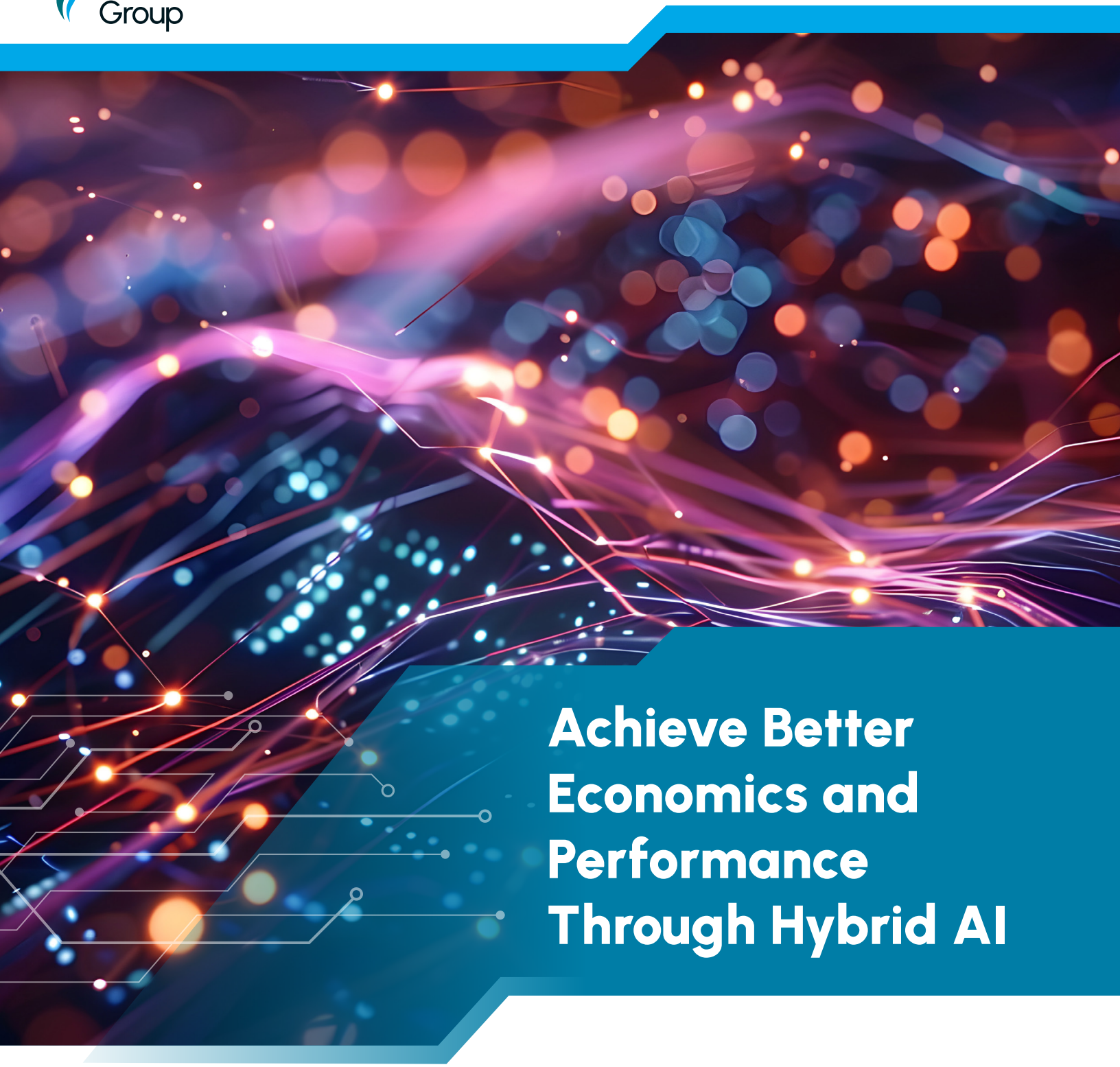
TheFuturum Group



Achieve Better Economics and Performance Through Hybrid AI

AUTHOR

Steven Dickens
Chief Technology Advisor | The Futurum Group

Ron Westfall
Research Director | The Futurum Group

IN PARTNERSHIP WITH

Lenovo | NVIDIA

AUGUST 2024

Executive Summary

Today's IT and business decision makers (ITBDMs) are overseeing the unprecedented task of guiding their organization's AI journey. The choices are broad, and the speed of innovation is intense. In exploring the best path for the AI journey, we see decision makers are increasingly choosing a hybrid AI model as the best overall course. Why hybrid AI?

Hybrid AI is the optimization or extension of AI models based on machine learning, deep learning, neural networks with human expertise to create domain-specific AI models with enhanced accuracy or AI use case-specific models with maximum accuracy or prediction probability. Hybrid AI is the idea of adhering to data gravity and bringing AI solutions and capabilities to where you need it - from mobile and desktop devices, private data centers, edge computing or even the cloud. As such, the use-case and data management strategy may dictate where and how to deploy AI capabilities, with consideration to factors such as security, cost, network infrastructure, data gravity and reliability.

We see that this approach overcomes limitations of single-model or single-infrastructure methods, aligning with customer demands for flexible data management and application support. Hybrid AI integrates open-source AI and closed-source capabilities, such as OpenAI's ChatGPT capabilities, boosting productivity, agility, data protection, and AI democratization.

For key decision makers -including CIOs, CTOs, Chief Data Officers, Chief AI/Data Officers, and Architects, the initial step involves defining problems, identifying data sources, and selecting use cases, AI applications, with the right mix of infrastructure and right-sized models. This process typically includes strategy development, building an AI stack, an AI factory, or center of excellence (COE), and deploying chosen solutions.

Implementing hybrid AI requires evaluating, piloting, and scaling emerging technologies with a broad partner ecosystem to be able to design a solution that has the right combination of AI techniques, the right architecture, and the right model for your business.

We examine why the Lenovo Smarter AI for All portfolio and vision can meet the unique demands of selecting and implementing hybrid AI organization wide. Lenovo Smarter AI for All prioritizes business outcome driven approach by providing the right mix and size of AI PCs, Workstations, AI-ready solutions, expertise, and AI-optimized storage, compute,

infrastructure, software, and partnering capabilities with a validated partner ecosystem. This includes the ability to deliver tailored, agile, scalable, and energy-efficient industry and use-case based AI solutions across personal and industry edge (PCs, workstations, business locations), data centers (CoLos, on-prem), and high-performance computing (HPC) environments, all with public cloud integration/ synergies.

Together, we found that Lenovo and NVIDIA build AI factories that provide an AI-optimized and enabled enterprise AI environments where AI applications can be developed, deployed, and managed securely, privately and at scale.

The Fast-Evolving Hybrid AI Landscape

Organizations and their employees are eager to harness the transformative power of Generative AI (GenAI), moving beyond initial experiments and AI/ML pilots to leverage their own data. However, widespread adoption is often hindered by legal, compliance, and governance challenges, expertise gaps, data silos, security concerns, inadequate AI infrastructure, and ecosystem complexity. Budget constraints further complicate matters. While 96% of CIOs anticipate increased AI investments in the coming year, only 20% expect overall IT budgets to grow by more than 10% (CIO Study 2024).

Hybrid AI is emerging as a crucial component of business AI strategies, enabling organizations to develop AI and data solutions that align with their goals and bring AI closer to data sources, real-time interactions, and experience creation points. Although free AI tools, public foundational models, and existing cloud subscriptions facilitate initial pilots, many organizations find themselves trapped in endless proof-of-concept cycles and escalating costs. This is often due to the lack of a robust Hybrid AI strategy that integrates AI across private and public cloud environments, absence of customized AI models, limited access to quality organizational data, and inadequate AI compute sizing and planning.

From our perspective, running GenAI and AI locally at the edge, including on devices, is becoming critical for organizations with distributed data and edge locations (e.g., retail, manufacturing, hospitals, hybrid workplaces), those requiring real-time analytics and access, and entities with disconnected compute needs (e.g., remote oil & gas operations or highly protected edge environments).

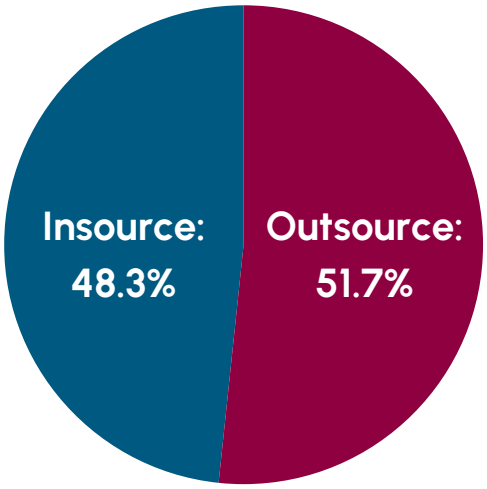
Edge locations and devices generate vast amounts of data, most of which remains unanalyzed. AI and GenAI present a significant opportunity to leverage this untapped resource. With AI applications developed for edge use cases, increasingly efficient smaller models, and AI-enabled devices like AI PCs, edge computing becomes a reality. This eliminates the need to transfer all data or AI models to public clouds or centralized data centers for real-time analysis, addressing security, latency, and bandwidth limitations.

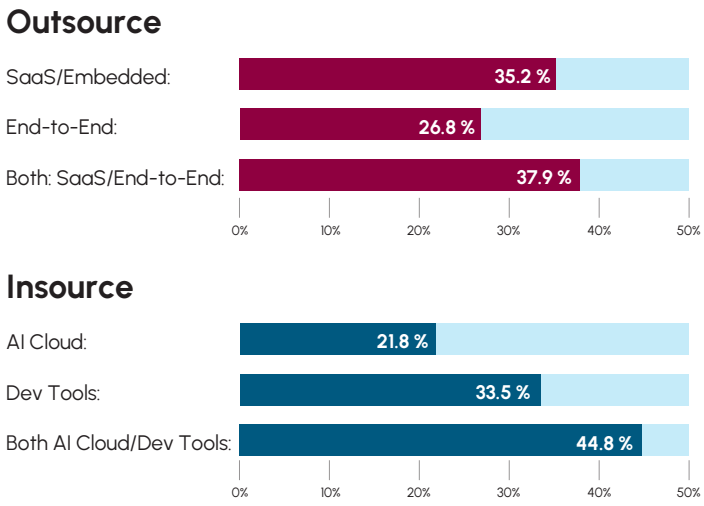
Local GenAI processing allows for data analysis at the point of creation, reducing latency and improving efficiency—crucial for real-time applications such as manufacturing processes and robotics. Additionally, privacy concerns often necessitate local processing to ensure sensitive data remains on-device.

Cloud-based Large Language Models (LLMs) present challenges that organizations increasingly seek to avoid. These include the risk of hallucinations—logically inconsistent but plausible-sounding outputs—and the propagation of inaccuracies from training data. Furthermore, organizations have limited control and visibility over cloud-managed services' infrastructure and implementation, raising security concerns.

A notable example occurred in March 2023 when OpenAI's ChatGPT experienced an outage due to a vulnerability in an open-source library, potentially exposing customer payment information. This incident underscores the importance of rigorous LLM safeguard testing and robust cyberthreat defense mechanisms when utilizing cloud-based AI services.

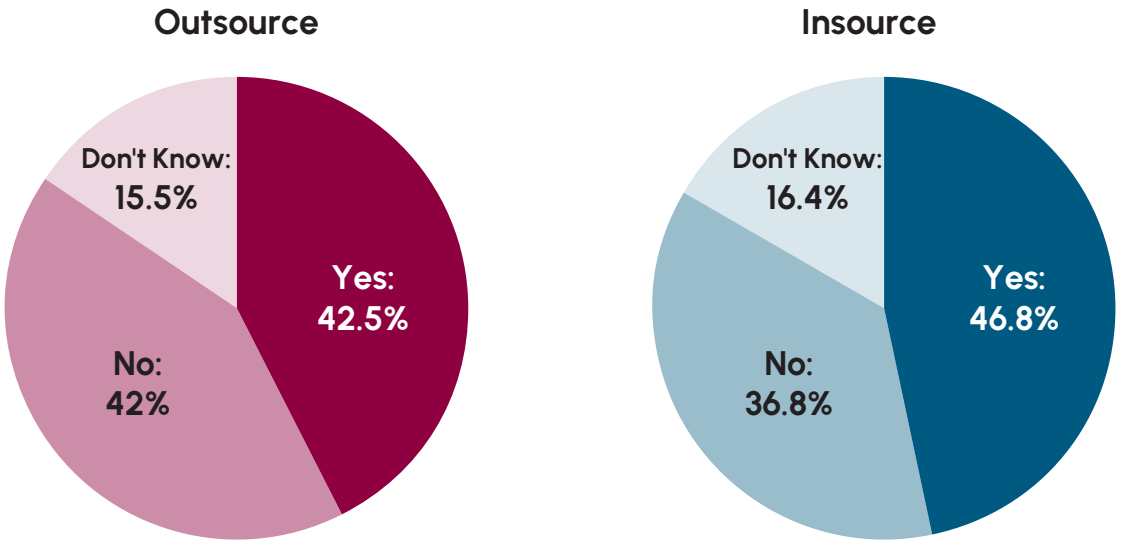
Furthermore, we see that organizations are taking a balanced approach on how they deploy AI services:





Source: The Futurum Group

The rapidly growing interest in hybrid AI is creating new competitive dynamics as organizations are increasingly exploring plans to change or add new vendors (2023-24):



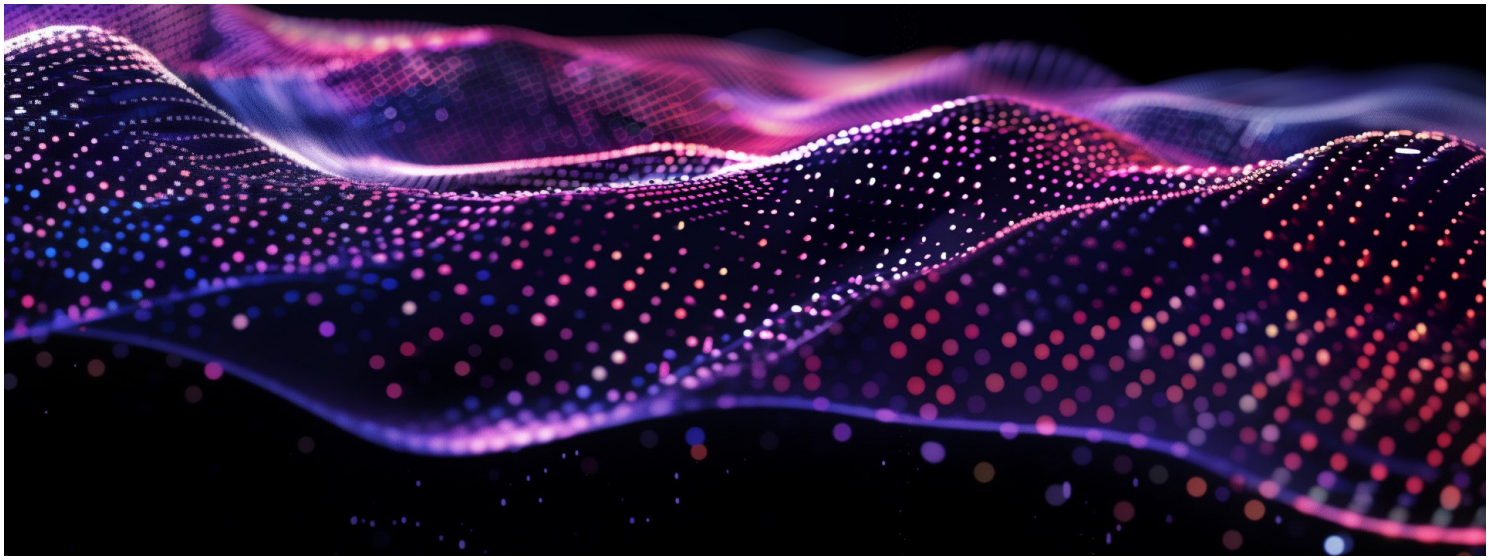
Source: The Futurum Group

The success of hybrid AI hinges on the critical role played by ecosystems in curating AI solutions tailored to organizational needs. Ecosystems unite a diverse array of value chain stakeholders, enabling the seamless integration of AI systems across hybrid environments, leading to the creation of made-to-order AI solutions. AI solutions frequently require a combination of technologies and expertise, reliant on ecosystems to provide access to the wide array of software, applications, tools, and capabilities to develop solutions that can directly address business challenges as the AI journey unfolds.

Moreover, Hybrid AI's inherent flexibility is essential in meeting the diverse needs of the various personas involved in AI within enterprises. AI-application developers and data scientists prefer to use familiar AI tools and frameworks, regardless of their distribution. At the edge, industries, such as retail stores and manufacturing, expect and require real-time response from AI systems alongside the ability to process data swiftly.

However, the effective use of hybrid AI requires meticulous planning and execution, considering technology, people, processes, and Responsible AI approaches. Businesses grapple with siloed initiatives, misaligned AI solution choices and budgets, while IT teams struggle to modernize infrastructure and meet diverse AI compute demands, compliance, security standards and the need to run a hybrid operation. As data management and public cloud use become uncontrolled, security, governance, and escalating cloud costs emerge as pressing concerns. These challenges must be balanced against the skills and change management required to achieve meaningful business outcomes from AI.

For instance, data scientists and ML engineers should not spend most of their time on data quality, regulations, and compliance. Businesses should involve IT&O teams in AI-related IT decisions. Similarly, a bottom-up approach by IT teams to introduce AI to the enterprise may result in over or under-provisioning without a thorough understanding of business needs and use cases.



Lenovo and NVIDIA's Co-Engineered Portfolio

The AI industry's growth is driven by multiple factors, with the increasing demand for automation and optimization across sectors being paramount. As outlined in our paper, companies are harnessing advancements in computing power and cloud infrastructure to implement more efficient AI applications. This shift has led to a surge in requests for customized AI solutions, including consulting, design, deployment, and maintenance services that can help them simplify that process.

Lenovo Smarter AI for All portfolio addresses these evolving needs with AI-enabled devices, smart infrastructure, and robust cloud solutions, facilitating seamless AI integration across various platforms. The strategic partnership between Lenovo and NVIDIA is accelerating innovation in the AI sector. Their co-engineered portfolio simplifies AI adoption, delivers faster outcomes, and enhances performance and efficiency in hybrid AI environments, further propelling industry growth.

Important Information About this Report

CONTRIBUTORS

Steven Dickens
Chief Technology Advisor | The Futurum Group

Ron Westfall
Research Director | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT LENOVO AND NVIDIA

Lenovo brings the new era of AI-powered innovation to everyone. Our full-stack portfolio delivers powerful, flexible, and responsible AI solutions to transform industries and empower individuals. We create a future of Smarter AI for all. At Lenovo, we believe the future of AI involves the co-existence of public and enterprise AI. Lenovo brings AI to you and your data.

In partnership with NVIDIA, hybrid AI solutions are purpose built through engineering collaboration to efficiently bring AI to customer data, where and when users need it the most, advancing Lenovo's vision to enable AI for all and delivering time to market support of breakthrough technologies and architecture for the next generation of generative AI. Lenovo hybrid solutions, already optimized to run NVIDIA AI Enterprise software for secure, supported and stable production AI, also provide developers access to NVIDIA microservices, including NVIDIA NIMs and NeMo Retriever.

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION

The Futurum Group LLC | futurumgroup.com | (833) 722-5337 |