DEVELOPERS CAN NOW DEPLOY MULTIMODAL MODALS ON HUGGING FACE USING FRIENDLIAI'S GPU INFERENCE

GPU-optimized AI inference platform enables developers to easily deploy, scale, and operate generative AI models without investing in complex infrastructure.

REDWOOD CITY, CALIFORNIA , UNITED STATES, May 21, 2025 /EINPresswire.com/ -- FriendliAI, an AI inference infrastructure company, today announced that Hugging Face developers can utilize FriendliAI's inference infrastructure service to deploy and serve multimodal AI models directly in the Hugging Face Hub. Hugging Face is the world's largest AI model and dataset platform with over 7 million users.

FriendliAI's GPU-optimized AI inference platform enables developers to easily deploy, scale, and operate generative AI models without investing in complex infrastructure. The company offers one of the fastest processing speeds among global GPU-based model API providers (according to Artificial Analysis benchmark results). With over 370,000 directly deployable models on Hugging Face, FriendliAI offers unmatched model coverage.

Hugging Face users can now deploy multimodal models as well as language models to Friendli Endpoints directly from the Hugging Face Hub with just one click, ensuring high performance and low latency for even the most complex, resource-intensive tasks.

Multimodal models represent the next frontier in AI, enabling systems that understand and interact across text, images, video, and audio. These models are powering an entirely new class of applications—from vision-language agents and voice-driven assistants to content generation tools that blend modalities in real time. However, deploying them reliably at scale is significantly more demanding than traditional language models. FriendliAI addresses these challenges with a streamlined, production-ready multimodal solution.

"Building on our successful partnership that brought high-performance language model inference to Hugging Face developers, we're thrilled to expand our collaboration to include multimodal AI capabilities," said Byung-Gon Chun, CEO of FriendliAI. "Hugging Face developers now have the power to leverage our optimized infrastructure to a broader range of AI models, allowing them to seamlessly deploy both text and multimodal models with the same efficiency and performance they've come to expect."

As multimodal AI powers a new wave of applications such as agents that understand images and respond with speech, or tools that blend text, visuals, and audio, developers need infrastructure that scales with the demands of multimodal workloads. By combining Hugging Face's expansive model ecosystem with FriendliAI's inference infrastructure, FriendliAI is making it simple and cost-effective for developers to push the boundaries of what's possible with multimodal AI without getting slowed down by infrastructure complexity.

"We're excited to deepen our strategic partnership with FriendliAI to bring their specialized inference capabilities to multimodal models," said Julien Chaumond, CTO at Hugging Face. "Our community has already benefited greatly from FriendliAI's high-performance infrastructure for language models, and this expansion represents a natural evolution as developers increasingly work with sophisticated multimodal AI systems. By offering one-click deployment of multimodal models through FriendliAI's optimized infrastructure, we're removing technical barriers and enabling our community to build more powerful and diverse AI applications while maintaining the seamless experience they value on the Hugging Face Hub."

Deploying generative AI models at scale can be a daunting task, with complex infrastructure and soaring operational costs standing in the way. Friendli Dedicated Endpoints simplify multimodal model deployment at scale—reducing both complexity and cost, so teams can focus on building. Powered by FriendliAI's GPU-optimized custom inference backend, this fully managed service offers lightning-fast, cost-efficient model serving with dedicated GPU resources and seamless, automatic resource management—no manual tuning required.

###

lisa langsdorf
GoodEye PR
+1 347-645-0484
email us here