The Aiera Leaderboard: A Research-Centric Benchmark for Financial LLMs

Bryan Healey Kieran Black Jacqueline Garrahan
Aiera

June 8, 2026

Abstract

General LLM leaderboards rank models on broad capability, but for institutional financial research the question that matters is narrower and harder: *how well does a model answer real, analyst-grade research questions when it is connected to professional-grade data?* We redesign the Aiera Leaderboard around that question. The headline **Aiera Score** is a weighted composite in which **Research**—key-facts accuracy on proprietary research questions answered with the model connected to Aiera's Model Context Protocol (MCP) server—carries 60% of the weight, with three established financial capability tasks (Q&A, Summary, Sentiment) making up the remaining 40%. The board is **research-gated**: a model must have a Research score to be listed. Across the 16 models currently qualified, the ordering is led by Anthropic's Claude-Opus family (4.6, 4.8, 4.7) and OpenAI's GPT-5.5/5.2, with the strongest open-weight models (Kimi-K2.6, gpt-oss-120B, GLM-4.6) forming a competitive middle tier. We find that research performance is only weakly correlated with general capability (Spearman $\rho \approx 0.34$) and is far more discriminating (research scores span 0.3–46.9 vs. 46.0–79.7 for capability), which is precisely why a research-weighted, research-gated board ranks LLMs on financial tasks differently—and, we argue, more usefully—than a general-capability board.

1 Introduction

Large language models are increasingly used as the reasoning layer over financial data, but the leaderboards practitioners consult were built to measure *general* ability. For a financial-data provider and its users, the operative question is different: connected to real research, transcripts, and filings through a standard interface, **which models actually produce correct, sourced answers to the questions an analyst would ask?**

We answer this by making **research the center of the leaderboard**, not an afterthought. The Aiera Leaderboard's headline metric, the **Aiera Score**, weights a model's research performance (answering institutional-grade research questions with Aiera-supplied proprietary data) at 60%, with the three capability tasks filling the rest.

We measure research as an absolute *level*—how well a model answers research questions when connected to the Aiera MCP server—rather than as the *improvement* it gains over a no-data baseline without access to Aiera data. While the improvement a model derives from the data connection is a natural way to quantify the value of that connection, the **level** is the right basis for ranking model quality, and is what the board currently reports.

Contributions.

1. A **research-gated, research-weighted leaderboard** for financial LLMs (§2–3).
2. A weighting **rationale grounded in a sensitivity analysis** rather than intuition (§4).
3. **Results and analysis** for 16 models, including the finding that research is a distinct, more-discriminating axis than general capability (§5).

2 Related Work

LLM leaderboards and evaluation frameworks. Holistic, multi-task evaluation of language models has been popularized by HELM [1] and the Open LLM Leaderboard [2], typically built on the `lm-evaluation-harness` [3]. These rank models on broad, general-purpose capability; our board instead centers a single domain-critical axis—research—and gates on it.

Financial NLP benchmarks. Numerical reasoning over financial documents is established by FinQA [4] and ConvFinQA [5]; broader financial suites include PIXIU/FLARE [6] and FinanceBench [7], and domain-specific pretraining by BloombergGPT [8]. Financial sentiment analysis traces to FinBERT [9]. Our Q&A, Summary, and Sentiment tasks follow this tradition; the Research task is, to our knowledge, novel in scoring grounded, data-connected research answering as the headline metric.

Retrieval, tools, and the Model Context Protocol. Connecting models to external knowledge spans retrieval-augmented generation [10] and tool/agent use: ReAct [11], Toolformer [12], and Gorilla [13]. The Model Context Protocol [14] standardizes the tool interface our Research task exercises, and lets us evaluate every model against an identical data server.

Automatic metrics and LLM-as-judge. Generation quality is measured with BLEU [15], ROUGE [16], and BERTScore [17] (the last is our Summary metric). For open-ended answers we adopt the LLM-as-judge paradigm [18], mitigating its known biases via a held-out, cross-family grader and atomic, fact-level scoring in the spirit of FActScore [19].

3 Tasks and Metrics

The Aiera Score combines four tasks, each scored 0–100.

Task	Dataset / source	Metric	Weight
Research	Aiera research eval (proprietary, 150-item)	key-facts entailment, model + Aiera MCP	**0.60**
Q&A	`Aiera/finqa-verified`	exact-match (manual)	0.24
Summary	`Aiera/aiera-ect-sum`	BERTScore-F1	0.10
Sentiment	`Aiera/aiera-transcript-sentiment`	accuracy	0.06

Research is the model's *absolute* performance with Aiera: for each proprietary research question, a held-out judge scores the fraction of required key facts the answer supports [19, 18]; the task

score is the mean over ∼116 proprietary items. We use the absolute with-Aiera *level* (not the baseline→MCP *lift*) because the leaderboard ranks model quality, not the size of Aiera's contribution. The three capability tasks are drawn from the established Aiera benchmark suite, unchanged in definition: Q&A follows the numerical-reasoning format of FinQA [4], Summary is scored with BERTScore-F1 [17], and Sentiment follows financial-sentiment practice [9].

Capability ordering. Within the 40% capability share, the three tasks are weighted in order of importance to financial work: Q&A (calculation-grounded reasoning) > Summary > Sentiment, i.e. 0.24 / 0.10 / 0.06—a 60/25/15 split of the non-research weight.

4 The Aiera Score

The score is the weighted mean of the task scores actually present for a model, normalized by the sum of those weights:

$$\text{AieraScore}(m) \;=\; \frac{\sum_t w_t \cdot \text{score}_t(m)}{\sum_t w_t}, \qquad t \text{ over tasks with a score for } m.$$

Raw, not normalized. We score on the raw 0–100 metrics rather than min-max normalizing each metric across models. §4 shows the ranking is invariant to this choice, so we keep the simpler, more interpretable form.

5 Why These Weights (Sensitivity Analysis)

The 60% research weight is a deliberate statement—"research matters most"—but we verified it is also well-behaved.

- **The top of the board is weight-insensitive.** Sweeping the research weight from 40% to 70% leaves the leaders unchanged: Claude-Opus-4.6, Claude-Opus-4.8, Claude-Opus-4.7, and GPT-5.5 occupy the top four at every setting (4.7 and GPT-5.5 swap 3rd/4th between 40% and 60%). These models lead because they are strong on research *and* capability; no defensible weight dislodges them.
- **The research weight governs the mid-pack.** Its real effect is how much an open-weight model's Aiera-research strength should outweigh general-capability gaps. At 60%, research-strong open models (Kimi-K2.6, GLM-4.6) rank above capable-but-research-weaker ones (Llama-4, Qwen3)—the intended, research-centric behavior.
- **Normalization-invariant.** Min-max normalizing each metric before weighting yields the same top-8 ordering, so we use raw scores.
- **Research is a distinct axis.** Across the 16 models, research correlates only weakly with capability-only score (Spearman $\rho \approx 0.34$). Research is also far more discriminating: scores span **0.3**–**46.9** (mean 19.6) versus **46.0**–**79.7** for capability. In practice **research separates the leaders while capability orders the pack**—5 of 16 models exceed 30 on research, while 7 of 16 fall below 10. A research-weighted score therefore ranks financial LLMs differently from a capability board by design, not by accident.

6 Results

6.1 The board (16 models)

#	Model	Aiera	Research	Q&A	Summary	Sentiment
1	Claude-Opus-4.6	**59.9**	46.9	86	67	74
2	Claude-Opus-4.8	57.3	46.5	76	66	76
3	Claude-Opus-4.7	55.3	42.9	77	65	77
4	GPT-5.5	53.7	36.4	87	67	72
5	GPT-5.2	47.3	28.8	81	63	71
6	Kimi-K2.6	41.1	31.6	86	9	11
7	gpt-oss-120B	38.8	17.7	84	43	64
8	Gemini-2.5-Flash	37.3	12.5	78	69	71
9	GLM-4.6	37.3	22.6	71	40	43
10	Llama-4-Maverick	33.8	6.2	81	67	64
11	Mistral-Large	33.1	7.6	74	64	74
12	Qwen3-235B	33.0	2.6	84	68	77
13	Qwen3-32B	25.8	3.5	73	39	40
14	Llama-4-Scout	24.5	0.3	57	68	64
15	Gemma-3-27B	22.6	0.6	63	25	77
16	LFM2-24B-A2B	22.5	6.9	34	60	70

6.2 Observations

- **An Anthropic sweep at the top, on research.** Claude-Opus-4.6/4.8/4.7 take the top three almost entirely on research strength (42.9–46.9, well clear of the field), with GPT-5.5 and GPT-5.2 next. The proprietary frontier models are simply best at turning Aiera's research data into correct, sourced answers.
- **Research separates; capability orders.** The leaders' margins come from research, where they score 2–4× the pack. Below them, several open and proprietary models cluster on capability while scoring near zero on research (e.g. Qwen3-235B: Q&A 84 but research 2.6)—consistent with the low research/capability correlation in §4.
- **Per-task outliers reward inspection, not just the composite.** Two are worth noting. *Kimi-K2.6* posts elite Q&A (86) and research (31.6) but near-zero Summary (9) and Sentiment (11)—it reliably emits empty or malformed output, which the per-task columns expose even as its research/Q&A strength lifts it to #6. *Gemini-2.5-Flash* is the inverse: even capability across the board but modest research (12.5), landing it at #8, just above open-weights.
- **Edge-optimized models rank low, as expected.** Liquid's LFM2-24B-A2B—a mixture-of-experts model with only ∼2B active parameters, built for on-device deployment—posts solid Summary (60) and Sentiment (70) but weak Q&A (34) and Research (6.9), placing it near the bottom.

7 Methodology & Reproducibility

- **Capability tasks** are run via the `lm-evaluation-harness` [3] (pinned); the four datasets are public on Hugging Face. Scores are published per model to `Aiera/aiera-leaderboard-results`.
- **Research** is run via Aiera's MCP research-evaluation harness against a private 150-item research benchmark (scores published, inputs withheld).

8 Limitations

- **Private research set.** Research inputs and gold answers are not published (licensing); only scores are. This limits external reproduction of the research task specifically.
- **Coverage.** The board lists only research-evaluated models (16 today). Notable current models remain unlisted until run through the research eval.
- **Capability serving configurations differ by model**, which can affect cross-model capability comparability at the margin; the research task, where models reach an identical MCP server endpoint, is more directly comparable.
- **Judge-based metrics.** Research and (implicitly) some capability metrics rely on an LLM judge; we mitigate with a held-out, cross-family judge and atomic key-facts scoring.

9 Conclusion

By gating on research and weighting it most heavily, the Aiera Leaderboard ranks LLMs by what actually matters for analyst work: producing correct, sourced answers from professional data. The result is a board that both reflects Aiera's research-centric offering and surfaces a finding a general-capability leaderboard would miss—that research ability is a distinct, more-discriminating axis, on which today's frontier models lead but a handful of open-weight models are competitive.

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, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, T. Wolf. "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] Q. Xie, W. Han, X. Zhang, et al. "PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance." *NeurIPS Datasets and Benchmarks*, 2023. arXiv:2306.05443.

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

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

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

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

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

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

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

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

[15] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation." *ACL*, 2002.

[16] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries." *ACL Workshop on Text Summarization*, 2004.

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

[18] 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.

[19] 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.