of process reward models: From outcome signals to process supervisions for large language models](#). *arXiv preprint arXiv:2510.08049*.
- 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. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023a. [Least-to-most prompting enables complex reasoning in large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023b. [Don't make your LLM an evaluation benchmark cheater](#). *arXiv preprint arXiv:2311.01964*.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. [TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Jie Zhu, Yuanchen Zhou, Shuo Jiang, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2025. [Fin-PRM: A domain-specialized process reward model for financial reasoning in large language models](#). *arXiv preprint arXiv:2508.15202*.

Tool	Returns
<i>Workbook</i>	
<code>list_sheets</code>	sheet names
<code>describe_sheet</code>	dimensions, populated region
<code>read_cell</code> , <code>read_range</code>	cached value + formula
<code>read_row_label</code>	row label of a cell
<code>read_column_header</code>	column header of a cell
<code>search_cells_by_value</code>	cells matching a value
<code>get_cell_format</code>	number format, styling
<code>find_formulas_referencing</code>	cells that consume a cell
<code>list_named_ranges</code>	named ranges
<code>compute</code>	arithmetic over literals
<i>Documents</i>	
<code>list_cim_documents</code>	documents in the package
<code>cim_outline</code>	section outline
<code>read_cim_page</code>	page text
<code>search_cim</code>	matching spans, by page

Table 5: The fifteen read-only tools through which the agent accesses the package.

A Agent tools and the document-exclusion protocol

The agent reaches the deal package only through read-only tools; it never receives a file. Table 5 lists the fifteen tools. `Workbook` reads return the cached computed value together with the formula where one exists, so hard-coded inputs are distinguishable from derived figures; document reads return page text the agent can quote and cite by page and span; and `compute` is a calculator over literals, so arithmetic is not done in prose. Masking is enforced at the tool layer, not the prompt layer: any call that touches a masked forecast cell returns a masked marker rather than a value, whether the cell is addressed directly, inside a range, through a named range, or via a formula that references it. The same layer applies the document-side guard at every stage, not only during projection: documents that restate the forecast (for example a projections section of the memorandum) are withheld from the tool index throughout the run, and page text returned by any retrieval call is scrubbed of figures that match masked-cell values, so the forecast cannot be recovered from a narrative restatement. Automated checks over the served tool outputs confirm that no masked value surfaced at any stage, and a human annotator further verified this on the retrieved pages. The critic in the outline + critique condition operates through exactly these tools with the same masks; it never sees ground truth (Section 3).

B Deterministic metric definitions

All deterministic metrics are computed per run and averaged across the ten deals. Every verdict is a mechanical comparison against the workbook, the documents, or the ground truth.

Extraction and reconciliation precision. The share of cited items whose claim matches the cited source, over items with a determinable verdict. An extraction item is correct when the value at its cited cell or span matches the asserted value within rounding, and a reconciliation item is correct when the two cited extraction items exist and the asserted relationship, match or mismatch and its direction, holds between them. Precision is the appropriate axis because the model chooses what to assert, and the metric asks whether those assertions survive checking. Items without a determinable verdict, such as claims about figures absent from the workbook, are excluded rather than guessed. Recall against the curated targets is reported in Appendix J.

Targeted forecast accuracy. The scored grid is a fixed set of canonical (field, year) cells, identical across deals, which makes forecasts comparable across deals and models. A predicted cell is correct at the five-percent level when $||\hat{v}| - |v||/|v| \leq 0.05$ against the true masked value v . Targeted accuracy divides correct cells by *all* cells the model was asked to predict, so an unattempted cell counts as a miss and selective answering is not rewarded.

Coverage. The share of asked cells the model predicted at all. Accuracy at ten percent and the mean absolute percentage error are computed on the same grid.

Error conventions. Errors are computed on magnitudes, so a sign-convention flip, such as costs reported as negatives, is not scored as a total miss. Sign agreement between prediction and truth is 96–100 percent in every model and condition, so this convention does not hide direction errors. The horizon-degradation figure (Figure 1) uses symmetric MAPE, $||\hat{v}| - |v||/(|\hat{v}| + |v|)$, which is bounded to $[0, 100]$ and stable when the true value is near zero.

Non-target accuracy and forecast quality. Predictions outside the scored grid that land on a workbook cell with a known value are graded at the same five-percent bar (Appendix K). Forecast quality is $0.8 \times$ targeted accuracy + $0.2 \times$ the within-five-percent rate over all predicted cells with a workbook value. The grid dominates the

weighting because an off-grid cell may have been readable directly rather than forecast; the smaller term credits correct work beyond the grid without letting it substitute for the required cells.

Grounding. The share of market claims whose quoted excerpt is verified on the cited page, and, for the recommendation, the share of citations that resolve to a real upstream item in the artifact store, with invented citations as the complement. Both checks grade the evidence trail itself, because a claim whose quote or citation does not resolve cannot be audited however plausible it reads.

Risk propagation. The share of misaligned reconciliation findings, the items that flag a genuine narrative-versus-model mismatch, that resurface in the recommendation. Matching is by the finding’s identifier in the recommendation’s citations, which makes propagation a deterministic check rather than a judgment about whether the memo addressed a risk. The identifier rule counts auditable carry-through: under the benchmark’s citation rule, a risk discussed without citing its finding cannot be traced to it, and such claims are graded unsupported at the per-claim level, so an uncited mention is a grounding failure rather than propagation.

Overall accuracy. The share of correct items over every deterministically gradable item in the run. Pooling weights each stage by how much checkable work it produces, and the result is the judge-independent summary reported beside the panel score in Table 1.

C Judge protocol and rubric

Each judge grades at two levels under an identical prompt, so differences between judges reflect the judge, not the protocol. **Per claim.** For every market and recommendation claim the judge receives the claim together with the exact upstream items it cites, and nothing else; the deterministic scorecard is not part of the per-claim input. It returns strict JSON: a support label (`supported`, `partially_supported`, `unsupported`), a faithfulness score (1–5, does the claim assert more than its evidence shows), and a relevance score (1–5, is the claim material to the investment decision). A claim citing no evidence cannot be supported; a claim contradicting its own evidence is unsupported. To limit position and verbosity bias, claims are presented in a fixed stage order and the judge scores support before fluency. **Per run.** The

judge is instructed as a senior partner on an investment committee grading an analyst’s full writeup, given the final recommendation text together with the deterministic scorecard and the specific items flagged wrong, the same verified record a reviewing partner would hold while reading the work. It first selects a rubric band, then places the score within it: 90–100 committee-ready; 75–89 strong, one revision; 60–74 competent but flawed, material errors to fix before trusting the conclusion; 40–59 unreliable, substantial rework; 20–39 poor, pervasive or wrong-target work; 1–19 unusable. A hard rule caps the score below the top bands only for degenerate output, a flat forecast, a forecast targeting the wrong entity, or a recommendation contradicting the analyst’s own numbers, regardless of prose quality; the rule is categorical and contains no graded accuracy term. The judge returns the band, the 1–100 score with a written justification, a reliability label (high/moderate/low), a binary committee-readiness ruling, and key failures and strengths, as strict JSON with a fixed generation budget. **The holistic score is not a restatement of the scorecard.** Because the per-run judge sees the deterministic scorecard, the stage-metric regression of Section 5 could in principle measure how faithfully judges follow their instructions, and the repair lifts of Section 5 could reflect a cleaner scorecard rather than a better memo. The repair grid separates the two. Every repair rung supplies the true output of one stage and therefore drives that stage’s scorecard entry to or near its ceiling in the re-run, yet the four rungs move the panel very differently: a true extraction changes the four models by -2.6 to $+10.8$ points, a true reconciliation by about one point, and a true market brief by essentially nothing, while the true forecast lifts every model by twenty points or more (Table 12). If the judges mechanically rewarded a cleaner scorecard, each rung would lift the score roughly in proportion to the entry it fixes; three of the four fix their entry and move nothing. Two further points bound the concern. The hard rule guards against degenerate output rather than grading accuracy: its triggers, a flat forecast, a forecast targeting the wrong entity, or a recommendation contradicting the analyst’s own numbers, do not include sparse coverage, so the near-empty but well-formed forecasts at the bottom of the observed quality range (0.03, Table 2) are penalized through the coverage metric and rubric-band placement rather than the cap, and the rule does not bind over the observed

range (0.03 to 0.47). And the per-claim level receives only the claim and its cited evidence, so half of the grading protocol is independent of the scorecard by construction.

D Prompt structure

Every system prompt is assembled from three parts: a shared preamble, the condition preamble, and the stage task prompt. This section summarises their operative content; the verbatim text ships with the released harness.

Shared preamble, identical in every stage and condition. The model is cast as a financial analyst on a private-equity or M&A transaction whose task is to produce accurate, grounded, structured claims. Five standing rules follow. *Reasoning first*: write a short plan before calling tools and reassess after each batch of calls. *Grounding*: every numerical claim must cite a specific cell or a specific document and page, and no number may be fabricated. *Units*: every reported number declares its unit, and uncertainty is stated rather than guessed. *Canonical cells*: when several cells hold the same value, bind to the source-of-truth cell, using number formats, referencing formulas, and named ranges to disambiguate. *Derivations*: any computed number that does not come directly from a cell or page is written in a structured derivation block giving its name, expression, and value, so the arithmetic is checkable.

Condition preambles. Unguided states that no schema, canonical field list, or draft is provided and that the model must discover the deal structure itself. Outline states that the canonical field list and unit conventions are provided and that claims must bind to those names. The critique pass instructs the critic to review the analyst’s draft for errors and the analyst to re-verify every numerical claim against the source data rather than copy the draft.

Unguided stage prompts. Each states the task and nothing else. Stage one, for example, asks the model to extract whatever facts it judges an investment committee would need, deciding for itself what matters and how thoroughly to read, with no list of required fields.

Outline stage additions. Extraction receives the canonical field list (enterprise value, entry EBITDA, entry multiple, sources and uses totals, IRR, MOIC, equity investment, and per-year revenue and EBITDA) and the instruction to bind

each field to its source-of-truth cell. Reconciliation is asked to locate the seller claim corresponding to each workbook field and classify the gap as seller-optimistic, seller-pessimistic, or agreeing within tolerance. Forecasting is asked to predict the masked cells from the workbook’s visible growth assumptions, the trajectory of prior-year actuals, and the model’s own methodology, with the explicit instruction not to invent values from priors but to reason from cells it can actually see. Market and recommendation receive per-stage descriptions of what a complete artifact contains, including the required citation form.

Judge prompts. The per-claim judge receives one claim with the exact upstream items it cites and returns strict JSON with a support label and faithfulness and relevance scores. The per-run judge is instructed as a senior partner on an investment committee, receives the deterministic scorecard and the recommendation text, selects a rubric band and then a score within it, and returns strict JSON with the score, a reliability label, a committee-readiness ruling, and key failures and strengths. Bands, the hard rule, and the schema are given in Appendix C.

Worked examples inside the released stage prompts reference specific deal packages and are therefore omitted here and sanitized in the release.

E Inter-judge agreement: four judges treated as equals

No single model should be privileged as the arbiter of quality, so every run of every condition, including every ablation run, is graded by all four frontier models as judges (Opus 4.7, Fable 5, GPT-5.6 Sol, and Gemini 3.6 Flash) and we analyse them symmetrically: none is a reference against which the others are measured. **The judges agree on ordering.** Pairwise Pearson correlation between judges over the 21 graded model-condition cells averages 0.89 and never falls below 0.83; rank agreement gives a mean Kendall τ of 0.70 and a four-judge concordance of Kendall’s $W = 0.88$ (Table 6, Figure 3). At the level of individual (deal, model, condition) gradings, every pair of judges correlates at $r = 0.82$ – 0.86 over the 206 complete four-judge gradings; a single judge carries an intraclass correlation of $ICC(2, 1) = 0.82$, and the four-judge panel mean reaches $ICC(2, 4) = 0.95$, which is the quantitative reason the paper reports the panel rather than

Figure 3: Pairwise judge agreement: Pearson r above the diagonal, Kendall τ below.

any one grader (Table 7). Agreement on the ablation grid is the same ($\text{ICC}(2, 4) = 0.95$ over 280 gradings). Agreement on the categorical rulings is moderate, as expected for thresholded labels (Fleiss’ $\kappa = 0.58$ on the reliability label, 0.55 on committee-readiness). **The judges disagree on level, not on placement.** The mean assigned score over the cells runs from 51.1 for Opus 4.7, the harshest judge, to 56.3 for Gemini 3.6 Flash, the most lenient, which also separates runs most widely (30.6 points against GPT-5.6 Sol’s 22.4; Table 8, Figure 6). Plotted against the panel mean, every judge tracks the panel at $r \geq 0.96$ with a level offset of a few points (Figure 4); the offset is that judge’s mean signed deviation from the panel, negative for a harsher grader and positive for a more lenient one. Model rank under each judge moves by at most one mid-field position. A benchmark keyed to an absolute committee-readiness threshold would have to fix and calibrate a judge; one that compares relative capability, as ours does, is safe under any of the four. **No judge favours its own family.** The own-family delta, a judge’s mean score on analysts from its own model family minus the other judges’ mean on those same runs, is small for every judge and negative for Opus 4.7 (−2.6): Opus is *harsher* on Anthropic analysts than the panel is. Fable 5, Gemini 3.6 Flash, and GPT-5.6 Sol sit at +1.1, +2.2, and +2.1 (Table 9, Figure 5). The self-enhancement bias that model-based judging risks (Zheng et al., 2023) does not appear here in any direction. Two consequences for how the paper is read. First, because ordering is judge-invariant and the panel mean is the most reliable available instrument, the body reports the four-judge panel mean throughout, alongside the per-judge columns of the main table; the per-judge ablation grids are in Appendix F. Second, absolute scores are judge-relative and are never compared across judges: a memo is not “62 good,” it is “62 under Opus, 67 under Gemini 3.6 Flash,” and every delta the paper quotes is a within-scale comparison.

F Every result under every judge

The body reports the four-judge panel mean; this section gives the same results judge by judge so

	Opus	Fable	Gem 3.6	Sol
Opus	1.00	0.84	0.94	0.86
Fable	0.66	1.00	0.90	0.86
Gem 3.6	0.81	0.72	1.00	0.93
Sol	0.70	0.63	0.76	1.00

Table 6: Pairwise agreement between judges over the 21 graded model-condition cells: Pearson r above the diagonal, Kendall τ below. Mean pairwise $r = 0.89$ (minimum 0.83); mean $\tau = 0.70$; Kendall’s $W = 0.88$.

Statistic	Value	Basis
Mean pairwise Pearson r	0.89	21 cells
Mean pairwise Kendall τ	0.70	21 cells
Kendall’s W (concordance)	0.88	21 cells
$\text{ICC}(2, 1)$, single judge	0.82	206 gradings
$\text{ICC}(2, 4)$, panel mean	0.95	206 gradings
$\text{ICC}(2, 4)$, ablation grid	0.95	280 gradings
Fleiss’ κ , reliability label	0.58	206 runs
Fleiss’ κ , committee-readiness	0.55	206 runs

Table 7: Inter-rater reliability of the four-judge panel, on the holistic score and on the judges’ categorical rulings.

any claim can be checked against a single grader. Table 10 repeats the knockout grid under each judge: the ordering, in which forecast removal costs at least as much as reconciliation removal, holds for all four models under all four judges. Table 12 and Figure 7 repeat the repair chains: the forecast rung lifts the score by 17 to 27 points in all sixteen judge-model cells, so the conclusion of Section 5 does not depend on the grader. The same grid carries the dissociation established in Appendix C: every rung perfects its own score-card entry, and only the forecast rung moves the score. Extraction and forecast repairs supply exact ground truth, whereas reconciliation and market repairs are best-effort approximations, so the conclusions of Section 5 rest on the two exact levels. Table 11 gives the pooled stage-metric regression quoted in Section 5, and Table 13 re-fits it under each judge, with forecast quality the dominant coefficient in every fit. The pooled fit treats observations that share a deal or a model as independent; re-fitting with deal and model fixed effects absorbed leaves forecast quality dominant (standardized $\beta = +0.73$, $p < 0.001$), so the association is within-deal and within-model, not an artifact of pooling. Table 14 recomputes the efficiency of the outline runs under each judge; the ranking of the six models is identical under all four.

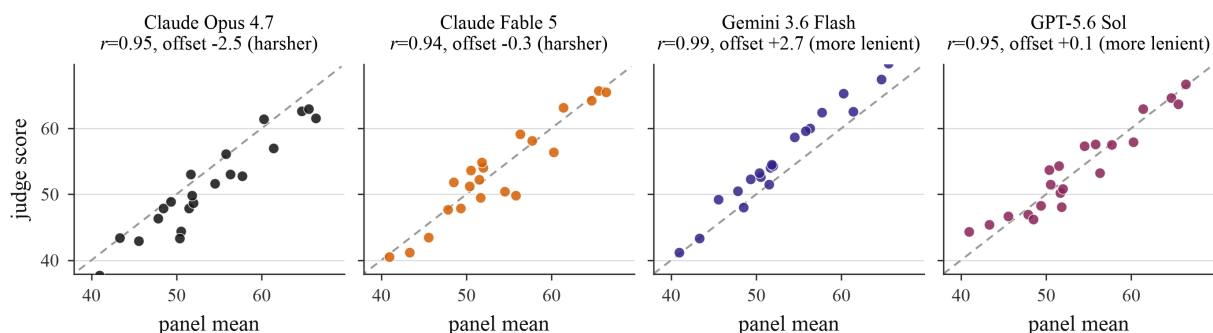


Figure 4: Each judge’s cell score against the four-judge panel mean, with the identity line. The offset in each title is that judge’s mean signed deviation from the panel over the same cells: negative means the judge assigns systematically lower scores than the panel (a harsher grader), positive a more lenient one. The high r with small offsets says the judges agree on where a run sits relative to the others and differ mainly in level.

Judge	Mean	SD	Spread	dev. from panel
Claude Opus 4.7	51.1	7.2	25.2	2.6
Claude Fable 5	53.3	7.4	25.2	1.2
Gemini 3.6 Flash	56.3	8.2	30.6	2.7
GPT-5.6 Sol	53.7	6.8	22.4	1.4

Table 8: Judge leniency and discrimination over the graded model-condition cells: mean assigned score, its standard deviation, the spread between that judge’s best and worst cell, and the mean absolute deviation from the panel mean.

Judge	Own-family delta
Claude Opus 4.7	-2.6
Claude Fable 5	$+1.1$
Gemini 3.6 Flash	$+2.2$
GPT-5.6 Sol	$+2.1$

Table 9: Own-family bias: a judge’s mean score on analyst runs from its own model family minus the other three judges’ mean on the same runs. Positive favours its own family; no delta is large enough to change any ordering in the paper.

G Hybrid critique: a small analyst with a flagship critic

Self-critique raises cost by a factor of two to five without a reliable gain, but it holds the critic and the analyst at the same capability. The hybrid condition therefore pairs a small analyst with a same-family flagship critic: the analyst produces the outline run, the flagship audits the run’s artifacts and citations, and the analyst revises once. Table 15 reports the three pairs, each graded by all four judges, alongside each family’s flagship running the task solo. A flagship critic helps only where the analyst is already strong. For Sonnet 5 with a Fable 5 critic the gain is $+5.9$ on the four-

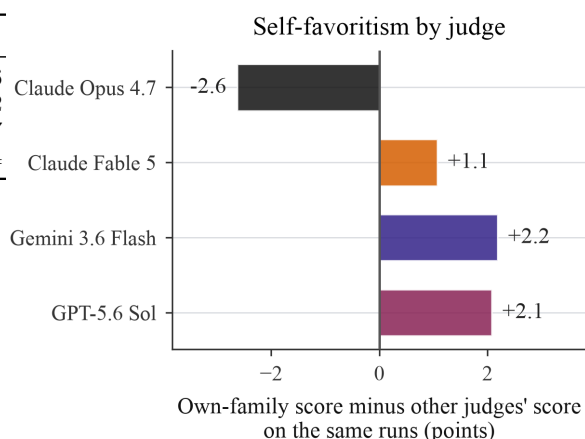


Figure 5: Self-favoritism by judge: the judge’s mean score on analyst runs from its own model family minus the other three judges’ mean on those same runs. Positive means the judge grades its own family above the panel’s view of the same work; negative means it is harsher on its own family.

Figure 6: Distribution of assigned scores by judge over all graded runs. The judges differ in leniency and in how widely they separate runs, which is why absolute scores are judge-relative and only within-scale comparisons are made.

judge panel and positive under every judge, from $+3.0$ under Opus to $+9.4$ under Sol, though at ten deals the paired gain does not survive correction ($p = 0.09$ uncorrected); for GPT-5.6 Luna with a GPT-5.6 Sol critic the panel change is -0.2 and mixed in sign across judges. The third pair sits at the capability floor: the Gemini 3.1 Flash-Lite analyst emits a median of about five forecast cells per deal against forty to fifty for the other two analysts, with correspondingly thin extraction and reconciliation and no errors or truncation, and

Figure 7: The repair chain of Figure 2 under each judge separately, one row per judge and one column per model. The forecast rung is the only one that moves the score, in all sixteen judge-model cells, so the causal claim of Section 5 does not rest on the choice of judge.

Judge	Condition	Opus	Fable	Sol	Flash
Claude Opus 4.7	Outline base	53.0	62.9	61.5	48.7
	–reconcile	50.9	57.3	63.7	45.9
	–forecast	46.0	56.0	56.1	44.1
	–both	43.6	55.4	56.4	39.7
Claude Fable 5	Outline base	59.1	65.7	65.5	54.0
	–reconcile	60.2	61.3	59.7	59.8
	–forecast	52.6	54.1	52.2	52.9
	–both	48.5	47.0	48.4	47.6
Gemini 3.6 Flash	Outline base	60.0	69.8	71.8	54.3
	–reconcile	66.1	63.5	64.2	65.4
	–forecast	54.8	55.4	56.0	55.9
	–both	52.7	51.7	55.4	51.8
GPT-5.6 Sol	Outline base	53.2	63.7	66.7	50.8
	–reconcile	57.7	58.2	56.6	58.1
	–forecast	55.9	56.0	56.1	56.0
	–both	51.4	52.7	50.5	53.6
Four-judge panel	Outline base	56.3	65.5	66.4	52.0
	–reconcile	58.7	60.1	61.0	57.3
	–forecast	52.3	55.4	55.1	52.2
	–both	49.0	51.7	52.7	48.2

Table 10: Knockout under every judge; each block is one judge and the last block is the four-judge panel mean reported in the body (Table 4).

Stage metric	Std. β	Univ. r	p
S1 extraction precision	+0.212	+0.423	< 0.001
S2 reconciliation precision	+0.135	+0.259	< 0.001
S3 forecast quality	+0.770	+0.852	< 0.001
S4 grounding	+0.184	+0.306	< 0.001
Risk propagation	+0.013	+0.050	0.380

Table 11: Association of each stage metric with the four-judge panel score, over all model, condition, and deal observations including the knockout and repair runs, with predictors and score standardized, $n = 313$. Forecast quality is the dominant correlate. Table 13 reports the same regression under each judge separately.

its critiqued run gains +2.4 panel points over its solo run, a floor effect carried by the near-empty artifacts rather than by critique. At the capability floor the binding constraint is the volume of structured work the analyst produces for a critic to audit. The comparison that is significant runs the other way: in every family the critiqued analyst remains below the flagship running the task solo (Table 15, Δ column; Holm-corrected paired tests by deal, pooled -11.2 points, $p < 0.001$), so a flagship critic recovers only part of the gap to the flagship itself. The cost asymmetry com-

Judge	Repair level	Opus	Fable	Sol	Flash
Claude Opus 4.7	main (outline)	53.0	62.9	61.5	48.7
	+true extract	54.4	60.8	63.3	55.9
	+true reconcile	53.1	61.9	62.5	58.0
	+true forecast	79.8	85.5	86.4	79.6
	+true market	78.0	84.2	85.3	82.8
Claude Fable 5	main (outline)	59.1	65.7	65.5	54.0
	+true extract	63.4	63.3	62.7	64.5
	+true reconcile	64.7	65.6	66.7	65.6
	+true forecast	85.4	86.1	85.1	85.2
	+true market	83.0	84.4	85.3	84.7
Gemini 3.6 Flash	main (outline)	60.0	69.8	71.8	54.3
	+true extract	68.1	68.7	69.0	69.5
	+true reconcile	67.2	68.4	69.0	69.9
	+true forecast	88.5	88.5	88.2	89.5
	+true market	89.7	89.0	88.7	88.4
GPT-5.6 Sol	main (outline)	53.2	63.7	66.7	50.8
	+true extract	62.6	60.7	60.3	61.3
	+true reconcile	59.3	62.5	62.6	60.5
	+true forecast	81.0	79.9	83.9	82.4
	+true market	80.1	82.1	81.9	81.4
Four-judge panel	main (outline)	56.3	65.5	66.4	52.0
	+true extract	62.1	63.4	63.8	62.8
	+true reconcile	61.1	64.6	65.2	63.5
	+true forecast	83.7	85.0	85.9	84.2
	+true market	82.7	84.9	85.3	84.3

Table 12: Repair under every judge; the forecast rung is the one that moves the score in every judge-model cell.

pounds this: the Sonnet 5 analyst cost \$537 over the ten deals against \$187 for its critic, while Fable 5 ran the same task solo for \$208 at a higher score.

H Provenance composition of the recommendation

Figure 8 decomposes each recommendation by the upstream stage its citations point to, one panel per condition and one bar per model. Under the unguided baseline the recommendation rests almost entirely on extraction and forecast items; under outline, reconciliation and market items enter the memo, which is the mechanism behind the broader evidence base discussed in Section 4.2. In the outline + critique panel the last three bars are the hybrid pairs (small analyst with the same-family flagship critic, Appendix G).

I Reconciliation precision by condition

The unguided baseline emits no gradable reconciliation citations, so reconciliation is gradable only

Stage metric	Opus	Fable	Gem. 3.6	Sol	Panel
S1 extraction precision	+0.26 (+0.46)	+0.16 (+0.37)	+0.23 (+0.42)	+0.17 (+0.37)	+0.21 (+0.42)
S2 reconciliation precision	+0.10 (+0.22)	+0.15 (+0.27)	+0.14 (+0.26)	+0.12 (+0.24)	+0.13 (+0.26)
S3 forecast quality	+0.75 (+0.84)	+0.77 (+0.84)	+0.71 (+0.79)	+0.72 (+0.79)	+0.77 (+0.85)
S4 grounding	+0.16 (+0.27)	+0.20 (+0.32)	+0.17 (+0.29)	+0.18 (+0.29)	+0.18 (+0.31)
Risk propagation	+0.01 (+0.04)	+0.01 (+0.05)	+0.01 (+0.04)	+0.02 (+0.06)	+0.01 (+0.05)

Table 13: The stage-metric regression of Table 11 re-fit under each judge and on the panel; cells are the standardized coefficient with the univariate correlation in parentheses ($n = 313$ – 314). Forecast quality dominates under every judge.

Model	Cost (\$)	Opus	Fable	Gem. 3.6	Sol	Panel
Opus 4.7	95.0	0.45	0.51	0.51	0.46	0.48
GPT-5.5	55.5	0.87	0.97	0.96	0.97	0.94
Gemini 3.1 Pro	26.5	1.22	1.33	1.31	1.38	1.31
Fable 5	207.8	0.27	0.28	0.30	0.27	0.28
GPT-5.6 Sol	28.8	1.87	1.99	2.18	2.02	2.01
Gemini 3.6 Flash	22.0	1.48	1.64	1.65	1.54	1.58

Table 14: Efficiency of the outline runs under each judge and on the panel: that judge’s score weighted by overall accuracy, per US dollar of measured API cost, the efficiency definition of Table 1 applied judge by judge. The Panel column matches the outline-block efficiency row of Table 1. The ranking of the six models is identical under all four judges.

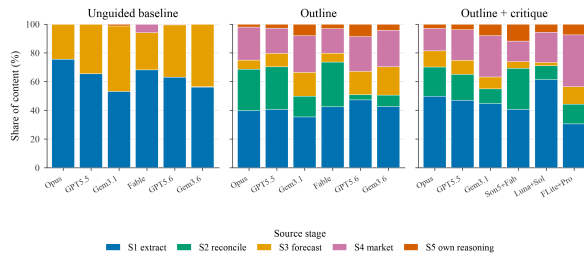


Figure 8: Provenance composition of the recommendation, one panel per condition, one bar per model. The unguided baseline rests on extraction and forecasting; scaffolding brings reconciliation and market evidence into the memo. In the outline + critique panel, the last three bars are the hybrid pairs (small analyst + same-family flagship critic).

under scaffolding. Pooled over (model, deal) runs, precision is 93.4 ± 10.1 percent under outline ($n = 60$ runs) and 88.1 ± 20.1 percent under outline + critique ($n = 58$). The critique pool includes the hybrid analysts’ runs; restricted to the models run solo in all three conditions, the critique figure is 92.5 percent, so the drop reflects the hybrids’ smaller analysts rather than the condition itself.

J Recall against the curated ground truth

Precision grades what the model chooses to assert; recall bounds the complementary risk, that a model asserts little and looks clean. We report recall where the ground truth defines a target set

and volumes where it does not. For extraction, the ground truth pins 113 workbook-critical facts across the ten deals, the figures an analyst must hold to reproduce the valuation, and recall is the share of these facts the model surfaces under outline, matched by the cited cell or by value within the fact’s tolerance on the same sheet. Recall is 48 (Opus 4.7), 49 (GPT-5.5), 28 (G. Pro), 71 (Fable), 50 (Sol), and 45 (GF 3.6) percent. Extraction is therefore high-precision and partial-recall, and the extraction repair of Section 5 bounds the downstream cost of the missed facts at a few panel points. For reconciliation, the curated list pins 30 expected verdicts (23 aligned, 7 misaligned); it exists to define the repair intervention’s true reconciliation, not to enumerate every discrepancy in a live package. Volume rules out the degenerate case of a short, clean finding list: under outline the models emit a median of 14 to 53 reconciliation findings per deal and flag a median of 7 to 37 as misaligned, several times the curated list, and precision is computed over all of them. The risk-propagation denominator is likewise the model’s own flagged set, so the percentages of Table 3 are computed over dozens of findings per run.

K Off-grid forecast predictions

Models frequently predict cells outside the scored canonical grid. These are excluded from the forecast metric, because a cell outside the masked re-

Family	Run	Judge score, by judge				Panel mean	Δ
		Opus 4.7	Fable 5	Gem. 3.6 F	GPT-5.6 Sol		
Sonnet 5	solo (outline)	49.8	54.8	54.5	48.1	51.8	
	+ Fable 5 critic	52.8	58.1	62.4	57.5	57.7	+5.9
	Fable 5 solo	62.9	65.7	69.8	63.7	65.5	-7.8 [‡]
GPT-5.6 Luna	solo (outline)	44.4	53.6	52.6	51.5	50.5	
	+ GPT-5.6 Sol critic	43.3	51.2	53.2	53.7	50.4	-0.2
	GPT-5.6 Sol solo	61.5	65.5	71.8	66.7	66.4	-16.0 [‡]
Gem 3.1 F-Lite	solo (outline)	37.7	40.5	41.2	44.3	40.9	
	+ Gem 3.1 Pro critic	43.4	41.2	43.3	45.4	43.3	+2.4
	Gem 3.1 Pro solo	47.9	52.2	51.5	54.3	51.5	-8.1 [‡]

Table 15: Hybrid critique, all runs scored by the four judges and their panel mean. Each block is one model family: the small analyst run solo under outline, the same analyst revised once after the family’s flagship critic audits its artifacts, and the flagship itself run solo under outline. The score columns are the four *judges*, not the models being run. Δ on the critic row is the panel change from adding the critic (critic – solo); Δ on the flagship row is how far the critiqued analyst still sits from the flagship working alone (critiqued analyst – flagship). [‡]Gap to the flagship significant under Holm-corrected paired *t*-tests by deal. The Gemini pair rests on near-empty analyst runs and is not comparable.

Condition	Off-grid preds	Checked	Acc@5%	Metric	Pair	Diff.	p_{Holm}
Unguided baseline	3332	2929	37.2	Panel	Fable 5 > Gem. 3.6 Flash	+12.2	.001
Outline	1192	1140	46.1	Panel	Fable 5 > Gem. 3.1 Pro	+12.1	.001
Outline + critique	1061	948	37.3	Panel	Fable 5 > Opus 4.7	+10.3	.004
				Panel	GPT-5.5 > Gem. 3.6 Flash	+11.7	.005
				Panel	GPT-5.5 > Gem. 3.1 Pro	+11.7	.006
				Panel	GPT-5.5 > Opus 4.7	+9.8	.018
				Panel	GPT-5.6 Sol > Gem. 3.6 Flash	+8.5	.049
				Acc.	GPT-5.5 > Gem. 3.1 Pro	+19.4	.002
				Acc.	Fable 5 > Gem. 3.1 Pro	+17.8	.003
				Acc.	GPT-5.5 > Gem. 3.6 Flash	+16.9	.011
				Acc.	Fable 5 > Gem. 3.6 Flash	+15.3	.022

Table 16: Off-grid predictions pooled over models and deals: how many were made, how many land on a workbook cell with a checkable value, and the share within five percent. The unguided baseline predicts almost three times as many off-grid cells and is less accurate on them.

gion may have been readable directly, but they are informative as a discipline diagnostic. Table 16 pools them over models and deals: scaffolding cuts off-grid predictions by almost a factor of three and raises accuracy on those that remain.

L Between-model statistics

One observation is a (model, condition, deal) run, pooled over the three conditions with the O+C observations for Fable, Sol, and GF 3.6 taken from the hybrid runs, exactly as Table 1 reports them. The one-way analysis of variance across the six models is significant on the panel score ($F = 7.64$, $p < 0.001$) and on accuracy ($F = 7.00$, $p < 0.001$); Table 17 lists every pairwise comparison that survives the Holm correction. All pairwise comparisons in the paper are Welch *t*-tests (paired *t*-tests by deal where the same deals are compared within a model), Holm-corrected at $\alpha = .05$ within each test family; where several analyses of variance address one question, the cor-

Table 17: Every Holm-significant pairwise comparison between models, pooled over the three conditions (panel score and overall accuracy). No other pair survives correction.

rection is applied to their *p*-values as a family before any post hoc is read.

M Cost and token usage

Costs are measured from per-call API usage ledgers written during the runs, in US dollars over the ten deals. Under outline, the full ten-deal run cost \$95.0 for Opus 4.7, \$55.5 for GPT-5.5, \$26.5 for Gemini 3.1 Pro, \$207.8 for Fable 5, \$28.8 for GPT-5.6 Sol, and \$22.0 for Gemini 3.6 Flash, on input volumes of 15.8, 7.7, 12.1, 12.1, 8.9, and 23.5 million tokens. The unguided runs cost \$63.6, \$46.7, \$23.7, \$146.4, \$20.5, and \$17.2 in the same order; the first three predate the ledger and are reconstructed from archived per-stage transcripts priced at the same per-token rates, a reconstruction that matches the ledger within

rounding for GPT-5.5 and Gemini 3.1 Pro and within six percent for Opus 4.7, whose transcripts do not separate cache reads. The critique pass raises cost by a factor of 1.9 (Opus 4.7, to \$183.6) to 5.2 (Gemini 3.1 Pro, to \$137.0), with GPT-5.5 at \$161.3. In the hybrid pairs, the Sonnet 5 analyst cost \$537 with a \$187 Fable 5 critic, GPT-5.6 Luna \$2 with a \$17 Sol critic, and Gemini 3.1 Flash-Lite \$13 with a \$4 Pro critic. Grading by the four-judge panel added \$523 over the Opus 4.7, GPT-5.5, Gemini 3.1 Pro, and hybrid grids and \$163 over the Fable, Sol, and GF 3.6 grids, dominated by the most expensive judge (Fable 5, roughly \$2.4 per outline memo against \$1.0 for Opus 4.7 and under \$0.8 for the other two).