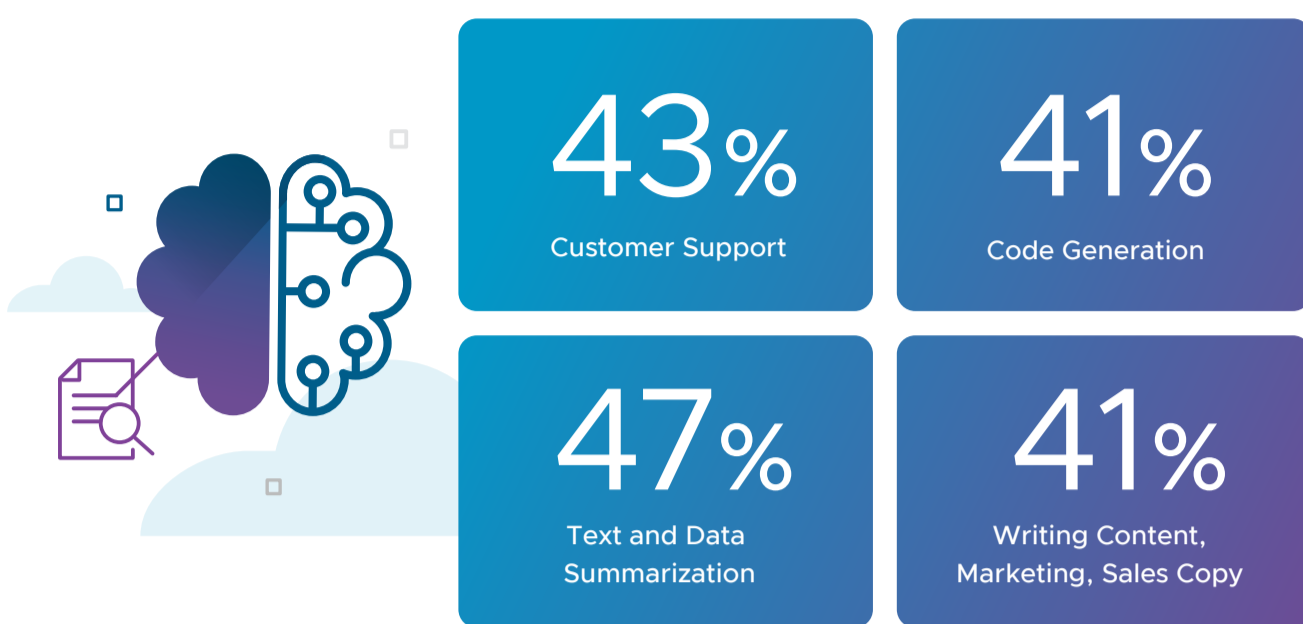**vmware®** by Broadcom

Powering Private AI with Virtualization

Virtualization Accelerates Private AI, Reduces Costs, and Lowers TCO

Generative AI Is Transforming Everything

In a recent survey[1] of 1700 enterprises, 80 percent reported evaluating GenAI use cases, such as AI applications for customer support or content creation for sales, marketing, and other functions. While much of the media spotlight remains on large language models (LLMs), most organizations require AI models that are fine-tuned and customized with their domain-specific data to meet their unique business needs. This is exactly what VMware Private AI offers.

Top Use Cases[2] for Private AI

43% Customer Support	**41%** Code Generation
47% Text and Data Summarization	**41%** Writing Content, Marketing, Sales Copy

Key Drivers for Private AI

Privacy and Control
Organizations maintain physical control of their data and run AI models adjacent to their existing data sources. No risk of data leakage, whether real or perceived.

Cost
Organizations can achieve cost savings with Private AI environments compared to public cloud AI services. With open-source models and managing their own AI infrastructure, organizations gain the benefit of a predictable cost model versus public AI token-based billing.
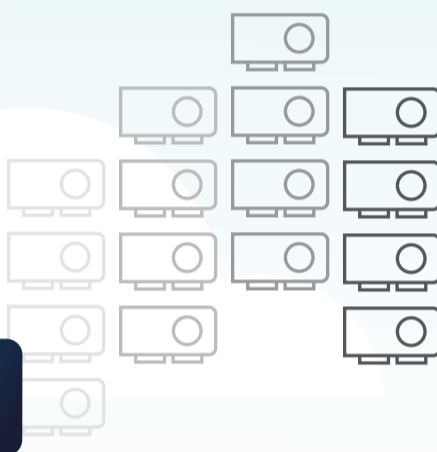
Compliance
Organizations in different industries have different compliance needs that enterprise solutions, including generative AI, must meet. Private AI allows greater access control and audit readiness when deploying a GenAI model.

Performance
AI model training and data queries can create substantial demand surges on networking infrastructure due to the enormous data sets of LLMs. Private AI enables organizations to design their AI models for their specific performance expectations.

VMware Private AI leverages virtualization to transform a single AI GPU into multiple virtual AI GPUs.

What Is Virtualization?

Virtualization is a software capability that allows organizations to place and run multiple virtual machines (VMs) on a single physical server. It is the technology secret sauce that allows organizations to do far more with much less hardware.

SECRET SAUCE

VMware Private AI

VMware Private AI is an architectural approach that balances the business gains of AI with the privacy and compliance needs of an organization. VMware's approach ensures privacy and control of enterprise data, choice of open-source and commercial AI solutions, optimum cost, performance, and compliance.

VMware Private AI significantly reduces the number of GPUs needed for a Private AI installation through the use of virtualization.

- Gain the benefits of AI without sacrificing privacy or control of data
- Take AI models to their data, allowing AI to run anywhere they do business
- Leverage a single AI platform to run whatever AI models or services they choose whether they are commercial or open source

> "Our goal is to help our customers move fast in their digital transformation to position themselves for success. We remain steadfast in our decision to focus our resources on R&D, and continuing to develop a true, seamless private cloud experience for customers – one that is competitive with the public cloud."[3]

Hock Tan
President and CEO, Broadcom

Benefits of VMware Private AI

Flexibility of Choice
Allows organizations a range of AI solutions to best fit their needs. VMware's platform approach future-proofs AI investments.

Privacy and Control
Allows enterprises to run AI services adjacent to their data — wherever it may be.

Intelligent Resource Sharing
Enables organizations to place and balance workloads across their AI infrastructure to ensure performance at the lowest possible cost.

Unified Operations
Allows organizations to use the same tools and processes to manage and operate AI and non-AI applications, further reducing costs and complexity.

Optimum Cost
Empowers enterprises to achieve the most cost-effective solutions for their generative AI models

Compliance and Governance
Uses popular open-source tools to securely govern AI models and data to eliminate data leakage, enforce enterprise access controls, and achieve cost-effective generative AI solutions with full regulatory compliance.

VMware-enabled Virtual Infrastructure Significantly Lowers Costs[4]

55%
Infrastructure Cost Savings

31%
Facilities Cost Savings

42%
Labor Productivity Cost Savings

51%
TCO Total (3 yrs) Compared to Traditional Infrastructure

[1] The Cube Research. Examining the VMware Private AI Foundation with NVIDIA Vision. April 15, 2024.
[2] VMware News & Stories: A changing market landscape requires constant evolution: our mission for VMware customers April 15, 2024.
[3] Analysis and TCO results from VMware Cloud Economics Team, 2024

Learn more at vmware.com/privateAI

vmware® by Broadcom