Press Release

# AI Innovators Worldwide Choose Oracle for AI Training and Inferencing

*Fireworks AI, Hedra, Numenta, and Soniox experience accelerated performance and cost efficiency with Oracle Cloud Infrastructure*

**Austin, Texas—Jun 18, 2025**

AI innovators across the world are using Oracle Cloud Infrastructure (OCI) AI infrastructure and OCI Supercluster to train AI models and deploy AI inference and applications. Fireworks AI, Hedra, Numenta, Soniox, and hundreds of other leading AI innovators have selected OCI for its scalability, performance, cost efficiency, choice of compute instances, and control over where to run their AI workloads.

As industries rapidly adopt AI to help drive innovation and efficiency, the AI companies that are providing these services require reliable, secure, and highly available cloud and AI infrastructure that enables them to quickly and economically scale out GPU instances. With OCI AI infrastructure, AI companies gain access to high-performance GPU clusters and the scalable computing power needed for AI training, AI inference, digital twins, and massively parallel HPC applications.

"Among AI innovators, OCI has rapidly become the destination of choice for training and inferencing needs of all sizes," said Chris Gandolfo, executive vice president, Oracle Cloud Infrastructure and AI. "OCI AI infrastructure delivers ultra high-speed networking,

optimized storage, and cutting-edge GPUs that AI companies rely on to power the next wave of innovation."

## Global AI Innovators Choose Oracle

Fireworks AI is an inference platform that empowers developers and businesses to build highly optimized and production-ready generative AI applications, serving over 100 state-of-the-art open models in text, image, audio, embedding, and multi-modal formats. Fireworks AI uses OCI Compute bare metal instances accelerated by NVIDIA Hopper GPUs and OCI Compute with AMD MI300X GPUs to help it serve over two trillion inference tokens daily on its platform and scale its services globally.

"Developers rely on Fireworks AI to integrate generative AI into their products, optimized for latency, throughput and cost per token," said Lin Qiao, co-founder and CEO, Fireworks AI. "With OCI AI infrastructure, we can deliver the ultra-fast response times and production-grade stability that developers expect. We're able to process AI workloads efficiently, minimize downtime, and help ensure AI applications run smoothly at scale so that our customers can focus on innovation without worrying about the underlying infrastructure."

Hedra, an AI-driven video creation company, enables users to create videos with life-like characters. By deploying its multimodal foundation models for generative image, video, and audio on OCI Compute bare metal instances accelerated by NVIDIA Hopper GPUs, Hedra reduced its GPU costs, experienced faster training speeds, and reduced its model iteration time.

"Creating expressive character videos at scale requires immense computational power and efficient multimodal processing," said Michael Lingelbach, founder and CEO, Hedra. "OCI handles our model training and inferencing across video, audio, and image data, while providing the rapid processing required for real-time character rendering and meeting the high storage demands of large datasets. This enabled us to release our latest model, Character-3, and content creation platform, Hedra Studio, quickly and without a hitch."

Numenta is an AI technology company focused on maximizing the performance and efficiency of deep learning systems. By using OCI Compute bare metal instances accelerated by NVIDIA GPUs, Numenta gained access to a range of reliable and high-performance training instances, achieving faster training speeds and increased cycles of learning.

"OCI provides the high-performance infrastructure and strong operational support we need to push the boundaries of AI without compromising speed or accuracy," said Dan Steere, CEO, Numenta. "With OCI, we've been able to confidently accelerate the

development of our next-generation technology, taking significant steps forward in Efficient Intelligence™."

Soniox, an AI company at the forefront of audio and speech AI, pioneers foundational AI models for audio, speech, and language comprehension. With its new universal multilingual speech AI model hosted on OCI, Soniox uses OCI Compute bare metal instances accelerated by NVIDIA Hopper GPUs to train its model to seamlessly recognize and understand speech across 60 languages in real-time with low latency and higher accuracy.

"A high-performance infrastructure that provides improved accuracy, speed, and cost efficiency was top-of-mind when selecting a cloud provider to support our growth," said Klemen Simonic, founder and CEO, Soniox. "OCI gives us access to the latest AI innovations that allow us to push the boundaries of speech recognition and audio understanding while significantly reducing deployment time and operational costs."

## Additional Resources

- Learn more about Oracle Cloud Infrastructure
- Learn more about OCI AI infrastructure
- Learn more about OCI Generative AI
- Learn more about Oracle's AI strategy

## Contact Info

### Adrienne Halford

Oracle PR
adrienne.halford@oracle.com
+1.916.531.1255

## About Oracle

Oracle offers integrated suites of applications plus secure, autonomous infrastructure in the Oracle Cloud. For more information about Oracle (NYSE: ORCL), please visit us at oracle.com.

## Trademarks