



Your data. Your agents. AI on your terms

It's easy to prove AI works. It's hard to make it work every day.

Across industries, enterprises have no shortage of AI pilots, proofs of concept, and experimentation. What they struggle with is turning that momentum into production systems that are fast, governed, and worth scaling. The bottleneck is rarely the model alone. It is everything around it: infrastructure, data access, orchestration, operations, compliance, and the ability to run AI reliably where the business needs it.

That challenge is becoming more urgent as the center of gravity in enterprise AI shifts to inference. Inference is where AI creates value: inside copilots, agents, search, automation, customer interactions, and operational workflows. It is also where latency, cost, governance, and control become real business issues.

This is exactly why private AI infrastructure is gaining traction. IDC points to rising enterprise demand for private AI environments that give organizations more sovereignty and control across the AI lifecycle, including both training and inference, especially for sensitive and regulated workloads.¹ For enterprises trying to operationalize AI, the question is no longer whether models are powerful enough. It is whether the platform behind them is production ready.

HPE Private Cloud AI is built for this moment.

Why AI progress stalls

Many AI initiatives lose momentum before delivering business value. It's not from a lack of imagination, but from the lack of a production ready foundation.

Four challenges show up repeatedly:

- AI pilots create excitement, but many stall before production because the infrastructure, data pipelines, and governance needed to operationalize them are not yet in place.
- Privacy, regulatory, and sovereignty requirements make it difficult to rely on cloud-first approaches when enterprises need tighter control over where data lives and how it is used.
- Limited in-house expertise across AI infrastructure, GPUs, and MLOps makes it harder to design, operate, and scale AI environments, adding risk, delay, and hidden people costs.
- Too many pilots are built on a patchwork of components and point solutions that make management, governance, and scale next to impossible in production.

These challenges are why so many promising pilots never become durable AI use cases. Enterprises do not need more disconnected pieces. They need a simpler, faster path to production.



¹ ["IDC Spotlight: The Great VM Reset,"](#) April 2026.

Inference-first platform for production AI

HPE Private Cloud AI, co-engineered with NVIDIA, is an enterprise platform for agentic AI, purpose built for production inference. It goes beyond simply hosting models, providing always on AI services, an integrated AI operations layer, and operational simplicity so you can move from pilots to production on a single, controlled platform.

This approach aligns with what IDC highlights in private AI infrastructure: tightly integrated, turnkey systems that reduce complexity and make it easier to operationalize AI without stitching together every layer yourself. IDC recognizes HPE as a Leader in private AI infrastructure systems, describing HPE Private Cloud AI as tightly integrated, fully turnkey, and requiring minimal on-site integration to get from rack to running.²

Why PCAI, in five points:

- 1. Built for inference, where enterprise AI creates value:** HPE Private Cloud AI's inference optimized foundation keeps latency, cost, and performance in check as usage scales, so teams avoid constantly redesigning the runtime environment.
- 2. Faster from pilot to production:** A turnkey stack with validated blueprints, self-service tools, and a single console lets teams deploy, orchestrate, monitor, and scale AI workloads—replacing months of custom integration with a standardized path from Day 0 through Day 2 operations.
- 3. Simpler to start, simpler to operate:** Build it yourself AI projects are replaced by a single platform that unifies development tools, orchestration, inference runtimes, governed connectors, and lifecycle management, so you begin with a usable production foundation instead of assembling your own.

² "IDC Spotlight: The Great VM Reset," April 2026.

4. Keep AI close to data and control: Running close to sensitive data and enterprise systems reduces the need to move or copy information into external services, especially critical for regulated industries and any organization with strong sovereignty, residency, or internal control requirements.

5. Simplified AI lifecycle: An integrated platform provides a faster starting point and cleaner model for running AI—letting you build and connect agents to data, tools, and APIs; deploy, monitor, and manage AI lifecycles; apply consistent policies, logging, and audits; and add new models and tools from a single control plane. That means teams do not have to rewire the platform every time requirements change.

The takeaway

HPE Private Cloud AI gives enterprises a sharper way to operationalize AI: an inference-first platform that simplifies how they get started, reduces the burden of integration, keeps AI close to data and control, and helps turn promising pilots into scalable production systems.

Learn more at

[HPE.com/private-cloud-ai](https://hpe.com/private-cloud-ai)

Visit [HPE.com](https://hpe.com)

[Chat now](#)

© Copyright 2026 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

a00151659ENW. Rev. 1

HEWLETT PACKARD ENTERPRISE

hpe.com

