

# Scaling domain-tuned AI without the cost of proprietary models

Llama helped Oxide AI improve domain precision and reinforce its explainable AI stack to deliver trusted insights at scale



# At a glance

Oxide AI upgraded its financial services app, Oxogen AI, with lightweight Llama models to improve domain-specific relevance and optimize performance. After fine-tuning with low-rank adaptation (LoRA), the small Llama model delivered results on par with GPT, fitting naturally into Oxide's integrative, white-box AI system for delivering intelligent and accessible financial insights.



## Use case

Enhancing financial research with fine-tuned, domain-specific large language models (LLMs)



## Goal

Deliver precise, transparent insights at scale with optimized LLM performance



## Llama version

Llama 2 7B and Llama 3.1 8B



## Deployment

AWS Cloud, Amazon Bedrock, IBM Cloud, IBM watsonx

## Results\*

**40%**

reduction in production costs for LLM-generated content and extractions

**37%**

faster average sequential response times than OpenAI GPT-4, while maintaining comparable quality levels

**95%**

accurate results after a month of LoRA tuning

# The challenge

## Using AI in finance demands precision and transparency at low cost

In modern finance, the volume, velocity and complexity of information far exceed human comprehension, making smart decision-making nearly impossible for the average investor. Oxide AI is transforming investing with Oxogen AI, an intelligent app that performs automated deep web research across entire markets to deliver high-value insights. At the core of Oxogen AI is an innovative large-scale AI system powered by objective-driven reasoning agents.

LLMs generate finance-focused articles based on input from Oxide's proprietary AI engine, EvoQ™, transforming evidence-based facts and reasoning into human-friendly text. This multi-layered approach ensures accuracy and full control over the generative process, eliminating any risk of hallucination. Early versions of Oxogen AI relied on GPT, but fine-tuning for the finance domain was necessary. GPT's API-only access and per-token fees also made large-scale deployment costly and impractical for processing very large data streams at high speed.

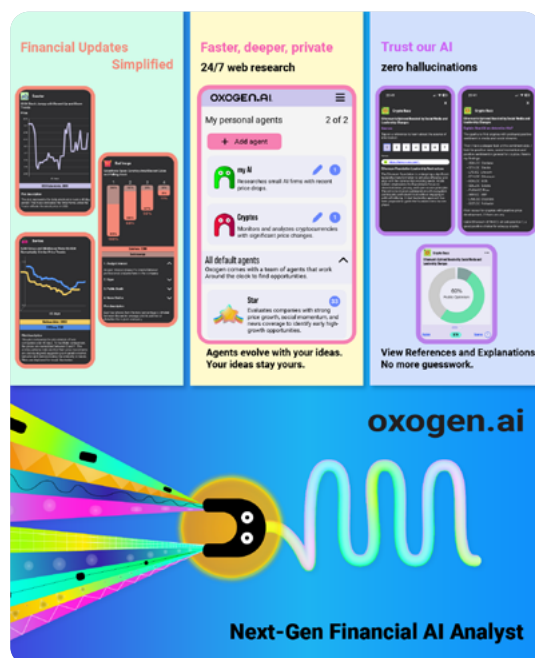
## Open-source AI gave Oxide AI the autonomy they needed

As open-source LLMs matured, Oxide saw an opportunity to bring fine-tuning, cost optimization and data security in-house. With the launch of production-ready Llama, Oxide had a viable alternative to proprietary models.

## OXIDE.AI®

Oxide AI's mission is to amplify individuals by giving them access to enterprise-grade AI technology that solves the critical problem of information overload. Oxide's solutions deliver precise, context-aware insights at the right time, empowering people to make smarter, faster decisions, supported by trusted and transparent AI.

- **Industry:**  
Technology
- **Company size:**  
10-person startup



The Oxogen AI app uses an agentic approach to find the best unbiased financial opportunities.

# The solution

## Llama integration delivered fine-tuned performance and total control

With the release of Llama, Oxide gained access to a family of AI models with open weights that could be fine-tuned and run on both hybrid cloud instances and local hardware.

Oxide fine-tuned the small Llama models in IBM Watson® Studio using LoRA. Unlike full weight fine-tuning, LoRA/QLoRA creates adapters that modify a fraction of the model's weight. This significantly reduced training time and computing costs, enabling the team to train task-specific adapters on their proprietary financial knowledge base and experiment rapidly.

Llama generated domain-specific results with 95% accuracy and reached parity with GPT. Unlike GPT, Llama provided a high degree of control over the model's behavior in a financial context, directly resulting in increased value from Oxide's intellectual property and custom data.

## Small Llama models and lower inference costs made scaling up affordable

Once fine-tuned, smaller Llama models generated domain-focused results comparable to proprietary models with hundreds of billions of parameters. Using only a tiny fraction of the computing resources, open-source Llama models don't come with per-token fees. The team was able to migrate the new, Llama-powered Oxogen AI app onto production infrastructure seamlessly and scale it globally without the massive cost increases associated with proprietary models and the vast data streams they must process.



“Llama stood out as an open, transparent model that we could fine-tune using our proprietary datasets. The smaller Llama model was also a great fit for our production needs, allowing us to maintain quality while optimizing compute cost and speed. Beyond performance, the fact that Llama is backed by Meta gave us confidence in long-term support and innovation.”

Kateryna Wikström  
CPO and Designer, Oxide AI

## AI agents use neuro-symbolic AI to fuse quantitative and qualitative insight

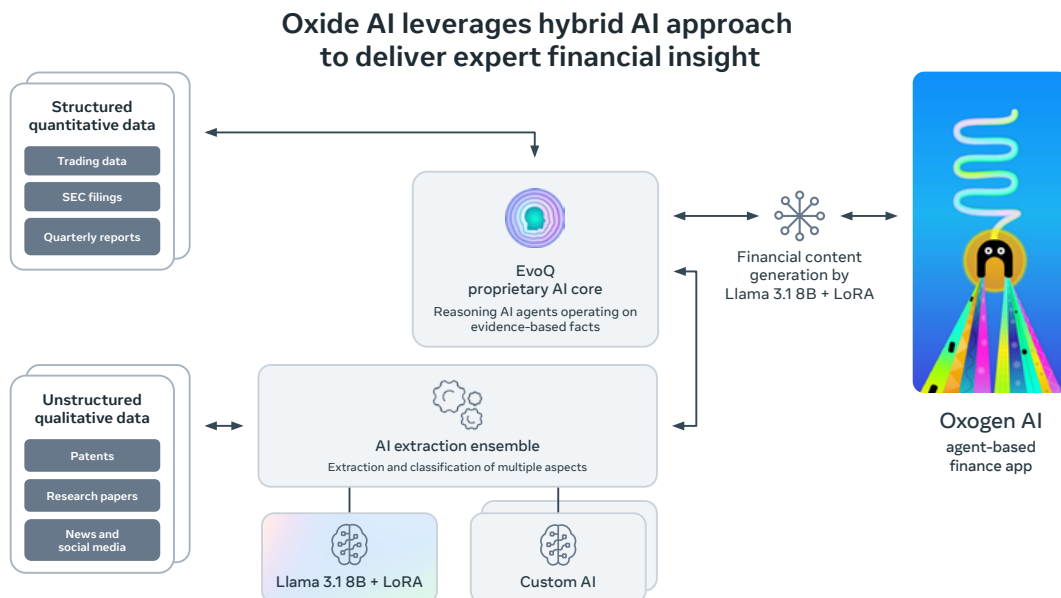
Financial information lives on a spectrum, from hard numbers like quarterly reports and stock prices to the latest wisdom-of-the-crowd sentiment trending on social media. To truly understand the value of assets, such as companies, bonds, options or cryptocurrencies, it’s essential to analyze the entire spectrum. That’s an immensely complex challenge given the vast volume and speed of financial information. Accuracy and explainability are paramount, as no hallucinations can be tolerated when making decisions.

To turn data into actionable insights, Oxide uses Llama-powered generative AI and other custom AI models to extract and classify qualitative information from patents, research, news and social media. This information is then combined with structured data from transactions, SEC filings and other quantitative sources.

The two streams converge in EvoQ, Oxide’s proprietary neuro-symbolic engine, where multiple reasoning AI agents operate at high speed over entire financial markets. These agents utilize innovative neuro-symbolic AI, combining pattern recognition with other control structures to extract high-value insights. The numerical insights produced by EvoQ are then translated by Llama into various content types, making them accessible and actionable for users.

“Our hybrid-AI approach combines highly refined domain-specific data, computational AI and Llama LLM support into a powerful, efficient system. It enables us to deliver precise, high-quality, context-aware outputs at a fraction of the cost, achieving results that would otherwise take hundreds of hours of manual research.”

Lars Hard  
Chief AI Officer, Oxide AI



Oxide AI’s EvoQ engine uses an ensemble of AI models to extract multi-perspective data and processes it through reasoning AI agents to generate deep financial insights.



# The outcome

## An AI upgrade that put Oxide AI in control of their data and performance

Shifting to open-source Llama helped address key challenges around privacy, domain-specific fine-tuning and production efficiency, all while achieving major cost savings. Small Llama models matched the accuracy of GPT but used just a fraction of the computing resources. With Llama powering the generative AI component of the EvoQ's neuro-symbolic AI engine, Oxogen AI is positioned to transform personal financial services on a global scale, far beyond traditional investing solutions.

**40%**

reduction in production costs for LLM-generated content and extractions

**37%**

faster average sequential response times than OpenAI GPT-4, while maintaining comparable quality levels

**95%**

accurate results after a month of LoRA tuning

“Meta is constantly updating Llama models and making them available across major cloud platforms quickly, so integration into our workflows is smooth and scalable. Regular improvements, easy integration and flexible fine-tuning made Llama the natural choice for building a trustworthy, efficient and domain-specific generative AI layer in our platform.”

Kateryna Wikström  
CPO and Designer, Oxide AI

# Conclusion

## Bringing intelligent financial agents to phones, watches and XR

Powered by EvoQ, the team has enabled intelligent financial agents within Oxogen AI, allowing users to create and run 24/7 deep web research missions, targeting over 6,000 financial market players.

The next step is bringing these AI agents to the edge and running user-specific workloads directly on smartphones, smartwatches and XR devices. By shifting compute away from the cloud, they aim to reduce latency and strengthen privacy without compromising intelligence. The ultimate goal is precision: delivering only the most relevant information on interfaces, such as AR glasses and smartwatches. Smaller Llama models, optimized for multi-modal inference, help make this possible.

### Oxogen on Meta Quest 3, mixed reality



“As the Llama ecosystem continues to mature, we see a path toward even tighter integration between Oxide AI’s proprietary AI stack and on-device intelligence, enabling more private, responsive and sustainable AI experiences.”

Lars Hard  
Chief AI Officer, Oxide AI

### Oxogen on Apple watch



Oxogen AI solutions support various edge devices.



## How can Llama help your business?

See how open-source Llama brings unmatched control, customization and flexibility to generative AI application development and deployment.

[Learn More](#)

[Related Stories](#) ▶