OVHcloud

Innovation for freedom

Press release

OVHcloud Launches AI Endpoints to Simplify and Democratize Access to AI Models

A new serverless platform offering a library of open-source models designed to meet a wide range of business and industry use cases

Roubaix – April 29th 2025 – OVHcloud today launches AI Endpoints, a serverless solution that enables developers to effortlessly add high-value AI functions to their apps. With over 40 advanced and powerful open-source LLM and generative AI models, addressing use cases such as chatbots, text-to-speech or coding assistance, AI Endpoints makes it easier to democratize AI whatever the size of the organization. Without the need to manage the underlying infrastructure or requiring Machine Learning expertise, AI Endpoints enables easy access to open-source AI models hosted in a trusted Cloud environment.

AI ready to consume: easily level-up business applications

OVHcloud AI Endpoints enables developers to test AI features in a sandbox environment before deploying at scale in applications, internal tools, or business processes. Use cases include:

- Integrating LLMs into applications: Easily add conversations through agents. With real-time, natural language interactions, AI Endpoints helps improve user engagement or automate customer service at scale.
- Text Extraction: Advanced machine learning models automatically extract, interpret, and organize unstructured data, playing a key role in ETL (Extract, Transform, Load) pipelines to improve operational efficiency.
- Bring speech to your app: Through APIs, the service converts spoken language into text and vice versa, enabling transcription and interactive voice-based transcriptions.
- Coding assistance: With tools like Continue, developers can integrate private, real-time AI assistance directly into their IDEs. Code suggestions, error detection, task automation all improve both productivity and code quality.

A serverless platform answering organizations need for strategic autonomy

OVHcloud's sovereign cloud infrastructure gives peace of mind to developers, assuring them that data is hosted in Europe, and is protected against non-European regulations, providing both technical and strategic autonomy.

The cloud is central to AI and AI Endpoints runs on OVHcloud's energy-efficient infrastructure, powered by water-cooled servers in ecofriendly data centers. This helps reduce the environmental impact of AI workloads, without compromising performance.

AI Endpoints promotes full model transparency through open weight AI models. This ensures that organizations can implement the same models on their infrastructure or deploy them on other Cloud services, whilst keeping control over data.

"*We are excited to launch AI Endpoints and are humbled by the incredible feedback we get from our amazing community. With support for the most diverse and sought after open source LLM models, AI Endpoints helps to democratize AI so developers can add to their apps the most cutting-edge models. Our solution enables them to do this easily in a trusted cloud environment with full confidence in OVHcloud's sovereign infrastructure*" said Yaniv Fdida, Chief Product and Technology Officer, OVHcloud.

Availability: AI in a pay-as-you-go model

After a preview phase, the service has been thoroughly developed and tested, incorporating customer feedback including a number of highly requested features such as support for stable open-source models, more choice and finer API key management. With more than 40 open-source cutting-edge AI models, AI Endpoints include:

LLMs: Llama 3.3 70B, Mixtral 8x7B, …
SLMs: Mistral Nemo, Llama 3.1 8B, …
Code: Qwen 2.5 Coder 32B, Codestral Mamba
Reasoning: DeepSeek-R1 (distilled Llama)
Multimodal: Qwen 2.5 VL 72B, …
Image generation: SDXL
Speech: ASR (speech-to-text), TTS (text-to-speech)

It is available now in APAC, Canada and Europe and deployed from the Gravelines datacenter. With a pay-as-you-go model, OVHcloud AI Endpoints pricing varies on a per model basis using the number of tokens consumed per minute as a unit.

Resources

- Learn more about OVHcloud AI Endpoints
- Learn more about OVHcloud
- Follow OVHcloud on X
- Follow OVHcloud on LinkedIn

About OVHcloud

OVHcloud is a global cloud player and the leading European cloud provider operating over 450,000 servers within 43 data centers across 4 continents to reach 1,6 million customers in over 140 countries. Spearheading a trusted cloud and pioneering a sustainable cloud with the best performance-price ratio, the Group has been leveraging for over 20 years an integrated model that guarantees total control of its value chain: from the design of its servers to the construction and management of its data centers, including the orchestration of its fiber-optic network. This unique approach enables OVHcloud to independently cover all the uses of its customers so they can seize the benefits of an environmentally conscious model with a frugal use of resources and a carbon footprint reaching the best ratios in the industry. OVHcloud now offers customers the latest-generation solutions combining performance, predictable pricing, and complete data sovereignty to support their unfettered growth.

CONTACT

Media relations

Julien Jay

Communications & Public Relations Manager

media@ovhcloud.com

+33 (0)7 61 24 46 67