Benchmark result · April 2026

An autonomous research agent for financial retrieval

Evaluation result on Daloopa FinRetrieval, 500-question benchmark

81.0%

Daloopa FinRetrieval · 500 questions · LLM-judge scoring

Our agent scored **81.0%** on Daloopa's 500-question FinRetrieval benchmark, evaluated under Daloopa's own LLM-judge scoring methodology. In the same public-web retrieval configuration, the best frontier model (GPT-5.2 with reasoning) scored **70.8%**, and Claude Opus 4.5 scored **19.8%**. This result uses **GLM-5.1, an open-weights reasoning model**, paired with our own agent architecture: domain-specific reasoning, targeted query construction, and careful handling of heterogeneous primary-source formats. The lift over the frontier web-retrieval baselines comes from how the agent reasons and searches, not from raw model scale.

Comparison: web + reasoning configurations

Configuration	Accuracy
Our agent + GLM-5.1	**81.0%**
GPT-5.2 + WebSearch + Reasoning	70.8%
Gemini 3 Pro + WebSearch + Reasoning	69.2%
Claude Opus 4.5 + WebSearch + Reasoning	19.8%

Benchmark composition

500 questions spanning **41 countries**, with 475 distinct issuers and fiscal periods from 2015 to 2026.

Top issuer HQ countries		Question category	
United States	42%	Balance sheet	20%
Japan	10%	Cash flow	19%
UK · Australia	5% ea.	Operational KPIs	18%
Brazil	4%	Income statement	16%
India · Canada	3% ea.	Guidance / outlook	15%
+34 others	28%	Segments / geography	13%

Benchmark: Daloopa FinRetrieval. Comparable-configuration scores from Daloopa's April 2026 paper (arXiv 2603.04403). All rows use public web retrieval + reasoning, with identical LLM-judge scoring. Geographic and category distribution derived from the released dataset.

How the agent answers a question

> *"What was Prestige Estates Projects Ltd's consolidated net profit for the period for calendar Q1 2025, in INR millions?"*
>
> Source: Daloopa FinRetrieval, Q21

1 · Translate the period before searching

Before any web lookup, the agent recognises Prestige as an Indian issuer and applies its domain knowledge that Indian companies use a March fiscal year-end. It translates the question's "**calendar Q1 2025**" into **fiscal Q4 FY25**, and its first search already uses the reformulated framing rather than the calendar label.

2 · Identify the primary source through reformulated queries

The agent issues its searches using the reformulated fiscal framing, not the question's surface-level calendar label. This distinction matters for what comes back: the company's own Q4 FY25 quarterly results filing appears among the top results, alongside news coverage and third-party aggregators that typically dominate financial queries.

3 · Reject a conflicting secondary source

One news article in those results reports the quarter's net profit as ₹25 crore. The agent recognises this as the **standalone** figure, not the **consolidated** figure the question asks for, and continues to the primary filing rather than accepting the secondary number.

4 · Read and search the scanned filing

The filing is a **19-page scanned image PDF** with no extractable text. The agent combines its OCR tool (powered by an open-weights small OCR model) with its document reading and in-document search tools, working through the filing to locate the consolidated net profit line for the quarter ending 31 March 2025.

Agent's answer

₹431 million

Value and unit both match Daloopa ground truth · primary source: company-hosted Q4 FY25 results

> **What this shows**
> The agent reasons about fiscal-calendar conventions across jurisdictions, generates queries in the relevant business framing, rejects conflicting secondary sources by reasoning about financial nuance (standalone vs consolidated), and uses its own tools, including an OCR tool powered by an open-weights small OCR model, to handle file-format edge cases. The same pattern scales across the 500-question benchmark: this is how we reached 81% with an open-weights reasoning model at the core.