DATADOG | Fintool

CASE STUDY

Fintool reduces customer churn by 84% and monitors complex operations with Datadog LLM Observability

ABOUT FINTOOL

Fintool is a financial copilot tailored for institutional investors. It leverages large language models (LLMs) to discover financial insights beyond the reach of timely human analysis.

Financial Services

San Francisco

> “By using Datadog LLM Observability, we’ve improved response accuracy and reduced latency, ensuring faster, more reliable insights for our customers.”
>
> **Nicolas Bustamante**
> CEO
> Fintool

WHY DATADOG?

- Identifies performance bottlenecks across multi-provider, multi-model LLM infrastructure
- Accelerates issue resolution through rapid LLM input/output analysis, out-of-the-box alerts, and comprehensive end-to-end LLM pipeline visualization
- Improves response speed while maintaining high answer quality and accuracy

CHALLENGE

Fintool wanted to deliver a top-tier customer experience and reduce churn by ensuring their multi-provider, multi-model LLM application consistently delivered accurate, low-latency responses.

USE CASE

LLM Observability

KEY RESULTS

84%

Reduction in customer churn

60%

Reduction in response latency

Conquering latency and complexity in financial data processing

Fintool is revolutionizing institutional investing with its innovative financial copilot. Through an intuitive chat interface, the platform analyzes earnings calls, SEC filings, and financial data to provide informed investment insights. Accomplishing this requires Fintool to process an unprecedented scale of data: 70 million chunks, 2 million documents, and around 5 TB of data for every 10 years of financial data and filings. They also process several billions of tokens per month.

Fintool's capabilities rest on a complex data pipeline. Each user query triggers dozens of LLM calls to classify companies, determine time periods, identify required data types, expand query keywords, and run specialized classifiers. The system executes multiple Elasticsearch queries to retrieve relevant documents, performs complex reranking steps to ensure that results are optimally ordered based on relevance and context, and runs through custom eval checks to verify every piece of information and numerical data.

The company recently faced two challenges that hampered its ambitious growth plans. Its multi-provider architecture—spanning Azure, OpenAI, Amazon Bedrock, Elasticsearch, Groq, and Cerebras—created a complex web of dependencies where a single provider's performance issues could cascade through the entire system, potentially disrupting service. The infrastructure complexity was further complicated by the need to maintain consistent performance across different geographical regions and varying workload patterns.

More critically, the team discovered a direct correlation between response speed and customer retention. Their detailed analysis revealed that slower chatbot responses led to increased churn rates, particularly during the critical evaluation period when potential customers were testing the platform.

Leveraging LLM Observability to deliver high-quality and accurate responses

After initially using basic log observability for AWS and Azure, Fintool turned to Datadog LLM Observability to tackle its challenges and gain insights into its LLM pipelines. Despite its complex, multi-provider infrastructure, the transition proved straightforward. Datadog's native auto-instrumentation support for OpenAI, Azure OpenAI, Anthropic, and Amazon Bedrock enabled Fintool to seamlessly capture all LLM calls alongside detailed operational metrics like cost, latency, and usage trends in minutes. This comprehensive view helped them efficiently pinpoint latency bottlenecks and track which models generated specific responses, including intermediate steps that contributed to the final outputs. Through careful analysis of its complex LLM architecture, Fintool efficiently pinpointed and addressed latency bottlenecks throughout its system.

With this enhanced visibility, Fintool implemented a sophisticated real-time rotation system between providers and models like GPT-4o or Llama 3.3 70B across different Azure regions. “The hard part is that latency could stem from at least 25 different operations,” says Nicolas Bustamante, CEO. “We needed to identify and monitor each one in real time."

The engineering team leveraged Datadog’s monitoring capabilities to gain granular insights into their LLM pipelines, tracking latencies, token consumption, and associated costs with precision. They implemented strategic optimization, including load balancing across data centers with higher GPU capacity.

With LLM Observability’s auto-clustering and topic modeling features, Fintool also improved the quality of their prompts and outputs. Fintool had built out a robust evaluation process with detailed eval checks to verify the precision and accuracy of their inputs. By submitting the custom evaluations into LLM Observability and layering evaluations onto LLM Observability’s Cluster Map, they quickly identified areas where response quality could improve, leading to refined prompt instructions and updated document sources. The team also implemented sophisticated caching mechanisms and query optimization techniques to reduce response times for frequently requested financial data.

Delivering a first-rate customer experience

Fintool’s optimization efforts yielded remarkable results that have transformed its operations. Response times for complex, multi-step queries improved by 60 percent—a significant achievement that directly contributed to higher conversion rates during free trials and reduced customer churn by 84 percent. Today, the ability to pinpoint and preemptively address failure points across the LLM pipeline enables proactive issue resolution and more streamlined operations.

In addition, the company achieved comprehensive visibility across its entire system, from LLM calls to Elasticsearch queries and reranking operations. This enhanced monitoring capability allowed the company to confidently scale operations to meet growing customer demands while maintaining high standards for accuracy and reliability. The improved system intelligence now adapts seamlessly to varying workloads, ensuring consistent performance, even during peak market hours. “By using Datadog LLM Observability, we’ve improved response accuracy and reduced latency, ensuring faster, more reliable insights for our customers,” says Bustamante.

Going forward, Fintool plans to continue to push the boundaries in financial technology. The team is currently investigating new approaches to natural language processing and developing new ways to optimize its multi-provider architecture. “Our goal is to make institutional-grade financial analysis accessible and lightning-fast,” says Bustamante. “Every millisecond counts when making investment decisions, and we’re committed to staying at the forefront of financial technology innovation.”

GET STARTED WITH A FREE TRIAL TODAY ›