Positron AI Raises $230 Million Series B at Over $1 Billion Valuation to Scale Energy-Efficient AI Inference

Co-led by ARENA Private Wealth, Jump Trading, and Unless, with strategic investment from Qatar Investment Authority (QIA), Arm Holdings, and Helena

Funding accelerates Positron's roadmap from shipping Atlas systems today to next-generation Asimov silicon, targeting tape-out in late 2026 and production in early 2027; announced at Web Summit Qatar

---

Reno, NV – Positron AI, the leader in energy-efficient AI inference hardware, today announced an oversubscribed $230 million Series B financing at a post-money valuation exceeding $1 billion.

The round was co-led by ARENA Private Wealth, Jump Trading, and Unless, and includes new and strategic investment from Qatar Investment Authority (QIA), Arm Holdings, and Helena. Existing investors Valor Equity Partners, Atreides Management, DFJ Growth, Resilience Reserve, Flume Ventures, and 1517 also participated. The financing validates Positron's mission to make AI inference dramatically cheaper and more energy-efficient at scale.

“We're grateful for this investor enthusiasm, which itself is a reflection of what the market is demanding,” said Mitesh Agrawal, CEO of Positron AI. “Energy availability has emerged as a key bottleneck for AI deployment. And our next-generation chip will deliver 5x more tokens per watt in our core workloads versus Nvidia’s upcoming Rubin GPU. Memory is the other giant bottleneck in inference, and our next generation Asimov custom silicon will ship with over 2304 GB of RAM per device next year, versus just 384 GB for Rubin. This will be a critical differentiator in workloads including video, trading, multi-trillion parameter models, and anything requiring an enormous context window. We also expect to beat Rubin in performance per dollar for specific memory-intensive workloads.”

Positron is building the infrastructure layer that makes AI usable at scale by lowering the cost and power required to run modern models. The company's shipping product, Atlas, is an inference system designed for rapid deployment and scaling. Atlas is also a fully American-fabricated and manufactured silicon and system, enabling fast production ramp and dependable supply for customers who need capacity quickly.

“Memory bandwidth and capacity are two of the key limiters for scaling AI inference workloads for next-generation models,” said Dylan Patel, founder and CEO of SemiAnalysis, an advisor and investor in Positron. SemiAnalysis is a leading research firm specializing in semiconductors and AI infrastructure that provides detailed insights into the full compute stack. “Positron is taking a unique approach to the memory scaling problem, and with its next-generation Asimov chip, can deliver more than an order of magnitude greater high-speed memory capacity per chip than incumbent or upstart silicon providers.”

Jump Trading Leads After Deploying Atlas

A key highlight of the round is Jump Trading's decision to co-lead after first becoming a customer.

"For the workloads we care about, the bottlenecks are increasingly memory and power—not theoretical compute,” said Alex Davies, Chief Technology Officer of Jump Trading. “In our testing, Positron Atlas delivered roughly 3× lower end-to-end latency than a comparable H100-based system on the inference workloads we evaluated, in an air-cooled, production-ready footprint with a supply chain we can plan around. The deeper we went, the more we agreed with Positron’s roadmap—Asimov and the Titan systems—as a memory-first platform built for future workloads. We invested because Positron combines traction today with a roadmap that can reshape the cost curve and capabilities for inference.”

“Jump Trading came to Positron as a customer,” said Agrawal. “As they saw our roadmap for Asimov, our custom silicon, and Titan, our next-generation system, they chose to step up as a co-lead investor. A customer becoming an investor is one of the strongest validations we can receive. It signals both technical conviction and real-world demand.”

Building Toward Asimov and Titan: A Memory-First Platform for Next-Generation Inference

Positron's next-generation custom silicon, Asimov, is designed around the reality that modern AI workloads are increasingly limited by memory bandwidth and capacity, not just compute flops. Asimov is designed to support 2 terabytes of memory per accelerator and 8 terabytes of memory per Titan system at similar realized memory bandwidth to NVIDIA's next-generation Rubin GPU. At rack scale, this translates to memory capacity of well over 100 terabytes.

This memory-first architecture unlocks high-value inference workloads, including long-context large language models, agentic workflows, and next-generation media and video models. Positron is on track to tape out its Asimov chip just 16 months after its June Series A financing gave it the resources to fully launch the design process, and the company intends to maintain this pace with future chips. “To us, development speed is an essential competitive advantage,” said Agrawal. “Competing with Nvidia means matching their shipping frequency, and we have designed our organization around that goal.”

“Positron is solving one of the most important bottlenecks in AI: delivering inference at scale within real-world power and cost constraints,” said Ari Schottenstein, Head of Alternatives at ARENA Private Wealth. “The combination of shipping traction today with Atlas, plus a credible path to Asimov, creates a rare opportunity to define a new category in AI infrastructure.”

Positron is building this platform with an ecosystem of industry leaders, including Arm, Supermicro and other key technology and supply-chain partners.

Momentum and Growth Trajectory

Positron expects strong revenue growth in 2026, positioning the company to become one of the fastest-growing silicon companies ever, achieving large-scale commercial traction in roughly 2.5 years from company launch. The company is working with multiple frontier customers across cloud, advanced computing, and performance-sensitive verticals, and continues to expand deployments and customer programs.

About Positron AI

Positron AI builds purpose-built hardware and software to make AI inference dramatically cheaper and more energy-efficient. Positron's shipping product, Atlas, is designed for rapid, scalable deployment, and the company's next-generation custom silicon, Asimov, targets tape-out toward the end of 2026 with production in early 2027. Positron's systems are built to serve long-context and next-generation AI workloads with leading economics. Learn more at positron.ai.

Media Contact:
Helen Cho
Bonfire Partners
press@bonfirepartners.io