# The Aiera Research Score: Scoring Financial LLMs on Facts, Grounding, and Depth

Bryan Healey  Kieran Black
Aiera

August 2, 2026

**Abstract**

Many leaderboards for language models aggregate a grab-bag of capability tasks, but for institutional research the question that matters is narrower: *how well does a model match the analytical quality of a research analyst, grounded in real data?* We have therefore redesigned (again) the Aiera leaderboard around that single question, dropping the general-capability tasks at which nearly all modern models already excel. The headline **Research Score** now combines three equal weight axes of research quality: **Facts** (does the answer get the substance right and complete?), **Grounding** (are its claims cited and defensible?), and **Depth** (is the analysis of the caliber an institutional professional would produce?). We decided to remove capability on two grounds: it is largely *orthogonal* to research quality and, on inspection, *artifact-prone*. We introduce a purpose-built **Depth** axis, validate it for discriminant validity ($r$=0.73 against coverage) and resistance to verbosity bias (within-model length correlation +0.21), and show it independently recovers expert practitioner judgment that accuracy metrics miss. Across 32 models the board is led by Claude-Opus-5, Kimi-K3, and Claude-Fable-5; notably, models that top the factual axes but write shallow, undifferentiated analysis (the GPT-5.6 cohort) rank below models that pair completeness with genuine depth.

## 1 Introduction

Language models are increasingly the reasoning layer over financial data, and the leaderboards practitioners consult were built to measure *general* ability. For a financial-research provider and its clients, the operative question is sharper: connected to real research, transcripts, and filings through a standard interface, **which models actually produce the complete, sourced, insightful analysis an institutional professional would be willing to sign their name to?**

An earlier version of this board answered a version of that question by weighting a proprietary research task at 60% and three established capability tasks (Q&A, summarization, sentiment) at the remaining 40%. This paper takes the design to its logical conclusion: **we drop capability entirely and make the board purely about research**. Two observations drive the change. First, modern frontier models have largely *saturated* the classic capability tasks—summarization scores cluster in a narrow band across the top tier—so those axes add little discriminating signal. Second, and more corrosive, capability metrics are *artifact-prone*: rigid exact-match scoring penalizes correct answers that are merely formatted unexpectedly, injecting noise that is easily mistaken for a gap (§6).

Instead, we decompose research quality into three axes that mirror how a strong analyst is actually judged—**Facts**, **Grounding**, and **Depth**—and weight them equally. The Depth axis is new, and is the crux: it captures whether a complete, grounded answer is *good*, not merely thorough.

**Contributions.**

1. A **pure-research leaderboard** for financial LLMs—a single coherent construct scored across three axes of grounded-analysis quality (§2–§3).
2. A purpose-built **Depth (Insight) metric**, engineered against verbosity bias and **validated** for discriminant validity and length-resistance rather than asserted (§5).
3. An **empirical case for removing capability**, including a documented scoring artifact that a capability axis mistook for a model weakness (§6).
4. **Results and analysis** for 32 models, including the finding that factual completeness and analytical depth diverge sharply at the top of the board (§7).

## 2 Related Work

**LLM leaderboards and evaluation frameworks.** Holistic, multi-task evaluation was popularized by HELM [1] and the Open LLM Leaderboard [2], typically built on the `lm-evaluation-harness` [3]. These rank models on broad capability; our board instead centers a *single* domain axis—research quality—and decomposes it rather than diluting it with general tasks.

**Financial NLP benchmarks.** Numerical reasoning over financial documents is established by FinQA [4] and ConvFinQA [5]; broader suites include FinanceBench [6], with domain pretraining by BloombergGPT [7] and sentiment tracing to FinBERT [8]. These target isolated capabilities; we instead score end-to-end grounded research answering, and argue the isolated-capability metrics are the ones worth *removing* from a research board.

**Retrieval, tools, and the Model Context Protocol.** Connecting models to external knowledge spans retrieval-augmented generation [9] and tool/agent use—ReAct [10], Toolformer [11], Gorilla [12]. The Model Context Protocol [13] standardizes the tool interface our evaluation exercises, letting every model answer against an identical data server.

**Automatic metrics and LLM-as-judge.** Beyond n-gram and embedding metrics such as BERTScore [14], open-ended answers are increasingly graded by LLM-as-judge [15, 16], with atomic fact-level scoring in the spirit of FActScore [17]. This paradigm has known failure modes—self-preference [18] and a verbosity bias [15]. Our contribution is to treat these as *engineering constraints*: we design the Depth judge explicitly against verbosity and report the falsifiable validity checks (§5), rather than assuming the judge is unbiased.

## 3 The three axes of research quality

Every axis is evaluated with the model *connected to the standardized financial-data server from Aiera via the Model Context Protocol*, answering through a client-side tool-use loop, with best-of-3

answer generation. All judged axes use a single fixed, cross-family panel of three held-out frontier models, graded atomically. Each sub-metric is scaled 0–100.

| Axis | What it measures / how |
|---|---|
| **Facts** | Did the answer surface the right substance, correctly and completely? |
| **Grounding** | Are the answer's claims tied to explicit sources, or asserted bare? |
| **Depth** | Is a complete, grounded answer actually *good* and of the caliber of a professional? |

**Facts: keyfacts and coverage.** *Keyfacts* scores entailment of specific gold facts in a model's grounded answer to proprietary research questions—accuracy on the facts that matter. *Coverage* scores a complementary set of complex, multi-step analyst prompts, each carrying a weighted rubric of required elements (all named entities, each requested dimension, the required verdict, no fabrication); the panel grades every element satisfied / partial / missing, and coverage is the weighted credit. Keyfacts asks *were the facts right*; coverage asks *was the whole job done*.

**Grounding: faithfulness.** Per answer, the panel estimates the share of material factual claims carrying an explicit source citation (most / some / few / none). This is the trust axis: it rewards answers a professional could defend and penalizes confident, uncited assertion.

**Depth: insight.** A judge panel, prompted in the tone and style of a senior institutional financial professional, rates each answer's professional caliber (exemplary / solid / shallow / weak). The prompt is *explicitly engineered against verbosity*: it directs that length is not a scoring vector, that a concise answer with a clear view outranks a long exhaustive one, and instructs the judge to weigh *insight density, not volume*.

## 4 The Aiera Research Score

Coverage and keyfacts are strongly correlated across the 32 models ($r$=0.91): they measure the same underlying "did it get the substance" from two critical angles. Weighting them separately would double-count, so we combine them into a single **Facts** axis. The score is then produced by an equal-weight mean of the three axes:

$$\text{Facts} = \tfrac{1}{2}\,\text{keyfacts} + \tfrac{1}{2}\,\text{coverage}, \qquad \text{Research} = \tfrac{1}{3}(\text{Facts} + \text{Faithfulness} + \text{Insight}).$$

**Equal weights, deliberately.** We considered a depth-weighted alternative (40% Depth / 30% Grounding / 30% Facts) that sharpens the editorial claim that depth is what distinguishes strong analysts. We ship equal weights instead: no single axis—least of all the novel Depth axis—should do more than a third of the work, which is the more defensible position. The choice is also low-stakes at the top: the leading model is first under both weightings.

## 5 Validating the Depth axis

Depth is the novel and softest axis, so it carries the burden of proof. We report two falsifiable checks over all 4,767 scored answers, plus an external corroboration.

**Discriminant validity.** If Depth merely relabeled completeness it would be redundant. The correlation of Insight with Coverage is $r$=0.73—related, as expected (a complete answer has more room to be insightful), but clearly distinct. Insight is not a proxy for coverage.

**Verbosity resistance.** A naive quality judge rewards length. The *cross-model* correlation of Insight with answer length is 0.61, but this is a capability confound—stronger models write longer *and* are more insightful. The decisive test is *within-model*: does a given model's own longer answers score higher? The average within-model length correlation is **+0.21**, and *negative* for some models—i.e. a model cannot meaningfully raise its Depth simply by padding. Strengthening the anti-verbosity language in the judge prompt did not move the cross-model figure, confirming the residual correlation lives in the data, not the judge's reaction.

**Blind expert corroboration.** Before this metric existed, domain experts formed hands-on impressions from a handful of complex prompts: certain models produced impressive analysis while others, though thorough, read as generic. The Depth metric—built with no access to those impressions—reproduced that ordering. Convergence of an instrument engineered against the obvious bias with independent expert judgment is strong evidence the axis captures real analytical quality that the factual axes miss.

## 6 Why we removed capability

The prior board reserved 40% of its weight for capability tasks (Q&A, summarization, and financial sentiment). We remove them, for two reasons.

**Orthogonality.** On the capability axis that is not artifact-prone (summarization), top-tier models cluster tightly—the metric barely discriminates. Where capability *does* vary, it tracks model breadth, not research quality, and pulls the ranking away from analytical performance.

**Artifacts.** Capability metrics are exact-match and format-sensitive. One strong open-weight model scored **10**% on sentiment—yet its summaries were coherent and full-length, its confusion matrix showed near-zero *actual* misclassifications (roughly 88% of its outputs matched *no* valid label token), and a separate run of the same model scored 100%. The 10% was a **label-parsing artifact**, not a capability: the metric was measuring format compliance, and it was materially depressing a strong model's rank. A pure-research score removes this class of noise entirely. This is not an argument that capability is uninteresting—only that it does not belong in, and actively muddies, a *research* ranking and the purpose of our leaderboard.

## 7 Results

### 7.1 The board (top of 32 models)

| # | Model | Research | Facts | Grounding | Depth |
|---|---|---|---|---|---|
| 1 | Claude-Opus-5 | **87.0** | 78.5 | 86 | 96 |
| 2 | Kimi-K3 | 86.6 | 74.0 | 96 | 90 |
| 3 | Claude-Fable-5 | 86.1 | 73.7 | 97 | 88 |
| 4 | GPT-5.6-sol | 84.1 | 77.2 | 94 | 81 |
| 5 | Claude-Opus-4.7 | 83.6 | 69.7 | 96 | 85 |
| 6 | Claude-Opus-4.8 | 82.6 | 69.7 | 94 | 84 |
| 7 | MiniMax-M3 | 81.5 | 74.3 | 93 | 77 |
| 8 | GPT-5.6-terra | 80.8 | 70.6 | 95 | 77 |
| 9 | GPT-5.5 | 80.7 | 76.1 | 95 | 71 |
| 10 | DeepSeek-V4-Pro | 80.1 | 71.1 | 92 | 77 |

### 7.2 Observations

- **Facts and Depth diverge at the top.** The GPT-5.6 cohort leads on the factual axes (GPT-5.6-sol: Facts 77, third-highest) but trails on Depth (81, 77, 77), and drops below Claude models that pair completeness with high insight. The prior blended score, which had no depth axis, ranked GPT-5.6-sol first; here it is fourth. Completeness is not caliber.
- **Grounding is the sharpest separator among leaders.** Coverage and keyfacts saturate near the top, so the frontier is separated less by *what* models say than by whether they cite it: the top tier grounds at 92–97, while several otherwise-strong models cover well but ground weakly and fall accordingly.
- **The top is an honest cluster.** The leading three (Claude-Opus-5, Kimi-K3, Claude-Fable-5) sit within one point; we report the ordering but do not over-claim a decisive #1.
- **Depth harshly separates the tail.** Weak models that produce filler or generic prose collapse on Depth (single-digit scores) even when their coverage is non-trivial—the axis is doing real discriminating work the factual metrics alone would not.

## 8 Methodology & Reproducibility

Answers are generated by each model connected to the data server through a client-side tool-use loop, best-of-3, with the highest-coverage attempt retained. All judged axes use a fixed, cross-family panel of three held-out frontier models, graded atomically (per rubric element for coverage; per answer for grounding and depth). The full run—32 models over 149 complex prompts, three attempts each—comprises roughly 14,300 answers and over 400,000 judge decisions; every answer, judge vote, and per-element score is archived, and the per-model onboarding recipe and canonical scoring function are version-controlled. One transient data-server outage during generation was detected via a throttle-signature scan and fully backfilled; the released data is clean.

## 9 Limitations

- **Depth is LLM-judged and subjective**, and now co-heaviest. We mitigate with a cross-family panel and the validated anti-verbosity design, but it is the softest measurement on the board and should be weighed as such.
- **Grounding measures citation *presence*, not source-verified correctness.** A model that cites lavishly but inaccurately is not yet penalized; the planned upgrade persists tool outputs and checks each claim against its cited source.
- **Ceiling effects** on coverage and insight compress discrimination among the strongest models, making top-of-board margins small.
- **Single best-of-3 run**; we do not yet report run-to-run variance.
- **Proprietary inputs are withheld.** The research questions, complex prompts, rubrics, and gold facts are not published (licensing and material-nonpublic-information constraints); we release scores and aggregate statistics only.

## 10 Conclusion

By dropping general capability and scoring only research quality—across Facts, Grounding, and Depth—the Aiera Research Score ranks financial LLMs by what actually matters for analyst work: producing complete, sourced, genuinely insightful analysis from professional data. The redesign surfaces a finding the prior blended board could not: that factual completeness and analytical depth *diverge*, and that some models which top the accuracy axes produce thorough-but-shallow work. A naive version of this study—one without a validated depth axis—would have concluded those models were the best, and been wrong in exactly the way a practitioner already suspected.

## References


[1] P. Liang, R. Bommasani, T. Lee, et al. "Holistic Evaluation of Language Models (HELM)." *Transactions on Machine Learning Research*, 2023. arXiv:2211.09110.

[2] E. Beeching, C. Fourrier, N. Habib, et al. "Open LLM Leaderboard." Hugging Face, 2023. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

[3] L. Gao, J. Tow, B. Abbasi, et al. "A Framework for Few-Shot Language Model Evaluation (lm-evaluation-harness)." Zenodo, 2023.

[4] Z. Chen, W. Chen, C. Smiley, et al. "FinQA: A Dataset of Numerical Reasoning over Financial Data." *EMNLP*, 2021. arXiv:2109.00122.

[5] Z. Chen, S. Li, C. Smiley, et al. "ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering." *EMNLP*, 2022. arXiv:2210.03849.

[6] P. Islam, A. Kannappan, D. Kiela, et al. "FinanceBench: A New Benchmark for Financial Question Answering." arXiv:2311.11944, 2023.

[7] S. Wu, O. Irsoy, S. Lu, et al. "BloombergGPT: A Large Language Model for Finance." arXiv:2303.17564, 2023.

[8] D. Araci. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." arXiv:1908.10063, 2019.

[9] P. Lewis, E. Perez, A. Piktus, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *NeurIPS*, 2020. arXiv:2005.11401.

[10] S. Yao, J. Zhao, D. Yu, et al. "ReAct: Synergizing Reasoning and Acting in Language Models." *ICLR*, 2023. arXiv:2210.03629.

[11] T. Schick, J. Dwivedi-Yu, R. Dessì, et al. "Toolformer: Language Models Can Teach Themselves to Use Tools." *NeurIPS*, 2023. arXiv:2302.04761.

[12] S. G. Patil, T. Zhang, X. Wang, J. E. Gonzalez. "Gorilla: Large Language Model Connected with Massive APIs." arXiv:2305.15334, 2023.

[13] Anthropic. "Introducing the Model Context Protocol." 2024. https://www.anthropic.com/news/model-context-protocol.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi. "BERTScore: Evaluating Text Generation with BERT." *ICLR*, 2020. arXiv:1904.09675.

[15] L. Zheng, W.-L. Chiang, Y. Sheng, et al. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." *NeurIPS*, 2023. arXiv:2306.05685.

[16] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu. "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment." *EMNLP*, 2023. arXiv:2303.16634.

[17] S. Min, K. Krishna, X. Lyu, et al. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation." *EMNLP*, 2023. arXiv:2305.14251.

[18] A. Panickssery, S. R. Bowman, S. Feng. "LLM Evaluators Recognize and Favor Their Own Generations." *NeurIPS*, 2024. arXiv:2404.13076.