

CUSTOMER USE CASE

Groq + Fintool: AI-powered Financial Insights

groqcloud™



Fintool is an AI equity research copilot for institutional investors. Fintool uses Large Language Models (LLMs) to discover financial insights beyond the reach of timely human analysis.

The Challenge

For finance professionals, every second counts. A brief delay can mean the difference between capitalizing on an opportunity or missing out. But they face a daunting task when it comes to analyzing the performance and projecting the future of public companies: there's terabytes of data! They have to sift through reams of financial reports, such as SEC filings, earnings call transcripts, and annual 10-K reports, to uncover the hidden insights that might move the market.

With hundreds of filings and documents produced per day, it's practically impossible for analysts and investors to keep up. They end up making important financial decisions based on imperfect, incomplete information.

The Solution

[Fintool](#) is an AI-based equity research copilot for institutional investors. Analysts can ask Fintool Chat any sort of financial research question, including something complicated ("find Russell 3000 companies with low gross margins and revenue growth under 10% that are exposed to US tariffs on China, either through top-line impact or cost structure"). The copilot gets to work to find the answer, relying on various LLMs for dozens of operations, including classifying companies, determining time periods, identifying required data types, expanding query keywords, and running domain-specific classifiers. These operations create queries to retrieve relevant data from publicly available filings, and the data is fed into yet another LLM to generate a precise response.

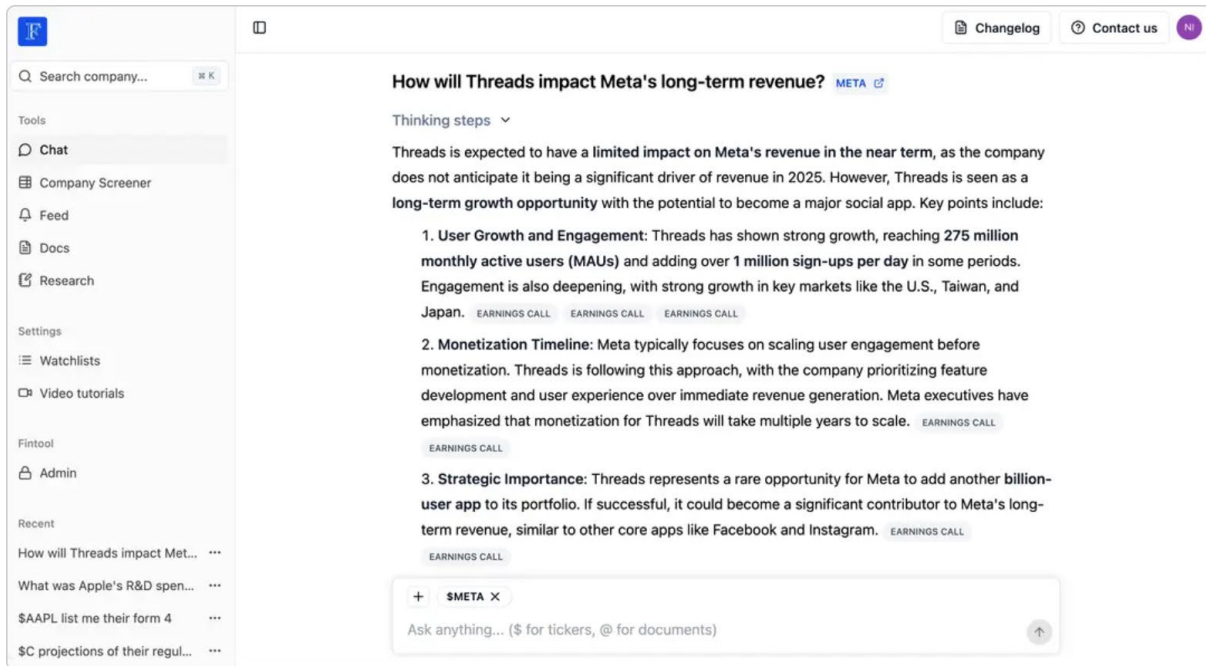
Finance professionals have zero tolerance for inaccuracies - one small mistake can lead to big losses. Before it shares any answer with a user, Fintool employs a network of LLM agents to verify the accuracy of every piece of information it generates.

Performing all these operations could take a lot of time, and finance professionals are not well known for their patience. Fintool partners with GroqCloud as its inference partner, to ensure its critical operations are executed with ultra-low latency.



Get Instant, Accurate Answers

Get precise answers in seconds, backed by millions of SEC filings, earnings calls, and conference transcripts. Every response includes direct source quotes and intelligent follow-up suggestions.



Previously, Fintool relied on Open-AI GPT-4o for key processes such as query understanding, expansion, and classification, but once its architects and developers got a taste of GroqCloud speed, they transitioned those operations over to Llama 3.3 70B, powered by GroqCloud. Processing speed jumped over 7X (111 tokens / second to 823). A few more months of testing, and Fintool moved more of its critical operations to GroqCloud, including its multi-agent verification system that employs adversarial checks to ensure the accuracy and validity of every response. Oh, and since GroqCloud is so efficient, Fintool's cost per token dropped by 89% when it moved to GroqCloud from OpenAI. Fintool poured some of those savings back into the solution: it can now feed even more context and tokens into its chat prompts, boosting quality.

Add it up, and Fintool got much faster, better, and more efficient as a result of moving its inference to GroqCloud. This change was pivotal in securing the highest score on the Finance Bench – the industry's premier LLM performance benchmark for financial queries. And they're not done yet: Fintool continues to explore how GroqCloud performance and its wide variety of available models can help it expand its performance and quality lead.



We optimized our infrastructure to its limits – but the breakthrough came with GroqCloud™. Overnight, our chat speed surged 7.41x while costs fell by 89%. I was stunned. So, we tripled our token consumption. We simply can't get enough.

—Nicolas Bustamante, CEO, Fintool

The Opportunity

Machine learning revolutionized quantitative finance by enabling high-frequency trading (HFT), where algorithms analyze terabytes of data in real time to exploit market inefficiencies. Now, large language models are set to transform investing by understanding textual data and context at scale. In seconds, Fintool will be able to assess a company's moat, evaluate management quality, perform sentiment analysis across dozens of filings, and generate a buy or sell recommendation—bringing the power of AI-driven analysis to fundamental investing.