TECHNICAL EVALUATION REPORT

HydraDB Performance Assessment

on FinanceBench

Prepared by: HydraDB Research Team

Version: 1.0 | May 2026

1. About This Report

This report shares the results of running HydraDB against FinanceBench, an industry benchmark for financial question answering. The goal is to give you a clear view of how HydraDB performs on real-world financial documents so you can judge its fit for your use case.

2. About the Benchmark

FinanceBench is a public benchmark built from 10-K, 10-Q, 8-K, and earnings documents of publicly traded companies. Each question has a human-verified answer and a pointer to the exact passage in the source document that supports it. The open-source sample covers a mix of direct metric lookups, domain knowledge, and reasoning-based queries.

HydraDB offers two retrieval modes, and we evaluated both:

- Fast mode is optimized for low-cost, high-throughput retrieval. We ran this mode against the full set of 150 open-source questions.
- Thinking mode performs additional reasoning over candidate passages to improve ranking quality. We ran this mode against a 120-question subset.

We evaluated HydraDB on two things that matter for production use: how often it finds the correct supporting evidence in the document, and how much context it sends to the downstream language model.

3. Retrieval Accuracy

Retrieval accuracy is measured using Recall@K. For a given value of K, this is the share of questions where the correct supporting passage appears somewhere in the top K results that HydraDB returns. The table below compares both modes side by side.

Top K Results	Fast Mode (%)	Thinking Mode (%)	Improvement
Top 1	44.1%	**50.3%**	**+6.2 pts**
Top 3	68.1%	**74.3%**	**+6.2 pts**
Top 5	78.9%	**84.3%**	**+5.4 pts**
Top 10	89.0%	**91.4%**	**+2.4 pts**

Table 1. Recall@K for fast mode and thinking mode.

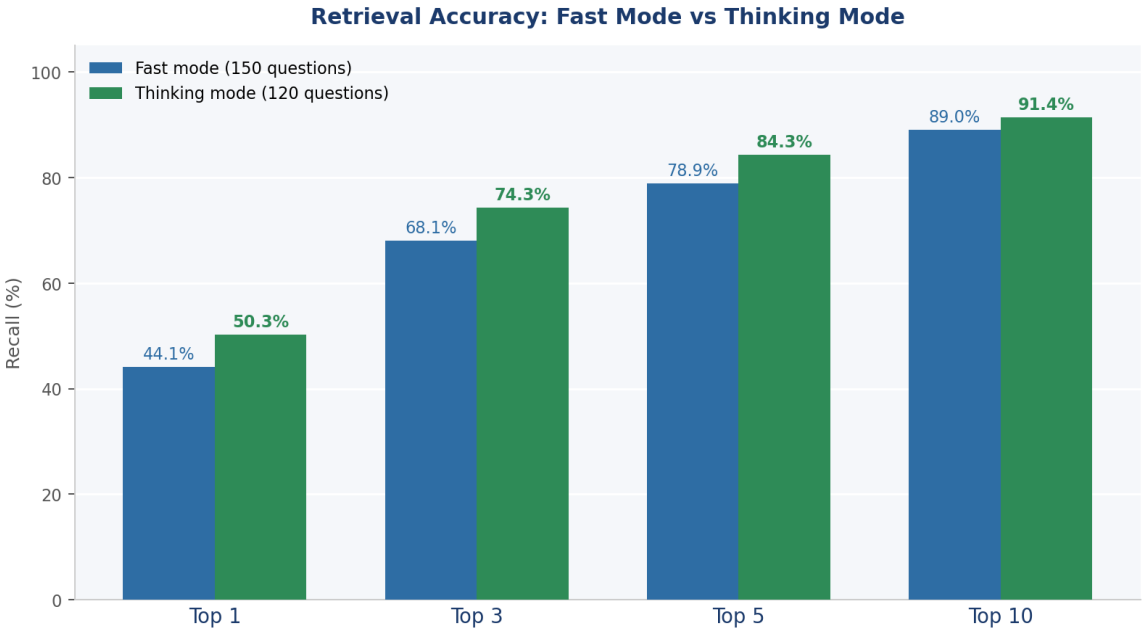

Figure 1. Recall@K across both modes.

In fast mode, HydraDB surfaces the correct piece of evidence within its top 10 results for nearly 9 out of 10 questions. At the top 5 level, the correct answer is present about 79% of the time, which is a practical working range for passing context into a language model.

Thinking mode improves recall at every level, with the biggest gains at the top of the ranking. Recall@1 rises by more than 6 points, and Recall@5 crosses 84%. This is the mode to choose when you need the first or first few results to be correct as often as possible, for example in workflows where the downstream model will only see the top result or where answer quality is sensitive to the order of retrieved context.

Fast mode remains the right choice when throughput and cost are the priority and the downstream model can work well with the top 5 or top 10 results. Both modes converge at Recall@10, which tells us the underlying retrieval is finding the correct evidence in nearly every case, and thinking mode is primarily reordering the results to put the right one higher up.

4. Context Size

When HydraDB returns results, it assembles them into a context package for the downstream model. Smaller and more predictable context sizes lead to lower inference costs and faster responses from whichever model you pair HydraDB with. The numbers below are from the fast mode run.

Measure	Tokens
Smallest context	**5,299**
Average context	**7,997**
Largest context	**12,175**

Table 2. Context size per query.

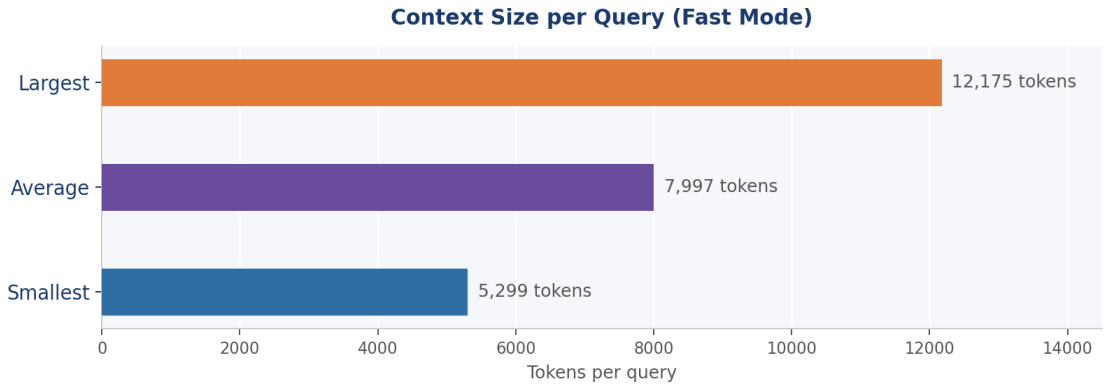

Figure 2. Context size per query in fast mode.

On average, HydraDB produces about 8,000 tokens of context per query. The largest context in the benchmark was around 12,000 tokens, which fits comfortably inside the context window of every major language model on the market today. The size is predictable across queries, which makes cost planning straightforward.

5. Summary

On FinanceBench, HydraDB retrieves the correct evidence in the top 10 results 89% of the time in fast mode and 91% of the time in thinking mode, with an average context size of roughly 8,000 tokens per query. Thinking mode adds 5 to 6 points of recall at the top of the ranking, which is the range that matters most when you want the first result to be correct.

The benchmark shows HydraDB can reliably find the right information inside long, complex financial filings and deliver it in a form that is ready to use with any modern language model. The two modes give you a direct choice between throughput and ranking precision based on what your application needs.

If you would like to discuss how these results translate to your specific documents and workflows, or arrange a pilot on your own data, please get in touch with your HydraDB contact.

References

[1] HydraDB. HydraDB: Technical Paper. benchmarks.hydradb.com/HydraDB.pdf

[2] Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., and Vidgen, B. (2023). FinanceBench: A New Benchmark for Financial Question Answering. arXiv:2311.11944. arxiv.org/pdf/2311.11944

[3] HydraDB FinanceBench Evaluation Code. GitHub repository. github.com/usecortex/hydradb-bench