



VMware Cloud Foundation Brings Leading AI Models to the Private AI Cloud

August 31, 2026

Nemotron 3, Gemma 4, cotomi, Qwen3.7-Max, and GLM-5.2 Validated to Run on VMware Cloud Foundation

LAS VEGAS, Aug. 31, 2026 (GLOBE NEWSWIRE) -- **VMware Explore 2026** -- Broadcom Inc. (NASDAQ: AVGO), a global technology leader that designs, develops, and supplies semiconductor and infrastructure software solutions, today announced that leading AI models from providers including Google, NVIDIA, NEC, Alibaba Cloud, and Z.ai, are validated to run on VMware Cloud Foundation (VCF), enabling customers to bring these AI models on-premises and deliver "model as a service" to their users. As enterprises accelerate the movement of inference workloads to private cloud driven by privacy, security, governance, and AI tokenomics, VCF provides the AI- and Kubernetes-native platform that gives organizations across industries a production-ready path to deploying a wide variety of AI models.

According to [Broadcom's Private Cloud Outlook 2026](#), 56% of enterprises are already running or planning to run production AI inferencing on private cloud. VCF is a unified private cloud platform capable of running inference workloads, agentic applications, containerized services, and traditional VMs together, eliminating the operational fragmentation of managing separate stacks. [Independent benchmark testing under ML Perf Inference v5.1](#) standards confirms that VCF delivers performance on par with bare metal, making it the ideal platform of choice for enterprises deploying AI at scale on premises.

"Broadcom is committed to giving enterprises the broadest set of AI models for their on-premises infrastructure, all validated on VCF," said Chris Wolf, global head of AI and advanced services, VMware Cloud Foundation Division, Broadcom. "Working with the world's leading AI model providers, we're giving organizations a clear path to data sovereignty and cost-effective AI at scale, with leading models available securely and delivered as a service to their user community through VMware Cloud Foundation's built-in services."

Leading Models Validated for VMware Cloud Foundation

VCF empowers enterprises to accelerate AI workload deployment at lower costs through an open and extensible ecosystem. Support for mixed compute across AMD, Intel, and NVIDIA frees enterprises to choose their preferred GPU and CPU hardware for their AI workloads. Leveraging vLLM as the default model runtime gives customers the ability to run more than 150 open source models on VCF. Today, Broadcom is announcing the following models have been tested and validated to run on VCF:

- **Nemotron 3:** The NVIDIA Nemotron 3 family of open, multimodal models delivers leading accuracy and efficiency to help agents complete tasks faster. Combining hybrid Mamba-Transformer MoE architecture, 1 million context and multi-environment reinforcement learning, Nemotron 3 enables scalable, long-running agentic workflows across enterprise applications.
- **Gemma 4:** Google DeepMind's latest open source, open-weight multimodal model family, purpose-built for developers and the research community for bringing local execution, and enabling enterprises to build and deploy autonomous AI agents.
- **cotomi:** NEC's proprietary AI model optimized for Japanese language, trained on curated, highly reliable datasets. It empowers enterprises by seamlessly combining high-speed processing with a 40% improvement in token efficiency.
- **Qwen 3.7-Max:** Alibaba's Qwen 3.7-Max is a proprietary multimodal model that offers impressive one-million-token context windows, advanced multimodal reasoning, and agentic-era design, giving global enterprises sovereign, on-premises access to one of the world's most capable AI model families.
- **GLM 5.2:** Z.ai (formerly Zhipu AI)'s open source General Language Model enables enterprises to deploy coding and reasoning agents locally for multi-step autonomous workflows with data sovereignty and optimal hardware performance.

Partner and Industry Commentary

"VMware Cloud Foundation gives enterprises the secure, governed foundation they need to put Gemma 4 to work across their most demanding workloads," said Olivier Lacombe, Director of Product Management, Google DeepMind. "By combining Gemma 4's multimodal reasoning and agentic capabilities with VCF's zero-trust security architecture and Model Context Protocol support, enterprises can build and deploy autonomous AI agents entirely within their own infrastructure without compromising on data sovereignty or operational control."

"NEC's cotomi was purpose-built for enterprises that need AI to truly understand the nuances of Japanese language and business context, and VMware Cloud Foundation gives those enterprises the private, secure infrastructure to deploy it at scale. With cotomi running on VMware Cloud Foundation, organizations gain an AI model capable of acting as a secure autonomous agent across their workplace systems — without sensitive data ever leaving their own environment," said Akio Yamada, Chief AI Officer, NEC Corporation. "Furthermore, NEC offers the NEC Private Infrastructure powered by VMware service, providing customers with a private cloud environment hosted in NEC data centers. We will continue to explore opportunities to further integrate cotomi with VMware Cloud Foundation."

"Enterprise customers need AI inference that respects data sovereignty and regulatory boundaries, particularly in markets where data sovereignty, not just residency, is non-negotiable," said Craig McLellan, CEO of ThinkOn. "By deploying advanced, enterprise-grade AI models on VMware Cloud Foundation, ThinkOn delivers a sovereign AI framework that gives organizations complete control over their intellectual property. Customers gain production-ready inference capabilities on private cloud infrastructure, with the governance and compliance controls their industries require."

Additional Resources

- Learn more about [VMware Cloud Foundation](#)
- Follow VMware Cloud Foundation social channels on [LinkedIn](#), [X, formerly known as Twitter](#) and [YouTube](#)

About VMware Explore

VMware Explore is the established cloud event for IT professionals to advance their skills and credentials. VMware Explore 2026 will connect attendees with technology experts and solution architects who are solving real-world challenges. Focused technical sessions, Hands-on Labs, and certification opportunities will enable attendees to apply their learnings and deliver greater value back to their organizations. Attendees will leave with the training, tools, and strategies to build and operate a modern private cloud that is AI-native—running any workload while keeping data local and secure. To learn more about VMware Explore, please visit: <https://www.vmware.com/explore>

About Broadcom

Broadcom Inc. (NASDAQ: AVGO) is a technology leader that designs, develops, and supplies semiconductors and infrastructure software for global organizations' complex, mission-critical needs. Broadcom combines long-term R&D investment with superb execution to deliver the best technology, at scale. Broadcom is a Delaware corporation headquartered in Palo Alto, CA. For more information, visit www.broadcom.com.

Broadcom, the pulse logo, and Connecting everything are among the trademarks of Broadcom. The term "Broadcom" refers to Broadcom Inc., and/or its subsidiaries. Other trademarks are the property of their respective owners.

Media Contact:

Roger T. Fortier

VMware Cloud Foundation Division, Broadcom

+1.408.348.1569

roger.fortier@broadcom.com