OVHcloud

Innovation for freedom

Communiqué de presse

OVHcloud lance AI Endpoints pour simplifier et démocratiser l'accès à l'IA et aux modèles d'entrainement

Cette nouvelle plateforme serverless offre une bibliothèque de modèles open-source, conçus pour répondre à un large éventail de cas d'utilisation

Roubaix – 29 avril 2025 – OVHcloud lance aujourd'hui AI Endpoints, une solution serverless qui permet aux développeurs d'intégrer facilement des fonctionnalités d'IA avancées à leurs applications. Grâce à plus de 40 modèles open-source de pointe incluant LLM et IA générative – pour des usages comme les agents conversationnels, modèles vocaux, assistants de code, etc. - AI Endpoints démocratise l'utilisation de l'IA, indépendamment de la taille ou du secteur de l'organisation. En allégeant les contraintes d'infrastructure nécessaire, et sa gestion, et sans besoin d'une expertise en machine learning, AI Endpoints permet un accès facile aux modèles d'IA open-source hébergés dans un environnement Cloud de confiance.

L'IA clé en main pour développer vos applications

OVHcloud AI Endpoints offre un environnement de test des fonctionnalités d'IA (sandbox) pour expérimenter les fonctionnalités IA avant un déploiement à grande échelle dans les outils métiers, applications ou processus internes. Les cas d'utilisation incluent :

- Intégration des LLM dans les applications : grâce à des interactions en temps réel et en langage naturel, AI Endpoints permet d'améliorer l'engagement des utilisateurs ou d'automatiser le service client à grande échelle.
- Extraction de texte : exploiter les modèles avancés d'apprentissage automatique pour extraire, interpréter et organiser automatiquement les données non structurées, optimisant ainsi les processus ETL (Extraction, Transformation, Chargement) et l'efficacité opérationnelle.
- Intégration de la parole : grâce aux API, le service convertit la langue parlée en texte et vice versa pour la transcription automatique, la prise de notes ou encore des interactions vocales enrichies.
- Assistance au codage : intégrée nativement aux environnements de développement intégrés (IDE), elle permet aux développeurs de bénéficier d'assistants et d'agents IA (en mode privé) et en temps réel améliorant à la fois la productivité et la qualité du code : suggestions de code, détection d'erreurs, automatisation des tâches…

Une plateforme serverless pensée pour l'autonomie stratégique des organisations

L'infrastructure cloud souveraine d'OVHcloud garantit l'hébergement des données Europe, et leur protection vis à vis des réglementations extra-européennes, pour une autonomie à la fois technique et stratégique.

Le cloud est essentiel à l'IA, aussi AI Endpoints s'exécute sur l'infrastructure d'OVHcloud laquelle bénéficie de serveurs refroidis à l'eau dans des datacenters respectueux de l'environnement. Cette approche permet de minimiser l'impact de l'IA, sans compromis sur la performance.

La transparence est au cœur de la plateforme avec des modèles open weight, offrant la possibilité de déploiements sur l'infrastructure propre d'une organisation, ou sur d'autres clouds, tout en gardant la maîtrise des données.

« *Nous sommes fiers de lancer AI Endpoints et reconnaissants des apports enthousiastes et riches de notre communauté. Grâce à l'intégration des modèles LLM open source les plus recherchés, AI Endpoints contribue à démocratiser l'IA et permet aux développeurs d'intégrer facilement les modèles les plus innovants, en toute confiance dans l'infrastructure souveraine d'OVHcloud*. », a déclaré Yaniv Fdida, Chief Product and Technology Officer, OVHcloud.

Disponibilité : IA avec tarification à l'usage

Après une phase *preview* et l'intégration des retours clients (support de modèles stables, gestion affinée des clés API…), AI Endpoints propose plus de 40 modèles IA open-source de dernière génération, dont :

LLM : Llama 3.3 70B, Mixtral 8x7B, ...
SLM : Mistral Nemo, Llama 3.1 8B, ...
Code : Qwen 2.5 Coder 32B, Codestral Mamba
Raisonnement : DeepSeek-R1 (Llama distillé)
Multimodal : Qwen 2.5 VL 72B, ...
Génération d'images : SDXL
Voix et discours : ASR (speech-to-text), TTS (text-to-speech)

Le service est dès à présent disponible en Europe, au Canada et dans la région APAC, déployé depuis le datacenter de Gravelines. La tarification est calculée à l'usage, basée sur le nombre de tokens (jetons) consommés par minute et par modèle.

Ressources

- En savoir plus sur OVHcloud AI Endpoints
- En savoir plus sur OVHcloud
- Suivez OVHcloud sur X
- Suivez OVHcloud sur LinkedIn

À propos d'OVHcloud

OVHcloud est un acteur mondial et le leader européen du Cloud opérant plus de 450 000 serveurs dans 43 centres de données sur 4 continents à destination de 1,6 million de clients dans plus de 140 pays. Fer de lance d'un Cloud de confiance et pionnier d'un Cloud durable au meilleur ratio prix-performance, le Groupe s'appuie depuis plus de 20 ans sur un modèle intégré qui lui confère la maîtrise complète de sa chaîne de valeur : de la conception de ses serveurs, à la construction et au pilotage de ses centres de données, en passant par l'orchestration de son réseau de fibre optique. Cette approche unique lui permet de couvrir en toute indépendance l'ensemble des usages de ses clients en leur faisant profiter des vertus d'un modèle raisonné sur le plan environnemental avec un usage frugal des ressources et d'une empreinte carbone atteignant les meilleurs ratios de l'industrie. OVHcloud propose aujourd'hui des solutions de dernière génération alliant performance, prévisibilité des prix et une totale souveraineté sur leurs données pour accompagner leur croissance en toute liberté.

CONTACTS PRESSE

RELATIONS MEDIAS OVHCLOUD
Julien Jay
Communications & public relations manager
+33 (0)7 61 24 46 67
MEDIA.FRANCE@OVHCLOUD.COM

AGENCE OMNICOM REPUTATION GROUP
Constance Kunicki - 07 85 93 58 59
Eugenie Dautel – 06 76 46 93 26
PAR.OVH@OMNICOMPRGROUP.COM