Avoiding AI Failure: A Proven Infrastructure Strategy for Sustainable Business Results

We often hear the analogy that AI is a race, that early adopters of GenAI gain a first-mover advantage and that companies fear being outpaced by competitors. There's a lot of truth to this. We've all watched retail giants like Amazon and Walmart harness GenAI to transform their businesses and set the pace for the rest of the world. It's no wonder the number of GenAI use cases in production doubled between October 2023 and December 2024. Nearly every business wants to harness the power of GenAI.

An Inconvenient Truth of GenAI

AI's influence on the economy, productivity, and security comes with endless projections and bold predictions. But behind all the excitement and race-day metaphors lies a sobering reality: Many AI projects never even make it to the finish line, much less the first lap. In fact, Gartner predicts that by the end of 2025, at least 30% of generative AI projects will be abandoned after the proof-of-concept stage.[2] So while the race for AI dominance is real, a significant number of companies never get out of the starting blocks. This, despite all their early enthusiasm and investment.

Why AI Initiatives Stall Before Reaching the Finish Line

Many companies are hesitant to even get started with AI due to confusion about how to begin. For those that do, the metaphor that things are a lot harder than they look is certainly true for advanced AI implementation. The reasons why so many organizations stall out in the race are many:

- **Infrastructure limitations:** Existing systems often lack the capacity or readiness for AI integration.
- **Specialized talent shortages:** The demand for AI expertise far exceeds supply, driving up costs and making skilled professionals difficult to find.
- **Unexpected pilot-stage obstacles:** Initial AI deployments often require more time, resources, and adjustments than anticipated.
- **Regulatory and security complexities:** Organizations must navigate evolving compliance standards and security vulnerabilities they hadn't accounted for.

The number of GenAI use cases in production doubled between October 2023 and December 2024.[1]

EP 45: Moving from Concept to Outcomes with HPE Private Cloud AI
LISTEN NOW

Hewlett Packard Enterprise

- **Rapid Development Pace:** The rapid acceleration and development of AI make it difficult for companies to stay up to date.

Business leaders find themselves caught between the fear of missing out on AI's potential and the inability to justify continued investment based on concrete business outcomes. Ambitious visions can quickly give way to frustration. But what if there were a way to bypass a lot of the prequalification race? What if there were a way to cut ahead of your competitors so that you could start leveraging GenAI?

Turning AI into a Turnkey Solution

Introducing HPE Private Cloud AI (PCAI), a turnkey solution designed to fast-track AI adoption and deliver immediate business impact. Developed through a strategic and collaborative partnership between HPE and NVIDIA, PCAI combines the expertise of two industry leaders to deliver a pretested, scalable, and AI-optimized solution. This joint initiative provides businesses with a tailored AI infrastructure that allows for seamless deployment and accelerated performance from day one.

With PCAI, organizations receive a ready-to-deploy system that arrives as a complete pretested and preconfigured unit. All components are fully integrated, validated, and optimized before delivery so there is no setup time involved. HPE and NVIDIA do all the heavy lifting beforehand to ensure that customers receive best-in-class performance from the outset.

Cut to the Finish Line with PCAI

AI doesn't just happen on its own. It requires infrastructure made up of specialized hardware, networking and storage that is designed for AI. It is a collective process that can require months of planning, procuring, deploying, testing, fine-tuning, and scaling.

It doesn't have to be that way though. PCAI lets your team bypass all this complexity and vastly reduces the timeframe so that the marathon is compressed into a sprint. Your rack arrives ready for action, allowing your business to focus on developing and scaling AI applications rather than troubleshooting infrastructure. With PCAI, your organization doesn't just enter the

competitive AI landscape. Your implementation is launched with a decisive head start that accelerates time-to-value and delivers measurable progress while competitors are still assembling their infrastructure.

The Ease of Cloud Manageability

Even though PCAI is deployed on-premises, it is managed through a cloud-native interface. This means you get the best of both worlds: security, control, and performance of on-prem infrastructure, combined with the ease of management typically associated with the cloud.

Cloud manageability is delivered through the HPE GreenLake platform, which provides intuitive self-service interfaces that enable organizations to manage AI workloads, infrastructure resources, and critical operations with the same ease and flexibility that enterprises have come to expect from modern cloud environments. With automated updates, centralized management, and seamless scalability, PCAI provides a proven, reliable path to production from day one, empowering your teams to start delivering tangible AI results ahead of the competition.

Start Small and Scale Later

GenAI is an uncharted territory for businesses of all sizes, a blend of promise and complexity that leaves many business leaders navigating without a roadmap. HPE PCAI eliminates this uncertainty by offering a pragmatic starting point that grows with your ambitions. If you're familiar with hyperconverged infrastructure, you'll appreciate that PCAI is built on the same modular foundation. Begin with a single rack and easily expand by adding additional modules as your AI initiatives grow. In fact, HPE has made the process of sizing as simple as purchasing a shirt from an online retailer as you can choose from small, medium, large and extra-large. The chart below outlines the different options available for each size.

Resources	Small	Medium	Large	Extra Large
Compute	4 or 8 NVIDIA L40S GPUs	8 or 16 NVIDIA L40S GPUs	16 or 32 NVIDIA H100 NVL GPUs	12 or 24 NVIDIA GH200 NVL2
Storage	30 TB to 248 TB	109 TB to 390 TB	670 TB to 1088 TB	670 TB to 1088 TB
Networking	100GbE NVIDIA Networking	200GbE NVIDIA Networking	400GbE NVIDIA Networking	800GbE NVIDIA Networking
Power	up to 8 kW rack	up to 17.7 kW	up to 25 kW x 2	up to 25 kW x 2

Small configuration designs are tailored for GenAI inferencing while larger configurations that support advanced use cases like Retrieval-Augmented Generation (RAG) and model training. The combination of PCAI's modular architecture and HPE GreenLake's cloud management platform enables organizations to begin with targeted pilot projects and scale intelligently as AI requirements evolve and business value becomes apparent.

Only Pay for What You Need

As many companies have learned in recent years, the cloud is not a destination. Instead, it represents another way to consume technology. While you can still purchase your PCAI solution outright, HPE also offers PCAI as a Consumption-as-a-Service model, allowing you to pay only for what you use and scale up as your needs grow. This pay-as-you-go approach is ideal for organizations that have variable capacity requirements and want to scale in sync with their business demands. Those with a defined capacity can choose to lease their solution. It's one more example of the flexibility that HPE PCAI is built around.

While the inclination may be to go to a public cloud, a private cloud often means more sense. While the public cloud does provide consumption pricing, it is complicated by spot pricing, data transfer fees, storage costs and

instance type. Thanks to its dedicated infrastructure, HPE Private Cloud AI provides a more predictable cost structure, especially for heavy AI workloads. In fact, a study found that even a small PCAI deployment could yield $700K in savings over three years compared to public cloud, alongside a 49% lower GPU hourly cost.[3] For large deployments the savings went as high as $6.1 million along with a 76% lower GPU hourly cost.

The Best of Both Worlds

HPE Private Cloud AI (PCAI) delivers the best of both worlds by combining the agility and user experience of the public cloud with the security, control, and compliance of an on-premises solution. It is a true hybrid approach that allows both rapid AI adoption and data sovereignty at the same time. The chart below highlights the advantages of PCAI's cloud and on-prem approaches:

Cloud Advantages	On-Prem Advantages
Cloud-Managed simplicity	Cost predictability
Eliminate data silos with a global namespace	Customization and control
On-Demand scalability	Sensitive data never leaves your premises
Provides self-service access to AI tools	Simplifies regulatory compliance like GDPR, HIPAA
Includes managed services including 24/7 reactive support	Low latency networking for applications

Use Cases for HPE PCAI

Business leaders are eager to harness AI. Even in its early market presence, PCAI is already proving its impact across industries, helping organizations accelerate AI adoption while maintaining control over costs and security.

- With self-serve access to essential AI tools, PCAI can boost developer productivity by up to 90%, enabling teams to build, test, and deploy AI models far faster than with traditional approaches.[4]

- PCAI delivers up to 50% total cost of ownership (TCO) savings across small, medium, and large AI infrastructure configurations compared to do-it-yourself (DIY) solutions, thanks to reduced engineering, operational, and software licensing costs.[5]

- HPE's own deployment of PCAI reduced executive reporting cycles by approximately 50% and cut processing costs by an estimated 25%.[6]

Many companies are already leveraging AI for a variety of applications, demonstrating the technology's versatility across business functions. Early adopters are implementing AI solutions for customer service automation through intelligent chatbots, predictive maintenance programs that prevent costly equipment failures, fraud detection systems that identify suspicious transactions in real-time, and personalized recommendation engines that enhance customer experiences.

Specific industry examples include financial institutions that are using AI loan application review automation, algorithmic trading, and enhanced risk management while healthcare organizations deploy AI for diagnostic imaging and improved operational efficiencies. Manufacturing companies leverage AI for quality control and supply chain optimization, and retail businesses utilize AI for inventory management and dynamic pricing strategies. Even public sector institutions are using advanced AI tools to create intelligent customer service experiences and content creation.

Existing Tool Integration

Another key advantage of PCAI is its extensibility, enabling organizations to deploy and operationalize AI applications such as virtual assistants and generative AI models in mere minutes. This is in stark contrast to the months that are typically required for assembling and integrating disparate components.

- PCAI includes NVIDIA's AI Essentials, a comprehensive suite of pre-configured tools and frameworks specifically optimized for AI workloads. This allows organizations to immediately leverage NVIDIA's powerful AI capabilities without the need for extensive setup and configuration

- PCAI also includes HPE AI Essentials, which offers a range of open-source tools and enterprise-grade software to handle everything from data engineering and analytics to visualization and model deployment

- PCAE integrates with Kubernetes, allowing developers and data scientists to leverage their existing Kubernetes infrastructure and tools while implementing AI solutions.

- PCAI integrates with HPE's Esmeral Data Fabric, which provides a unified data platform that can handle various types of data from different sources across hybrid environments. This integration allows companies to create a data lake where they can store and manage their data more efficiently.

WEI Provides the Customization Element

While preconfigured pretested rack solutions provide turnkey access to the transformation power of GenAI, these solutions won't deliver their full potential unless properly sized for your business and its objective needs. This is where WEI's expertise becomes invaluable.

WEI specializes in helping organizations harness advanced analytics, big data, IoT, and cloud computing technologies to extract real-time insights and enable agile decision-making. Rather than leaving you to navigate complex AI implementation alone, we invest time in understanding your unique business challenges and AI aspirations.

With extensive experience deploying AI solutions across diverse industries, we offer proven methodologies and cross-industry insights that help your organization accelerate AI adoption and avoid common pitfalls. By partnering with us, you can confidently move forward with AI initiatives, leveraging best practices to achieve faster, more reliable results.

Conclusion

While alternative solutions may offer select features, none match HPE PCAI's comprehensive, fully curated stack combined with a secure, cloud-like experience for your data. Developed through the collaboration of industry-leading innovators, PCAI delivers a turnkey AI infrastructure that enables you to start leveraging inference AI in minutes, not months. To discover how you can gain a competitive edge and accelerate your AI journey, contact WEI today and put your organization at the forefront of innovation.

Resources:

1. Survey: Generative AI's Uptake Is Unprecedented Despite Roadblocks | Bain & Company

2. Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025

3. AI that pays for itself: real cost savings with HPE Private Cloud AI

4. HPE Private Cloud AI: A strategic necessity for enterprises

5. AI that pays for itself: real cost savings with HPE Private Cloud AI

6. HPE gains enterprise traction with private cloud, AI-powered servers | CIO Dive

About WEI

WEI is an innovative, full service, customer centric IT solutions provider.

Why WEI? Because we care. We go further.

WEI is an expert in business technology improvement, helping clients optimize their technology environments and work efficiently. WEI works with clients to understand goals, integrate strategy with technology solutions, and leverage their current IT environment into one company-wide model to increase utilization and efficiencies around their unique business processes.