VULTR

Case Study

ConfidentialMind AI Enables Secure, Scalable AI Deployments with Vultr

ConfidentialMind simplifies AI integration for enterprises, offering one-click deployment of large language model (LLM) endpoints, databases, and data connectors. With an OpenAI-compatible API, developers can seamlessly integrate AI into applications without requiring prior AI or Kubernetes expertise.

By partnering with Vultr, ConfidentialMind gains access to high-performance AMD Instinct™ GPUs, enabling secure, scalable, and cost-effective AI deployments.

Powering AI innovation with flexible cloud infrastructure

Enterprises rely on ConfidentialMind to deploy AI solutions across on-premises, private, and public clouds, ensuring complete control over their technology stack while avoiding vendor lock-in. However, scaling AI workloads at an enterprise level presented several challenges:

- Limited GPU availability for AI inference workloads
- High upfront hardware costs
- Excessive cloud overhead for LLM processing
- Complexity in managing and optimizing GPU infrastructure

To overcome these challenges, ConfidentialMind sought a cloud provider with high AMD Instinct GPU availability, flexible configurations, and enterprise-grade support – particularly in the U.S. market.

Unlocking AI scalability with Vultr

ConfidentialMind CTO Severi Tikkala first connected with Vultr at the AI Infra Summit in San Francisco, where Vultr's newly launched AMD Instinct GPU cloud services aligned perfectly with their needs. The AMD Instinct™ MI300X GPUs stood out for their ability to power AI workloads efficiently and cost-effectively.

"At ConfidentialMind, we specialize in deploying generative AI systems like semantic search, RAG, and AI agents across on-premises and cloud environments. Vultr Cloud GPUs, including AMD Instinct™ MI300X, provide the computational power we need for running AI models. Vultr Kubernetes Engine streamlines deployment and scaling, enabling our customers to build AI solutions with data security and operational efficiency," states Tikkala.

ConfidentialMind

Industry

Artificial Intelligence

For use cases in:

- Financial Services
- Healthcare and Life Sciences
- Manufacturing
- Entertainment
- Gaming
- Media and Entertainment
- Retail
- Telecommunications

About

ConfidentialMind AI enables enterprises to easily deploy secure, scalable, generative AI applications. It leverages high-performance GPUs and a Kubernetes-based infrastructure to streamline AI adoption.

confidentialmind.com

> "Vultr Cloud GPUs, including AMD Instinct™ MI300X, provide the computational power we need for running AI models. Vultr Kubernetes Engine streamlines deployment and scaling, enabling our customers to build AI solutions with data security and operational efficiency."
>
> **Severi Tikkala,** CTO at ConfidentialMind

Building AI solutions with Vultr's cloud services

ConfidentialMind leverages Vultr's infrastructure to support enterprise-grade AI deployments, utilizing:

- **GPU compute nodes:** A mix of NVIDIA and AMD Instinct GPUs, with AMD Instinct™ MI300X, favored for high VRAM and efficiency.
- **Vultr Kubernetes Engine (VKE):** Simplifying AI workload deployment and scaling.
- **Supporting services:** Block storage, load balancing, and cloud-native tools for optimized performance.

Depending on customer needs, ConfidentialMind deploys configurations ranging from dual NVIDIA L40S or A100 GPUs to large-scale AMD Instinct™ MI300X clusters with up to eight GPUs per instance.

Maximizing cost efficiency and performance

By choosing Vultr over hyperscalers like AWS, GCP, or Azure, ConfidentialMind estimates at least 50% cost savings on GPU compute costs. Vultr's combination of high-performance AMD Instinct GPUs and competitive pricing allows them to scale AI workloads efficiently without unnecessary operational overhead.

"Generative AI deployments are mostly restricted by the amount of VRAM that the LLMs require," said Tikkala. "The AMD Instinct™ MI300X offers 192 GB of VRAM per GPU, which allows us to deploy the whole AI system on a single GPU. Vultr's affordable prices and scalable on-demand cloud allow us to serve customers with high-performance AI deployments affordably. The cost savings are estimated at least 50% compared to the hyperscalers."

This strategic combination of hardware and cloud services enables ConfidentialMind to optimize performance while maintaining cost efficiency, making high-end AI deployments more accessible for enterprises.

Driving AI adoption and business growth

With Vultr's scalable cloud infrastructure, ConfidentialMind enables enterprises to:

- **Accelerate AI deployments:** Customers can launch AI applications without infrastructure bottlenecks.
- **Expand market reach:** Vultr's presence supports seamless AI adoption in North America.
- **Ensure enterprise-grade security:** ConfidentialMind optimizes AI performance while maintaining data sovereignty.

Looking ahead

With Vultr's scalable infrastructure and cost-efficient cloud GPUs, ConfidentialMind continues to drive AI adoption for enterprises, ensuring secure, high-performance deployments across industries.

> "Vultr's affordable prices and scalable on-demand cloud allow us to serve customers with high-performance AI deployments affordably. The cost savings are estimated at least 50% compared to the hyperscalers."

Get started with your own Vultr success story.

Contact us at sales@vultr.com or visit vultr.com/sales →