The Lift Was Understated: Recomputing Aiera's Value Under a Corrected Harness

Bryan Healey Kieran Black
Aiera

July 7, 2026

Abstract

The Aiera Lift measured how much connecting a model to Aiera's MCP server improves its answers to proprietary, analyst-grade research questions—the *lift* over the same model working with only web search [1]. A companion study later found three defects in the shared evaluation harness (bifurcated MCP transport, an elicitation confound, and a judge self-preference check) and corrected them [2]. Because the Lift is a *difference*—MCP score minus baseline score—and the corrections raise the MCP condition while leaving the web-search baseline untouched, the published lifts were *understated*. We recompute the Lift for all 13 models under the corrected harness, holding everything else fixed. The result strengthens the original finding in every direction: the mean lift across the originally-reported lifters widens from **+0.19** to **+0.28**, the number of statistically-significant lifters grows from six to eight (Mistral-Large and Gemini-2.5-Pro cross into significance), and **GPT-5.5** overtakes Claude-Opus-4.8 as the largest lifter at **+0.38**. Crucially, the study's central claim is unchanged and, if anything, sharper: *the discriminator is tool-use competence, not capability tier*—the models that do not lift remain the ones that cannot reliably drive the tools. We report the corrected numbers.

1 Introduction

The Aiera Lift asked a single question: when connected to Aiera's MCP server (as compared to only having access to web search), how much better does a given model answer the kind of proprietary research questions an analyst actually asks? [1] It reported a *lift*—the paired difference in key-facts accuracy between the two conditions on questions the open web cannot answer—for 13 models, and found that a handful of models gained substantially while the rest were flat, with the discriminator being tool-use competence rather than model tier.

A companion paper subsequently audited the shared research-evaluation harness and found three defects that distorted the MCP condition: models from different providers were scored through *different tool transports*; the agentic loop *rewarded whether an answer was written rather than whether it was understood*; and the single grader had become a ranked entrant, prompting a judge-panel check [2]. Those corrections raise MCP-condition scores—some sharply—while the Lift's *baseline* condition (a generic first-party web search, run without the research tool loop) is unaffected. A lift is MCP − baseline; raising the minuend and not the subtrahend can only widen it. The published lifts were therefore conservative. This paper goes into detail on the corrections applied, and quantifies by how much those corrections impacted the lift.

Contributions.

1. We **recompute the Lift for all 13 models** under the corrected harness, isolating the effect by reusing the original (unchanged) baselines and holding the judge fixed (§2–3).
2. We show the corrected lifts are **uniformly wider** (12 of 13 grow), the significant-lifter count rises from six to **eight**, and the ordering reshuffles within the lifter group (§4).
3. We show the study's **thesis strengthens**: tool-use competence, not tier, still separates lifters from non-lifters—and we report the one model whose lift *narrowed* (§4–5).

2 The correction and why it moves the Lift

We summarize the three harness corrections only briefly; the companion paper will provide the full diagnosis and evidence [2].

Bifurcated transport. Some providers' models reached the MCP server through a first-party managed connector (provider-run tool loop), others through a client-side loop. These are not equivalent, and the managed path under-served several models. The correction routes every model through one consistent client-side loop.

Elicitation confound. The client-side loop returned a model's last text. A model that ran the full research workflow but ended by *offering* its analysis rather than *writing* it was scored as if it had produced nothing—most severely a model whose MCP score more than doubled once fixed. The correction adds a mandatory, tool-free terminal-synthesis turn that adds no new information, only surfaces what was already gathered [4].

Judge panel. The companion study re-graded every stored answer with a three-model panel and found self-preference negligible on this atomic, per-fact metric [5, 6]; single-judge and panel scores were effectively identical. We therefore hold the original single, pinned, held-out judge fixed here, so the judge cannot be a source of the Lift's movement.

3 Method

We rerun the MCP condition for all 13 models on the same 150-item research set the original study used, under the corrected client-side loop with the terminal-synthesis turn, and re-derive each lift as (corrected MCP) − (original baseline) on the proprietary, web-unanswerable items ($n \approx 116$ per model). Reusing the frozen baselines isolates the correction: any change in a lift is attributable to the MCP-condition fixes alone, not to baseline re-sampling or a judge swap. Scoring, item set, and the held-out judge are unchanged from [1]. We report each lift with a paired 95% confidence interval and call a model a *lifter* when that interval excludes zero. To match the original study's protocol we score a single attempt per cell (best-of-1).

4 Results

The eight models above the rule are significant lifters; the five below are not. Four findings:

Model	MCP	Base	Lift (corrected, 95% CI)	Published
GPT-5.5	0.54	0.17	**+0.38** [+0.30, +0.45]	+0.20
Claude-Fable-5	0.51	0.13	**+0.37** [+0.30, +0.45]	+0.24
Kimi-K2.6	0.43	0.13	**+0.30** [+0.23, +0.37]	+0.19
Claude-Opus-4.8	0.40	0.17	**+0.24** [+0.17, +0.31]	+0.30
GLM-4.6	0.33	0.11	**+0.22** [+0.15, +0.29]	+0.12
gpt-oss-120B	0.25	0.08	**+0.17** [+0.10, +0.23]	+0.10
Mistral-Large	0.17	0.11	**+0.07** [+0.02, +0.11]	−0.03
Gemini-2.5-Pro	0.14	0.09	**+0.05** [+0.00, +0.10]	+0.02
Llama-4-Maverick	0.07	0.05	+0.02 [−0.02, +0.06]	+0.01
Qwen3-235B	0.10	0.09	+0.00 [−0.04, +0.04]	−0.07
Llama-4-Scout	0.03	0.04	−0.01 [−0.03, +0.01]	−0.04
Qwen3-32B	0.06	0.08	−0.02 [−0.06, +0.01]	−0.05
Gemma-3-27B	0.01	0.09	−0.09 [−0.11, −0.06]	−0.08

- **The lifts are uniformly wider.** Twelve of thirteen lifts grow or hold; the mean lift across the six originally-reported lifters rises from **+0.19** to **+0.28** (∼1.5×). The models that widen most (GPT-5.5, Claude-Fable-5, Kimi-K2.6) are exactly those whose MCP answers the elicitation and transport bugs had most suppressed.
- **Two models cross into significance.** Mistral-Large moves from a published *negative* lift (−0.03) to a significant **+0.07**, and Gemini-2.5-Pro's small lift becomes significant—taking the count from six lifters to **eight of thirteen**.
- **The order reshuffles, and one model narrows.** GPT-5.5 (**+0.38**) overtakes Claude-Opus-4.8 as the largest lifter. Opus-4.8 is the lone model whose lift *shrinks* (+0.30 → +0.24): its corrected MCP score came in slightly *lower*, i.e. the managed transport had served it marginally better than the standardized loop—the same isolated exception observed for the Opus family on the corrected leaderboard [2]. We report it rather than smooth it.
- **The thesis strengthens.** The reshuffling is entirely *within* the tool-competent group; the non-lifters remain the models that cannot reliably drive the tools (Llama-4-Scout, Qwen3-32B, Gemma-3-27B stay flat or negative, with Gemma significantly negative). *Tool-use competence, not capability tier, is still the discriminator*—now with a wider margin between the groups.

5 Discussion and limitations

That an honest correction *increases* the headline effect is worth stating plainly: the published Lift understated Aiera's contribution because the harness had been under-eliciting the very condition the study exists to measure. The asymmetry is the reason—baselines were untouched, so every fix to the MCP condition flowed straight into the lift. The correction also does not rescue weak models: the genuine non-lifters stay put, and one strong model (Opus-4.8) even narrows, which is the behavior a faithful measurement should show.

- **Single corrected transport.** As in the companion work, we standardize on one client-side loop; a different loop could shift absolute MCP levels, though not the held-constant baseline the lift is measured against [2].

- **Best-of-1.** We match the original study's single-attempt protocol to isolate the correction; a best-of-N protocol (the current leaderboard standard) would raise MCP scores further and is expected to widen lifts more, not less.
- **Private research set.** Research inputs and gold facts remain unpublished; only scores are able to be released, as in the original study [1].

6 Conclusion

Recomputing *The Aiera Lift* under a corrected harness moves every number in the same direction: lifts widen, more models clear significance, and the tool-use-competence thesis sharpens. The published figures were conservative because the measurement had been quietly penalizing the MCP condition; fixing it shows the data connection is worth more, to more models, than we first reported—while leaving the honest negative cases exactly where they were.

References

[1] B. Healey, K. Black. "The Aiera Lift: Measuring the Value of a Financial Data MCP Server Across Foundation Models." Aiera, 2026.

[2] B. Healey, K. Black. "Measuring What the Model Understood: Three Methodology Corrections to the Aiera Research Benchmark." Aiera, 2026.

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

[4] S. Yao, N. Shinn, P. Razavi, K. Narasimhan. "τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains." arXiv:2406.12045, 2024.

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

[6] 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.