VULTR

DATASHEET

Vultr Serverless Inference

Train anywhere, infer everywhere

VULTR.COM

VULTR

Vultr Serverless Inference

Deploy and scale Generative AI (GenAI) models quickly and efficiently, with the ability to use your proprietary data, powered by the simple-to-manage Vultr Serverless Inference's global acceleration.

Train anywhere, infer everywhere

Vultr Serverless Inference, tailored for GenAI applications, offers a global, self-optimizing platform for effortless deployment and serving of models. Leveraging a serverless architecture, it eliminates the complexities of scaling and managing infrastructure, allowing businesses to prioritize innovation. Turnkey retrieval-augmented generation enables users to upload their data or documents as inputs for deployed models without requiring model training or risking data leakage to public models. Vultr Serverless Inference's OpenAI-compatible API makes deployment and integration simple. Its extensive global network guarantees efficient, reliable performance with minimal latency across six continents. Users benefit from superior performance, reduced operational costs, and autonomous scalability.

Why it's important right now

The demand for deploying sophisticated AI models globally with minimal latency is escalating. Businesses need a cost-effective and automatically scalable solution that is easy to manage. Vultr Serverless Inference offers these solutions, plus the ability for users to harness their unique information in their proprietary data without training a model, all while adhering to strict security protocols and compliance regulations.

Deploy and deliver pre-trained AI models

Flexibility and choice

With Vultr Serverless Inference, speed-to-market is enhanced by eliminating the need to retrain or reconfigure AI models for deployment, and operational costs are consequently reduced. With turnkey RAG, companies can securely upload their data to the included private retrieval-augmented generation vector database and leverage a pre-trained public model for custom outputs without the cost of training a model. These outputs are highly customized and directly relevant to the business's specific nuances, leading to more accurate, practical solutions tailored to the company's needs.

Data sovereignty and protection

Deploying in-region and performing retrieval-augmented generation using a private vector database allows businesses to maintain control over sensitive information, reducing data privacy and security risks. When using Vultr Serverless Inference, businesses maintain ownership and control and can have confidence in Vultr's dedication to data protection, security, and privacy.

Advanced security measures

Vultr is dedicated to meeting our customers' diverse global risk and compliance needs, including server availability, security, data protection, and privacy. Our commitment to compliance is demonstrated through our CSA Star Level I assessment (CAIQ), SOC 2 Type 2 attestation, and ongoing adherence to the SOC 2 framework.

© Vultr 2024

Key features built to address critical enterprise challenges

Autonomous scalability and global reach

Vultr Serverless Inference autonomously scales GenAI applications, matching demands and optimizing performance without manual intervention. Its global infrastructure ensures AI solutions are accessible with low latency across six continents, making it the ideal platform for international businesses.

Operational and overhead efficiency

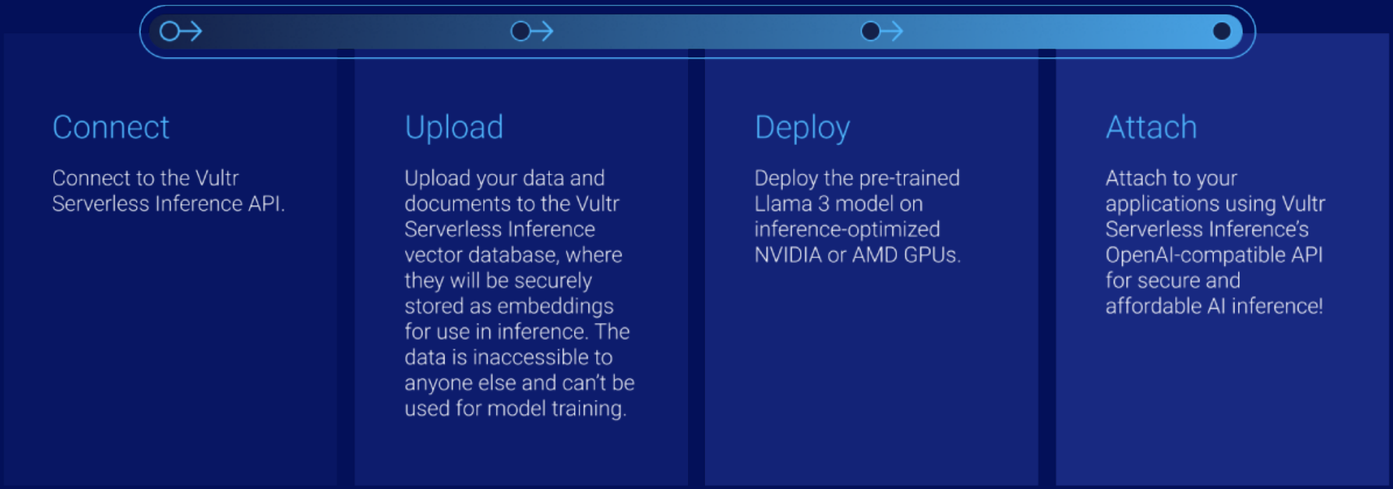
Leveraging our serverless architecture significantly reduces the complexity and overhead of managing AI infrastructure. This streamlines the deployment process, making advanced AI capabilities accessible to companies without extensive in-house expertise.

Unbeatable price-to-performance

Inference pricing varies significantly across the industry. Vultr dynamically adjusts resource allocation and provides a pay-as-you-go model, allowing businesses to scale their AI applications in line with their growth without needing to make significant capital expenditures.

Vultr Serverless Inference

Deploy AI securely without the complications of infrastructure management.



Connect

Connect to the Vultr Serverless Inference API.

Upload

Upload your data and documents to the Vultr Serverless Inference vector database, where they will be securely stored as embeddings for use in inference. The data is inaccessible to anyone else and can't be used for model training.

Deploy

Deploy the pre-trained Llama 3 model on inference-optimized NVIDIA or AMD GPUs.

Attach

Attach to your applications using Vultr Serverless Inference's OpenAI-compatible API for secure and affordable AI inference!

© Vultr 2024

AI-driven transformation: Key industry use cases

AI deployment for the modern enterprise

Vultr Serverless Inference offers innovative businesses the agility and scalability required to deploy GenAI applications globally without the complexities of infrastructure management and, with turnkey RAG, without model training. Enabling seamless integration, with adherence to stringent security and compliance standards, empowers companies to focus on innovation and growth. This platform ensures businesses can leverage the full potential of AI with unmatched efficiency, operational flexibility, and cost-effectiveness, meeting the dynamic demands of today's digital landscape.

Financial services

Fraud detection

Utilize Vultr Serverless Inference for real-time analysis of transaction data across global banking systems to identify and prevent fraudulent activities. Banks and financial institutions can instantly flag and investigate suspicious activities by deploying sophisticated AI models that can learn from transaction patterns, thereby reducing economic losses and enhancing customer trust.

Real-time risk assessment for loans and insurance

Leverage Vultr Serverless Inference to deploy AI models that instantaneously assess the risk profiles of loan applicants or insurance policy seekers based on various data points. This allows financial institutions and insurers to quickly make more informed, data-driven decisions, optimizing their risk management strategies while providing faster customer service.

Healthcare & life sciences

Predictive patient care and monitoring

Deploy AI models via Vultr Serverless Inference to continuously monitor patient data from wearable devices and electronic health records, enabling predictive analytics for patient care. This use case can identify potential health issues before they become critical. This allows for timely intervention and personalized care plans, improving patient outcomes and operational efficiency in healthcare facilities.

Telecommunications

Network optimization and anomaly detection

With Vultr Serverless Inference, telecom companies can analyze real-time network traffic, identifying patterns that indicate potential issues or inefficiencies. AI-driven models can predict network congestion points, detect anomalies indicating security breaches, and optimize resource allocation to ensure high-quality, uninterrupted service for customers.

Retail

Dynamic pricing and inventory management

Utilizing Vultr Serverless Inference, retailers can deploy advanced AI models to analyze real-time data streams from multiple sources, including sales trends, customer behavior, market conditions, and inventory levels. This enables dynamic pricing strategies where prices are adjusted in real-time to reflect demand, competition, and stock levels, maximizing profitability while ensuring competitive pricing for customers. Simultaneously, the system can optimize inventory management by accurately predicting future product demand, ensuring optimal stock levels across locations to meet consumer demand without overstocking or stockouts. This dual approach enhances operational efficiency and customer satisfaction while significantly boosting revenue and market responsiveness for retailers in highly competitive landscapes.

Media & entertainment

Content recommendation engines

Vultr Serverless Inference enhances content discovery on streaming platforms by deploying AI models that analyze viewing habits, content preferences, and engagement metrics. This delivers personalized content recommendations, improving viewer satisfaction and engagement rates and driving platform loyalty and growth.

Learn more about
Vultr Serverless Inference

Contact us at vultr.com to get started. →

© Vultr 2024