The Aiera Lift: Measuring the Value of a Financial Data MCP Server Across Foundation Models

Bryan Healey Kieran Black
Aiera

June 16, 2026

Abstract

Frontier language models are increasingly used for financial research, yet their knowledge is frozen at training time and the open web does not surface the proprietary research that professional analysts rely on. We quantify how much a single intervention—connecting a model to Aiera's Model Context Protocol (MCP) server—improves research-question answering, and whether that improvement holds across model providers and capability tiers. We construct a **private, source-grounded benchmark of nearly 500 analyst-style questions**, each with gold facts traceable to a verbatim Aiera source and phrased so as not to reveal the source. We evaluate each model in two conditions—*baseline* (the model with its own native web search) and *MCP* (the same model connected to Aiera)—and score answers by automatic **key-facts entailment** plus a held-out LLM judge. Across **13 models** (a 150-question eval, 3,892 graded answers) we find that connecting a model to Aiera produces large gains on the ∼78% of questions the open web cannot answer—*but only for models that can drive the tools.* Six of thirteen models lift significantly (paired 95% CI excludes zero)—three frontier (Claude-Opus-4.8 **+0.30**, Claude-Fable-5 **+0.21**, GPT-5.5 **+0.20**) and three open-weight (Kimi-K2.6 **+0.19**, GLM-4.6 **+0.12**, gpt-oss-120B **+0.10**)—while the other seven, *including frontier Gemini-2.5-Pro (+0.02)*, gain nothing or regress. Aggregated, the six "lifters" go from 0.13 → 0.32 (**+0.18**, ≈2.4×) while the remaining seven move −0.03. The discriminator is **not capability tier but tool-use competence**. On web-answerable controls the lift collapses to ∼0 for two of the three native-web frontier models, confirming the harness does not manufacture a gap.

1 Introduction

When asked a specific question about a company's most recent quarter, an analyst's thesis, or an expert interview, an LLM must either recall it (often impossible—the data postdates training, or was never made public) or search the open web (which will not surface sell-side research, and the accessibility of non-research data can vary widely). Retrieval-augmented and tool-using agents address this, but the practical question for a data provider is:

If you let a model access rich financial data through a standard interface (MCP), how much better does it get, and for whom?

We answer this with a controlled, model-agnostic measurement that we call **the lift**: the difference in answer quality between a model alone (with web access) and the same model with access to financial data from Aiera's MCP server.

Contributions.

1. A **source-grounded, source-agnostic** private benchmark composed of nearly 500 financial-research questions (§3), built so both conditions get a fair attempt and gold answers are verifiable against verbatim sources.
2. A **provider-agnostic lift harness** (§4) that runs the *identical* question through providers' first-party MCP connectors and a generic MCP tool-calling loop (open-weight models, Gemini)—every model reaching the same server.

2 Related Work

Retrieval-augmented and tool-using LLMs. Parametric models store facts but cannot reliably access or update them; retrieval-augmented generation (RAG) couples a generator with a non-parametric index to improve factuality on knowledge-intensive tasks [1]. A parallel line gives models *agency* over external tools: Toolformer shows models can teach themselves when to call APIs [2], while ReAct interleaves reasoning traces with actions to plan and incorporate tool results [3]. Benchmarks of tool- and function-calling competence—e.g. Gorilla's API-calling evaluation [4]—establish that the ability to *drive* tools varies widely across models, a finding our tool-use-competence gradient (§5.3) echoes in the financial-retrieval setting. The Model Context Protocol [5] standardizes this tool interface, which is precisely the intervention we measure: rather than build a bespoke agent, we expose one data source over a standard protocol and ask how much *any* model improves when connected to it.

Financial NLP benchmarks. Financial QA has focused largely on *numerical reasoning over a provided document*: FinQA pairs questions with tables and text from filings and gold reasoning programs [6], and ConvFinQA extends this to multi-turn conversations [7]. Domain-specific models such as BloombergGPT show that finance pretraining helps on these public benchmarks [8]. These evaluations assume the relevant source is already in context. Our benchmark differs in two ways: the agent must *retrieve* the source itself—often proprietary research or point-in-time transcript detail absent from the open web—and questions are deliberately source-agnostic so both conditions get a fair attempt. This shifts the measured quantity from in-context reasoning to grounded retrieval-and-synthesis, and lets us isolate the value of a data *source* rather than of a model's reasoning.

LLM-as-judge and its biases. We score with an automatic judge, following work showing strong LLM judges agree with human preferences at roughly 80% on open-ended tasks [9]. That work also catalogs judge failure modes—position, verbosity, and self-enhancement bias—and Panickssery et al. [10] show that judges can recognize and favor their own generations, with a measured link between self-recognition and self-preference. We mitigate these directly (§4.2, §6): a pinned, held-out judge; a narrow, atomic *key-facts entailment* score rather than holistic preference; and cross-family spot-checks.

3 The Benchmark

3.1 Design principles

- **Grounded gold.** Every item's gold answer is derived from a real Aiera source—an earnings-call transcript, SEC filing, proprietary research note, expert interview, or company publication—and each *key fact* cites a verbatim excerpt from that source.

- **Source-agnostic questions.** Questions are phrased the way a practitioner would ask them ("What are the demand drivers for X heading into 2026?"), never naming the source. This is essential for fairness: a question that names a specific bank's research note would be unanswerable by the baseline and would leak the source.
- **High-lift vs. control.** Each item is tagged `web_findable`. ∼75% are *not* web-findable (proprietary research, point-in-time detail)—where lift is expected; ∼25% are web-answerable controls (mega-cap headline guidance)—where lift should be ∼zero.
- **Private.** Inputs and gold never publish; only scores do. This sidesteps licensing concerns and is why the dataset lives in a private repository.

3.2 Construction

A drafting pipeline retrieves real Aiera sources via API, then an LLM drafts a schema-conforming item—`question`, `gold_answer`, atomic `key_facts` (each with a `source_ref`), `sources` (type, id, date, verbatim excerpt), and metadata (tickers, sector, event type, difficulty, `web_findable`). Items are validated (every key fact traces to a source excerpt) and deduplicated by source id. A 30-item pilot was human-reviewed; the full set are method-validated with sampling QA. Distribution: 378 research / 121 transcript; 5 question types; 24 sectors; ∼75% high-lift.

4 Methodology

4.1 Conditions

- **Baseline** — the model with its *native* web search (each provider's first-party web tool where available; a generic third-party web-search tool on the open-weight path).
- **MCP** — the same model connected to Aiera's MCP server. Frontier models use their providers' *native* MCP connector; open-weight models and Gemini use a **generic function-calling loop** that connects to the same MCP server and bridges its tools to a standard function-calling schema. All models have access to the same tools and data.

4.2 Scoring

- **Key-facts entailment (headline).** For each gold key fact, a held-out judge decides yes/no whether the answer supports it; the score is the fraction supported—atomic, fact-level scoring in the spirit of FActScore [11]. Objective, low-variance, and explainable.
- **Holistic judge (secondary).** A 0–1 agreement score of the answer vs. gold, as a cross-check.
- **Judge.** A single pinned, neutral frontier model, held out of the leaderboard, grades all answers; we probe for self-preference via a cross-family spot-check.

4.3 Experimental setup

A fixed **150-item eval** (116 proprietary / 34 web-answerable control) × 13 models × 2 conditions = 3,900 cells; 3,892 graded after a handful of persistent connector errors. Models: Claude-Opus-4.8, Claude-Fable-5, GPT-5.5 (via first-party MCP connectors); Gemini-2.5-Pro and 9 open-weight models—Llama-4 (Maverick, Scout), Qwen3 (235B, 32B), Gemma-3-27B, gpt-oss-120B, GLM-4.6, Kimi-K2.6, Mistral-Large—via the generic MCP tool-calling loop. Open-weight and Gemini models

are served through a single common host with a shared third-party web-search tool as the baseline, for comparability; the generic loop is capped at 14 tool turns and a fixed token budget across models. The judge is a single pinned, held-out frontier model, grading every answer fact-by-fact. The 150-item eval is a balanced split drawn from the full benchmark.

5 Results

5.1 Headline lift (key-facts, proprietary items)

Per-model mean key-facts score, baseline (model + native web) → MCP (model + Aiera), on the 116 proprietary items, with the paired 95% CI on the difference.

Model	Tier	base	MCP	Δ lift	95% CI
Claude-Opus-4.8	frontier	0.17	0.47	**+0.30**	[+0.22, +0.37]
Claude-Fable-5	frontier	0.14	0.34	**+0.21**	[+0.14, +0.27]
GPT-5.5	frontier	0.17	0.36	**+0.20**	[+0.12, +0.27]
Kimi-K2.6	open	0.13	0.32	**+0.19**	[+0.12, +0.26]
GLM-4.6	open	0.11	0.23	**+0.12**	[+0.06, +0.18]
gpt-oss-120B	open	0.08	0.18	**+0.10**	[+0.04, +0.15]
Gemini-2.5-Pro	frontier	0.09	0.10	+0.02	[−0.02, +0.06]
Llama-4-Maverick	open	0.05	0.06	+0.01	[−0.02, +0.04]
Mistral-Large	open	0.11	0.08	−0.03	[−0.07, +0.01]
Llama-4-Scout	open	0.04	0.00	−0.04	[−0.06, −0.01]
Qwen3-32B	open	0.08	0.04	−0.05	[−0.08, −0.01]
Qwen3-235B	open	0.09	0.03	−0.07	[−0.10, −0.03]
Gemma-3-27B	open	0.09	0.01	−0.08	[−0.11, −0.06]

Table 1: Per-model key-facts lift on proprietary items (paired 95% CI). The six significant lifters span both tiers; the only frontier non-lifter (Gemini-2.5-Pro) sits with the laggards.

Aggregating by behavior rather than tier: the **six lifters** average 0.13 → 0.32 (**+0.18**, ≈2.4×); the **other seven** average 0.08 → 0.04 (−0.03). The split does not fall along the frontier/open boundary—three of the six lifters are open-weight, and the single frontier model that fails to lift (Gemini-2.5-Pro) sits with the laggards.

5.2 Controls

On the 34 web-answerable items the lift should vanish. For two of the three **native-web** models this holds cleanly: Claude-Opus-4.8 0.61 → 0.56 (−0.05) and GPT-5.5 0.53 → 0.42 (−0.11)—both high in baseline and not improved (slightly reduced) by MCP. This is the honest check that the harness does not manufacture a gap where the open web already suffices. The newest connector model, Claude-Fable-5, is the exception: its control lift is *positive* (0.58 → 0.72, +0.15 [+0.02, +0.27])—MCP improves it even where the open web suffices, so its controls are not a clean null. We therefore rest the null-check on the two established models and report Fable-5's controls transparently; its proprietary-item lift (+0.21) remains the comparable headline quantity.

A caveat we report transparently: the open-weight/Gemini **baseline** uses a generic third-party web-search tool, markedly weaker than the providers' native search—these models score ≈0.00–0.02 on control items in baseline, i.e. they barely retrieve from the open web at all. Consequently MCP "lifts" even their controls (e.g. Kimi +0.51, GLM +0.38). The control-as-null-check is therefore

clean only for the native-web models; for the generic-web path, baseline web quality is itself a confound, and the proprietary-item comparison (§5.1)—where neither path can reach the answer on the open web—is the defensible headline.

5.3 The tool-use-competence gradient

To explain *why* lift concentrates where it does, we re-ran a 15-item sample of proprietary questions in the MCP condition with full tool-call tracing—recording, per item, how many and which MCP tools the model invoked, and the resulting answer (180 traced runs). The holistic judge confirms the gradient and sharpens its tail: strong tool-users gain (Claude +0.19, gpt-oss +0.06, Kimi +0.06) while the weakest *regress* (Gemma −0.36, Qwen3-235B −0.30, Llama-4-Scout −0.26)—issuing tool calls but ending up worse than the parametric answer they would have given unaided.

The traces overturn the naïve hypothesis that lift is simply a matter of *calling the tools more*. Across the traced models, the number of tool calls per item is essentially uncorrelated with realized lift (Pearson $r \approx 0.36$): several non-lifters call **more** than the lifters—Qwen3-235B averages 12.7 calls/item and Mistral-Large 9.0, versus 7.6 for top-lifter Claude. Effort, it would seem, is not directly correlated with competence.

What *does* separate them is a within-item fact: **on items where the model lands at least one successful tool call, mean key-facts is 0.225; on items with none, it is 0.000** ($n = 21$ no-retrieval items in the sample). Two distinct failure modes produce the low-lift tail:

Failure mode	Models	Trace signature
Ignores the tools	Gemma-3-27B, Llama-4-Scout	call the tools on only 36% / 43% of items (0.8 / 1.1 calls/item)—they answer from parametric memory, ungrounded
Calls but won't synthesize	Qwen3-235B, Mistral-Large, Qwen3-32B	call heavily (9–13/item, a healthy share to retrieval tools) but emit thin, low-fact answers; Qwen3-32B exhibits a degenerate loop (e.g. repeated `get_research_providers`) that exhausts the turn budget and returns empty

Table 2: The two failure modes behind the low-lift tail.

Lifters, by contrast, both invoke retrieval and convert it: Claude, GLM-4.6, gpt-oss, and GPT-5.5 call retrieval tools and produce substantial, fact-bearing answers (mean ≈2.0k–9.4k chars). The core practical finding: a data interface is **necessary but not sufficient**—realized value is gated by the model's ability to both *drive* agentic retrieval and *synthesize* what it returns. *(Sample is 15 items/model, MCP-only; we use it as mechanistic evidence for the gradient, not a re-estimate of lift, which §5.1 reports on the full subset.)*

5.4 By question type

Lift (key-facts) for the six lifters on proprietary items, by item type:

Question type	base → MCP	Δ
Cross-doc synthesis	0.15 → 0.41	+0.26
Materiality	0.11 → 0.33	+0.22
Thesis	0.10 → 0.27	+0.17
KPI extraction	0.20 → 0.31	+0.11
Guidance change	0.27 → 0.31	+0.05

Table 3: Lift by question type. Largest where proprietary research adds the most.

The lift is largest exactly where proprietary research adds the most—multi-document synthesis and analyst theses—and smallest on guidance-change items, which are closer to public/headline information that can be web-retrievable.

6 Discussion & Limitations

- **Variance.** Agentic MCP retrieval is stochastic—the same model/item can find or miss a source across runs. We mitigate with larger n and report CIs; repeated trials per item are future work.
- **Judge bias.** LLM-as-judge can self-prefer; mitigated via a held-out judge, blinding, and cross-family checks. Key-facts entailment is narrower and less bias-prone than holistic.
- **Tool-use confound.** The lift bundles data quality and the model's ability to use the tools; we treat the latter as a first-class finding rather than a nuisance.
- **Baseline web asymmetry.** Frontier models use their providers' native web search; the open-weight/Gemini path uses a weaker generic web-search tool. This depresses the open-weight baseline (especially on controls, §5.2) and means cross-path baseline levels are not strictly comparable. The proprietary-item lift—where no web tool can reach the answer—is robust to this; the control null-check is clean only for native-web models.
- **Subset size.** Results are on a 150-item eval (116 proprietary); CIs are reported and the run is designed to expand to the full 499-item benchmark.
- **Web nondeterminism.** Native web results drift; runs are date-stamped.
- **Contamination.** Low—inputs are private.

7 Conclusion

A standard data interface (MCP) delivers a large lift on exactly the questions the open web cannot answer: across 13 models, the six that can drive the tools gain +0.18 on average (≈2.4×) on questions that require proprietary research or expert interviews, while web-answerable controls show no lift for the established native-web models. Crucially, the lift is **not** a frontier-vs-open story—three of the six lifters are open-weight, and a frontier model (Gemini-2.5-Pro) is among the non-lifters. The size of the lift is gated by a model's **tool-use competence**: the interface is necessary but not sufficient. For a data provider, the implication is concrete—exposing data over MCP makes *capable* models substantially better at real analysis, regardless of who built them.

References

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

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

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

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

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

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

[7] 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.

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

[9] 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.

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

[11] 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.