BENCHMARK LANDSCAPE REVIEW

FinRetrieval: No Independent Results Outside Daloopa

A landscape review of who has run, cited, or reported on Daloopa's FinRetrieval benchmark

FinRetrieval is a 500-question financial-data-retrieval benchmark published by Daloopa in January 2026. This memo surveys every public engagement with it and finds that, as of July 2026, no party outside Daloopa has published its own results. It documents who has used the benchmark — the authors, the press, and two academic citations — and explains the structural reasons independent reproduction has not occurred.

Research Memo · Generated July 2026
Prepared for external sharing

Contents

Audience and Scope

Written for readers evaluating FinRetrieval as an independent measure of AI-agent performance on financial retrieval. It covers only the question of external reproduction and reporting, not the benchmark's internal methodology, which is documented in the original paper. Findings reflect a public-source review conducted in July 2026 across web search, the arXiv and Semantic Scholar citation graph, the project repositories, and general press and community channels.

1 Summary Finding

As of July 2026, **every reported FinRetrieval number traces back to Daloopa itself**. There is no independent run, no third-party leaderboard entry, and no practitioner reproduction. The only non-Daloopa engagement is two academic papers that cite the benchmark as related work and a set of press outlets that repeat Daloopa's own figures.

> **THE CLAIM UNDER REVIEW**
>
> Daloopa reports that frontier agents reach roughly 90% accuracy when retrieving through its structured-data API. Claude Opus 4.5 reaches 90.8%. Web search alone falls as low as about 20%, a gap of up to 71 percentage points. No party outside Daloopa has independently produced either number.

The absence of independent results changes what the benchmark is. A score that only its author can reproduce works less as an independently contested benchmark and more as a vendor demonstration. The sections below establish that from the public record.

2 The Landscape of Engagement

Every public engagement with FinRetrieval falls into one of three groups. Only the first produces actual benchmark numbers. The other two restate or merely reference them.

Group	Who	Relationship to the results
Daloopa-origin	Paper, blog, repo, dataset	The sole source of actual numbers; authored and hosted by Daloopa.
Press	PR Newswire, Yahoo Finance, A-Team Insight, CompleteAITraining	Repeat Daloopa's figures; no independent testing.
Academic	BigFinanceBench; IPO Finance Agent	Cite it as related work; run their own separate benchmarks.

2.1 Daloopa-origin sources

These are the only places actual scores appear, and all are Daloopa's own:

arXiv paper Kim & Huang — both Daloopa employees

Daloopa blog benchmark write-up pages

GitHub under the `daloopa` org

HF dataset under the `daloopa` org

Author disclosure. The paper's authors are Daloopa employees, and the paper itself lists the single-source, self-referential design as a limitation in its Section 7.3.

2.2 Press coverage

The press cycle originates from a single PR Newswire release and its mirrors. Each restates the same 90.8% and 71-point figures from that release. None re-ran the benchmark or produced independent measurements.

ResearchGate, which also surfaces in searches, hosts only a mirror of the same Daloopa-authored paper, not a separate evaluation.

2.3 Academic citations

The arXiv and Semantic Scholar citation graph shows exactly two citing papers. Both reference FinRetrieval in passing and build their own, unrelated benchmarks. Neither reports a FinRetrieval score.

CITING PAPER 1

BigFinanceBench: A Workflow-Grounded Benchmark for Financial-Research Agents

Wang et al. · *arXiv:2606.03829* (2026).

Cites FinRetrieval in a related-work comparison table only. Runs its own 928-item, expert-authored benchmark, and uses no Daloopa data or tools.

CITING PAPER 2

IPO Finance Agent: Benchmark of LLM Financial Analysts

Benhenda · *arXiv* (2026).

Cites FinRetrieval, then builds its own IPO-diligence benchmark. Reports no FinRetrieval results.

3 Why No One Else Has Attempted It

Three structural factors explain the absence of independent results, and the first is decisive. The benchmark's strong numbers come from one configuration an outsider cannot run without Daloopa. The paper labels it MCP, after the Model Context Protocol, the standard interface an agent uses to reach an external data source. Here that source is Daloopa's own API, and reaching it requires an access token.

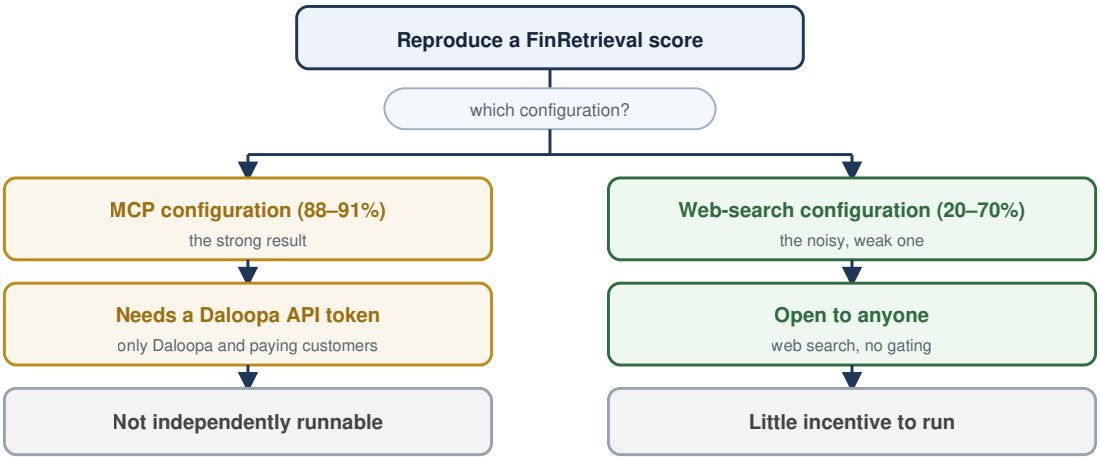

Figure 1. The path to reproducing a score splits by configuration. The strong MCP configuration is gated behind Daloopa's API; the open web-search configuration is weak and noisy. An outsider has neither the access to reproduce the strong result nor a reason to run the weak one.

3.1 The strong configuration is not independently runnable

The 88–91% results all come from the MCP configuration. Only Daloopa and its paying customers hold a token for the API, so only they can reproduce those numbers. An outsider can run only the web-search configuration, the noisy 20–70% one. No one outside Daloopa has both the access to reproduce the strong result and a reason to run it.

3.2 Self-referential design, disclosed by the authors

Daloopa is the data source, the question generator, the retrieval target, and the author. The paper itself lists this as a limitation in Section 7.3. Independent researchers tend to cite that kind of setup as prior work rather than adopt it. That is exactly the pattern the citation graph shows.

3.3 Recency

The benchmark is roughly six months old, published in January 2026. Even setting aside the access barrier, independent reproduction typically has not landed this fast.

4 Bottom Line

There is no example of anyone outside Daloopa providing FinRetrieval benchmark results. The only non-Daloopa engagement is two papers citing it as related work and press outlets repeating Daloopa's own figures.

> *FinRetrieval currently functions more as a vendor demonstration than as an independently contested benchmark.*
>
> — conclusion of this review

None of this speaks to whether Daloopa's numbers are wrong. It speaks to a narrower, verifiable point: no one else has checked them, and by construction almost no one else can. A reader weighing the 90% figure should treat it as a self-reported number from an interested party, pending independent reproduction that the design does not currently make possible.

5 Sources

Reviewed July 2026.

Source	Type	URL
FinRetrieval (Kim & Huang)	Paper	`arxiv.org/abs/2603.04403`
FinRetrieval repository	Code	`github.com/daloopa/finre-trieval`
FinRetrieval dataset	Data	`huggingface.co/datasets/da-loopa/finretrieval`
BigFinanceBench	Citing paper	`arxiv.org/abs/2606.03829`
Daloopa benchmark release	Press	`prnewswire.com/news-releases/daloopa-benchmark-report…`
A-Team Insight coverage	Press	`a-teaminsight.com/blog/ai-agents-need-better-data…`
CompleteAITraining coverage	Press	`completeaitraining.com/news/daloopa-benchmark-finds…`