

Fast-track AI workflows with Lenovo & NVIDIA® Accelerated Solutions

A complete guide for enterprises wanting to leverage the power of generative AI in their private data centers using AI-optimized solutions from Lenovo and NVIDIA.

Lenovo

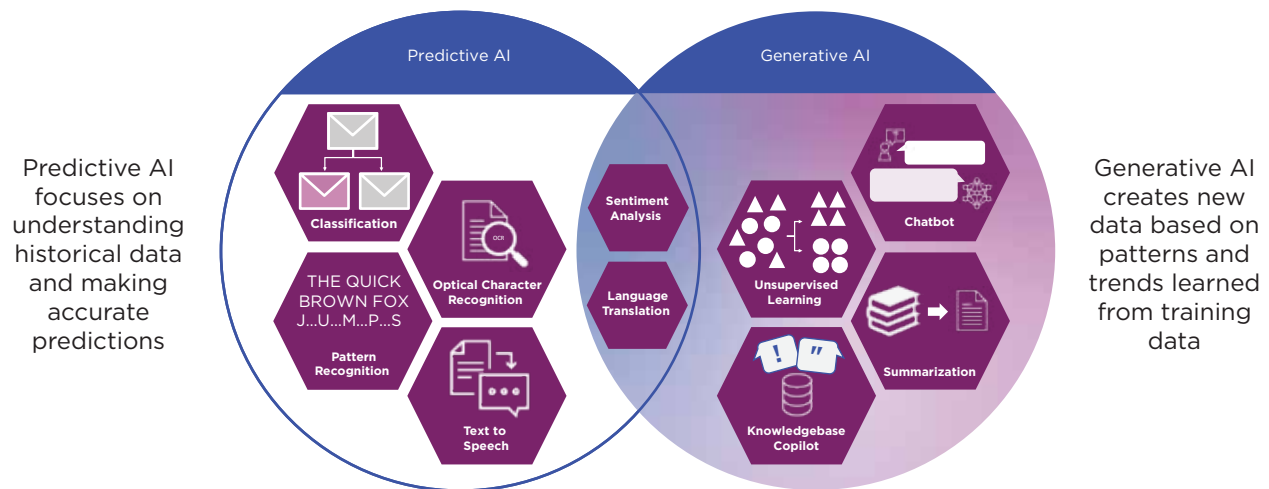
 **nVIDIA**

From concept to reality in a click with generative AI

The digital age has unleashed an explosion of data, with around 330 exabytes produced daily¹ — equivalent to a 75-million-year video call² or almost 300 million years of music.³ This unprecedented data abundance, fueled by big data technologies, the Internet of Things (IoT), and our connected devices, has sparked a rapid and transformative AI (artificial intelligence) revolution. AI models, trained on massive datasets, can now solve complex problems and create new possibilities once unimaginable.

Until recently, AI has primarily performed predictive tasks, such as forecasting future sales, detecting performance irregularities, or analyzing financial markets. However, since late 2022, there has been a dramatic shift with the emergence of generative AI (GenAI).

The buzz and excitement for GenAI stems from its ability to create new content indistinguishable from human work, such as text, images, audio, video, music, art, and code. GenAI demonstrates the potential to transform many industries and aspects of our lives, creating new commercial products and services, supporting new scientific discoveries, and improving everything from education to entertainment and gross domestic product to our overall existence.



One of the key breakthroughs in GenAI has been Large Language Models (LLMs), popularized by Open AI ChatGPT and Google Bard. LLMs are trained on text and code to learn the patterns and structures of human language through machine and deep learning. They are used for writing text, translating languages, composing creative content, and more. The speed and output quality have captured the world's imagination, sparking an incredibly broad range of successful use cases and a spate of media stories bordering science fiction.

The exponential growth of AI

\$150B The global AI industry was valued at \$150 billion in 2023 and is predicted to grow at a compound annual growth rate (CAGR) of 36.8% from 2023 to 2030. ⁴	34.3% 34.3% of workers identify as regular GenAI users across the four industries of financial services (42%), retail (30%), advanced industries (32%), and healthcare (33%). ⁵	\$1.3T By 2032, the GenAI industry will grow to \$1.3 trillion, with hardware the largest segment at \$640bn and software revenue hitting \$280bn. ⁶	70% GenAI has the potential to automate up to 70% of business activities by 2030. ⁷
---	--	---	--

The creative process reimaged





GenAI models use neural networks to analyze existing data and generate new and original content. They are trained using unsupervised or semi-supervised learning, allowing them to leverage large amounts of unlabeled data. Models can learn the underlying probability distribution of the data, allowing them to produce realistic and diverse outputs. They can also be conditioned on a prompt or seed text to create content on a specific topic or style.

Here is a simplified example of how a GenAI model might work to compose text:

- 1. The model is trained on a large dataset.
- 2. The model learns the data’s patterns, structures, and probability distribution.
- 3. The model is given a prompt, which can be text, image, audio, or video.
- 4. The model samples the probability distribution to generate tokens (words, images, code, etc.) as output.

Transformative impacts for every industry

GenAI is already having a significant impact on a wide range of industries, including:

Industry	Application	Use case	Generated outputs
 Financial services	Regulatory compliance	Automating compliance tasks such as data gathering, risk assessment, and reporting	Automatically collect and organize financial transaction and customer behavior data. Assess risk factors and generate compliance reports per regulations. Reduce manual checks and minimize errors.
 Retail	Customer service	Automated chatbots for customer inquiries, complaints, and recommendations	Interact with customers in real time, addressing common questions about products, orders, and services with 24/7 availability, reducing the workload on human customer service agents.
 Manufacturing	Supply chain	Optimization of inventory levels, delivery routes, and vendor management	Predict future inventory needs based on past trends and current demand. Optimize delivery routes using real-time data to reduce fuel consumption and speed up delivery times. Automate vendor communication for ordering and restocking, resulting in a more efficient and cost-effective supply chain.
 Healthcare	Diagnostics and treatment planning	Automated analysis of patient data for early and accurate disease diagnoses	Process and analyze various medical data, including electronic health records, lab tests, and medical imaging, to identify patterns indicative of specific diseases, enabling earlier and more accurate diagnoses. Provide healthcare professionals with personalized data-driven treatment options.

The forerunners in GenAI are gaining a competitive advantage through operational cost-efficiencies, increased sales, faster product development, and better quality control, all enabled by greater productivity, personalization, insight, and automation.



Getting started with GenAI

There are three AI model options, which range in cost, complexity, and value. Organizations will likely use all three types, often multiple instances of each, with models optimized and trained for specific departments, functions, or applications, e.g., a chatbot for customer services, an image generation model for product development, a text generation and translation model for sales and marketing, a code generation model for IT. The base model types are:

- **Option #1 — General purpose model:**

Often referred to as GenAI as a Service, this is an off-the-shelf option based on a pre-trained foundation model and will be used by every company. It is pay-as-you-go and the most straightforward GenAI deployment, but offers little customization and control. Designed for generic use cases, where information is readily available and little organizational context is required. Example: OpenAI ChatGPT/Open AI API or Stable Diffusion.

- **Option #2 — Moderately customized model:**

A foundation GenAI model trained on the company's data; a process referred to as fine-tuning. This option offers more customization and control, with upfront investment for infrastructure and development and ongoing maintenance costs. It is ideal when the GenAI model requires unique data to improve responses to specific needs. Example: A chatbot based on Meta Llama 2 and trained on a company's data, such as Lenovo's Pre-Trained AI Chatbot, with information from Lenovo service manuals, tech specs, and warranties.

- **Option #3 — Extensively customized model:**

A model trained from scratch on a unique dataset tailored to a specific use case, offering complete customization and control. This model will present upfront costs to build and develop. If designed for a particular application, it could benefit from a lower Total Cost of Ownership (TCO) through reduced ongoing compute requirements and overheads. Ideal for unique use cases that fundamentally rely on proprietary data. Example: A drug discovery system trained on proprietary data or a private financial data platform like BloombergGPT.

Once built, organizations can efficiently roll out multiple custom GenAI applications across departments by changing the data source, unlocking benefits for all stakeholders. With a strategic approach to GenAI, organizations will simplify complex tasks, drive automation, amplify innovation, and power productivity, leading to better business outcomes and, ultimately, a competitive edge.



Enterprise innovation with a private GenAI model

The journey starts with data

Data is essential for building a customized GenAI model. GenAI models can have billions of parameters, and private hosting provides confidence in data security and training, allowing organizations to configure data to get the most value. It also gives organizations control over the inference process and model outputs to ensure responsible and ethical usage.

A comparative analysis of private and public models

Organizations looking to take advantage of GenAI can utilize public or private GenAI models or both. Public models are easy to access, available off the shelf, and generally offer text, image, video, or code generation capabilities. Private models are hosted either on a self-hosted or cloud-based platform. These models are often developed using a public foundation model as a base, with custom applications designed to provide tailored outputs and results, increased data controls, and improved security measures.

Feature	GenAI foundational model	GenAI customized model
Data	Trained on public data with pre-built applications	Trained or fine-tuned with proprietary data with pre-built or custom applications
Access	Publicly accessible	Private to the organization that manages it
Cost	Pay as you go	Upfront investment and ongoing development to maintain
Expertise	Does not require AI expertise	Requires AI expertise
Customization	Limited customization	Can be customized to meet domain-specific information and skills
Security	Less secure, as the data is shared with the public	More secure, as the data is not shared with the public
Management	As a service	Private management and ownership

An organization will require a private model if:

- **Data privacy is imperative:** A healthcare provider could use a GenAI model to develop a patient diagnosis system trained on patient data. This would allow the provider to deliver patients more accurate and personalized private diagnoses.
- **Up-to-date transactional information is required:** A retailer could use a GenAI model to develop a dynamic pricing system that is updated with the latest sales data. This would allow the retailer to optimize prices and maximize profits.
- **There is an opportunity to leverage proprietary data and intellectual property:** A financial services provider could use a GenAI model to develop a proprietary investment trading strategy trained on its data and research.

Many large organizations are building private, customized GenAI models to drive business value despite the costs and challenges. Private models with custom applications will provide tangible returns on intellectual property, support the increasingly rigorous data protection and compliance requirements, and produce the foundation for a long-term sustainable competitive advantage.

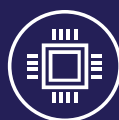
Leverage expertise for speed and savings

Private GenAI models are complex. Design and deployment require deep expertise in AI, machine learning, data science, and the underlying technologies. With the rapid trajectory of AI developments, all GenAI project teams should include individuals with an intrinsic understanding of the latest AI trends and real-world deployment experience. Expert guidance will speed up decision-making and development while reducing the potential for costly errors.

From concept to competitive advantage

Planning to build a sophisticated GenAI model starts with understanding business goals and data capabilities. Although this important process is multi-faceted and can be resource-intensive, the following six steps can lead to project success:

1. **Identify the business problem and desired outcome**
2. **Source and analyze data**
3. **Build the business case**
4. **Plan infrastructure, model design, and deployment**
5. **Build and train the model**
6. **Deploy, monitor performance, and refine as needed**



1. Identify the business problem and desired outcome

Start by clearly identifying the business problem to solve and the desired outcome. A well-defined problem statement will give the project direction, ensure alignment with strategic goals, and encourage leadership and stakeholder commitment.

A manufacturer could use a GenAI model to optimize production schedules, leveraging internal data such as demand forecasts, inventory levels, and machine availability to minimize costs and maximize output. The desired outcome might be to reduce production costs by 10% or increase output by 5%.



2. Source and analyze data

Data quality and relevance are crucial for building private models. Companies should utilize internal and proprietary data to gain an edge.

For example, financial firms can use internal data (historical trading data, customer behavior, risk factors) to build models to predict customer churn, boost revenue, detect potential fraud, and reduce liability and exposure.

Once data has been sourced and analyzed, a data strategy for AI can be developed. This data strategy should define the goals of the AI project, the data that will be used, and how the data will be managed and governed.



3. Build the business case

The business case is essential once the business outcomes and data competencies have been identified. A business case should plan the design, development, and deployment process and will enable productive internal conversations around the investment required, risks involved, and targeted project returns. A robust business case will:

- Set clear objectives and expectations
- Identify risks and challenges
- Encourage project funding/investment
- Outline the deliverables and project timeline
- Identify and align stakeholders, partners, and experts
- Plan risk and change management
- Enable the tracking of progress and measuring of success



4. Plan software, infrastructure, design, and deployment

GenAI models are computationally intensive, requiring significant computing resources and energy to train and deploy. For example, BloombergGPT took 1.3 million hours of GPU time to train.⁸

Software

The software required to develop, deploy, and manage a GenAI model includes:

- **Foundation models:** LLMs, image models, video models, etc., that can be fine-tuned for specific tasks, such as text generation, translation, and code completion (e.g., GPT-3, PaLM, and LLaMA).
- **Data curation and training tools:** Curation tools will support the data's cleansing, preparation, and organization. Training tools will enable model development through machine and deep learning.
- **Inference tooling:** With features such as model optimization, batch processing, and latency reduction.
- **Guardrails:** Topical, safety, and security guardrails ensure AI models stay on track by preventing them from straying into undesired areas, ensuring accurate and appropriate responses, and restricting connections to trusted third-party applications.

In addition to these essentials, many other software applications can help a GenAI build, such as model monitoring, optimization, and debugging tools, and guardrails to help mitigate the risks of bias and misuse.

Infrastructure

The infrastructure for private deployments can be on-premises, co-location, cloud, or edge. The best choice for a particular project will depend on factors such as the size and complexity of the model, the desired performance and scalability, and the budget.

- **On-premises deployment:** The model is deployed on the company's servers in internally managed server storage. This gives the company more control over the model and its data. However, on-premises deployment can be more expensive and complex than cloud deployment.

- **Cloud deployment:** The model is deployed on a cloud platform as a service. Cloud deployment is typically more affordable, easier to manage, and more cost-effective to scale than on-premises deployment; however, this option may not be suitable for all organizations.

- **Edge computing:** Ruggedized, compact edge servers can be used to deploy GenAI models closer to the data source, reducing latency, speeding up inference to application, and improving performance.

Accelerated computing

GPUs are pivotal in advancing GenAI by providing the computational power, parallel processing capability, and hardware acceleration required to efficiently train and deploy complex models.

Consider the following factors when choosing computing hardware:

- **Performance:** GenAI and LLMs are extremely computationally intensive. The hardware should provide the necessary performance to train, deploy, and run the model promptly.
- **Cost:** Accelerated computing hardware will require a sizeable investment, so select cost-effective hardware for the project's specific needs. Factor the TCO and the upfront costs.
- **Scalability:** The hardware should be scalable to support the growing needs of the model as it is trained and deployed.





Networking and connectivity

The networking and connectivity requirements for private GenAI deployments will vary depending on the specific infrastructure chosen. Network performance is required for training data import, model management, and security. Many challenges can be overcome with the latest switches, ethernet, and InfiniBand hardware, and optimizations such as neural network compression.

Security

GenAI models are vulnerable to a variety of security threats, including:

- **Adversarial attacks:** Adversarial attacks are attempts to manipulate or fool a model by providing it with carefully crafted inputs to cause the model to generate incorrect or harmful outputs.
- **Data poisoning:** Data poisoning is the act of injecting malicious data into the training data for a GenAI model. This can cause the model to learn incorrect or biased patterns.
- **Model theft:** Model theft is the unauthorized copying or distributing of a GenAI model. This can allow attackers to use the model for malicious purposes.

In addition to building a private model, implement appropriate security measures to protect models from these threats. This may include:

- **Data security:** Ensuring the training data is secure and protected from unauthorized access using the latest network hardware, software, and security controls.
- **Model monitoring:** Monitoring the model's performance to detect any anomalies that may indicate an attack.
- **Model hardening:** Using adversarial training and input validation to make the model more resistant to attacks.



5. Build and train the model

Once the infrastructure is in place and the model design is complete, the model can be built and trained. This process is time-consuming and compute-intensive, but can be expedited by leveraging optimized networking and GPU power. Some common training algorithms include:

- **Supervised learning:** The model is trained on a labeled dataset, where each data point has a known output. The model learns to predict new data points' outcomes by identifying patterns in the labeled dataset.
- **Unsupervised learning:** In unsupervised learning, the model is trained on an unlabeled dataset, where the data points do not have known outputs. The model learns patterns in the data without prior knowledge of the desired outputs.
- **Reinforcement learning:** In reinforcement learning, the model learns to perform a task by trial and error. The model is rewarded for desired outcomes and penalized for undesired outputs.

Once the model is trained, its performance can be tested and evaluated.



6. Deploy, monitor performance, and refine as needed

Production deployment involves integrating the model into the company's infrastructure, existing systems, and processes.

Following deployment, the model should be monitored for performance and refinement opportunities as needed. Optimization techniques include retraining on new data or with new hyperparameters.

For further information, read Lenovo's technical reference architecture [here](#).

GenAI is transforming every industry

How private GenAI and custom applications are impacting finance, retail, manufacturing, and healthcare

GenAI is bringing significant benefits to the world of business. Custom GenAI applications on private models are leading the charge, with impacts including more accurate decision-making in finance, increased sales and customer retention in retail, quality assurance and cost efficiency in manufacturing, and accelerated drug development and enhanced patient care in healthcare. Across all industries, GenAI is driving tangible improvements, underlining its critical role in the modern workplace.

The impacts of GenAI deployment:



Finance and financial services

- **Improved performance:** Intelligent insights enhance decision-making, leading to optimized performance.
- **Improved customer satisfaction:** Automated and personalized responses lead to quicker and more accurate issue resolution.
- **Enhanced operational efficiency:** With AI handling routine tasks, human agents can focus on more complex customer queries.

[Read more >>](#)



Retail

- **Increased sales:** Tailored product recommendations often result in higher conversion rates.
- **Enhanced customer engagement:** Virtual product advisors and omnichannel strategies provide a seamless and engaging customer experience.
- **Customer retention:** Personalized experiences make customers more likely to return.

[Read more >>](#)



Manufacturing

- **Quality assurance:** Automated visual inspections reduce the rate of defects and improve product quality.
- **Enhanced safety:** Real-time monitoring can quickly identify safety hazards, reducing workplace accidents.
- **Cost reduction:** Automation minimizes the need for manual inspections, leading to significant cost savings.

[Read more >>](#)



Healthcare

- **Speed of research:** GenAI accelerates the pace at which new drugs can be developed and brought to market.
- **Quality of care:** Advanced predictive models improve patient treatment plans, enhancing healthcare outcomes.
- **Data security:** Customized private GenAI builds ensure that sensitive patient data remains secure and compliant with regulations.

[Read more >>](#)

Lenovo





Finance and financial services



The finance industry has always been at the cutting edge of technological advancements. Financial institutions continuously seek ways to outperform their peers in a market characterized by high competition. Embracing innovative technologies offers them a competitive edge and aligns them with their customers' evolving needs and expectations.

Industry trends

- **\$200B to \$340B** can be saved annually by GenAI for banks and the financial services sector.⁷
- **Up to 20%** reduction in loss and default rates can be achieved with credit risk optimization.⁹

Retail banking is significantly changing as customers demand personalized and convenient services. According to Deloitte,¹⁰ customers expect more from their banks — more technology, guidance, support, and cross-channel experiences. McKinsey's¹¹ research supports this by stating that banks that offer a better digital experience lead in customer satisfaction and excel in key financial metrics.

Featured GenAI application story: Intelligent automation

The finance industry utilizes intelligent automation within this competitive and customer-focused landscape. This technology is deployed to enhance the customer experience and optimize call center operations. Customized private GenAI models have become the go-to solution in an industry where data privacy is non-negotiable and proprietary data is abundant.

These private AI models are trained on internal datasets, ensuring that sensitive financial and personal data never leaves the organizational boundary. They can automate routine customer inquiries, route calls to appropriate departments, and provide real-time data analytics to monitor service quality, all while maintaining stringent data security protocols.

Selected additional industry applications

Industry application	Description	Impact
Enterprise Document Search AI	Optimizes information retrieval by evaluating multiple data sources and summarizing results. Also generates reports to streamline office tasks.	Increases organizational efficiency by making information readily accessible.
AI Banking Assistant	Personalizes customer experience, incorporating fraud prevention, compliance, and credit risk management functionalities.	Elevates customer satisfaction and trust while managing risks effectively.
Investment Insights	Uses Natural Language Processing (NLP) to analyze trading research and summarize real-time data streams.	Speeds up decision-making processes in trading and investment, potentially leading to higher returns.



Retail



The retail industry has undergone dramatic transformations over the past two decades. Innovations in technology, shifts in consumer behavior, and the rise of digital channels have reshaped the retail landscape, making it more complex and filled with new opportunities. The retail industry is turning to AI-powered technology in response to these shifts.

Industry trends

- **Up to 40%** of worldwide retailers and brands are in the experimentation phase of GenAI.¹²
- **Up to 59%** improved profitability by 2035, using AI retail solutions.¹³

The meteoric rise of e-commerce has been one of the most significant changes in retail. According to Statista,¹⁴ online sales accounted for 23% of all retail sales in 2023, up from 20% in 2022. Retail is a dynamic and competitive marketplace, with retailers using technology to improve margins and customer loyalty.

Featured GenAI application story: Hyper-personalized marketing

GenAI is transforming the retail and e-commerce industries by enabling hyper-personalized marketing strategies. Customized private GenAI models analyze customer data to generate highly personalized shopping experiences, including product recommendations, virtual product advisor services, and omnichannel intelligent automation.

Personalized product recommendations are created by understanding customer behavior, preferences, and buying patterns, increasing the likelihood of purchase. Virtual product advisors engage customers in real time, guiding them through product selection and suggesting complementary items. Omnichannel intelligent automation ensures a consistent, personalized experience across all platforms while automating customer interactions such as customer service inquiries and post-purchase follow-ups.

Selected additional industry applications

Industry application	Description	Impact
Employee Concierge	AI-powered system designed to assist employees in various tasks, such as scheduling and information retrieval.	Improves employee productivity and job satisfaction.
Personalized Customer Service	Uses AI to offer tailored customer service based on individual behavior and preferences.	Increases customer satisfaction and fosters loyalty.
Supply chain management	Optimizes transportation routes, predicts delivery times, and identifies potential disruptions; communicates with stakeholders and customers.	Improves supply chain efficiency, reduces costs, and mitigates risks.



Manufacturing and advanced industries



The manufacturing industry has faced various challenges recently, from globalization and increased competition to complex supply chains and stringent regulatory compliance. The sector also battles escalating operating costs and mounting cybersecurity threats. Against this backdrop, adopting innovative and sustainable practices has become crucial for survival and growth.

Industry trends

- **Up to 27%** of manufacturers are already investing in GenAI technologies.¹⁵
- **Up to 45%** reduction in downtime with predictive maintenance.¹⁶

Many manufacturers are looking to technology to alleviate some of these pressures. Over half of manufacturing companies plan to increase their use of Internet of Things (IoT) applications, automation, inventory management, and predictive maintenance in the coming years.¹⁷ Additionally, the industry faces a worldwide skills shortage, making labor retention a critical issue. Three-quarters of manufacturers cite this as a challenge, with one in three executives stating that retaining high-performing employees is a strategic priority for 2023 and beyond.¹⁸

Featured GenAI application story: Synthetic data for defect detection

The manufacturing industry uses GenAI and synthetic data to train advanced computer vision models for various applications, including product defect detection.

In the past, obtaining real-world data has been expensive or logistically challenging. However, with synthetic data, GenAI models are trained on available real-world data to understand the statistical characteristics before generating data that closely resembles its attributes. This provides a valuable resource for computer vision model training, with significant cost advantages, improved model accuracy, and ultimately results in better quality and worker safety.

Manufacturers are using GenAI to create synthetic images of products with different types of defects to train models to identify and flag issues in real-world situations.

Selected additional industry applications

Industry application	Description	Impact
AI-Enabled Design Collaboration	AI facilitates real-time collaboration and review processes among engineering teams, improving product design.	Accelerates time-to-market and optimizes design.
LLM for Product Development	Lifelong Learning Machines are trained to assist in internal product development, continuously adapting to new data.	Enhances innovation and reduces product development cycles.
AI-Driven Insider Threat Detection	AI algorithms monitor internal network behavior to detect anomalies that could indicate insider threats and act accordingly.	Bolsters cybersecurity and protects intellectual property.



Healthcare and pharma



The healthcare industry is grappling with many challenges, including increased competition, rising operational costs, global supply chain disruption, and many regulatory hurdles. Faced with these complexities, healthcare relies on technology to optimize operations, enable efficiencies, and drive improved outcomes.

Industry trends

- By 2025, Gartner expects more than **30%** of new drugs and materials to be systematically discovered using GenAI.¹⁹
- Up to **30x** faster and **99%** accurate mammograms using GenAI translation techniques.²⁰

Data plays a critical role in healthcare, with the industry accounting for 30% of the world's data and growing at 36% every year, according to the OECD.²¹ Investment in AI software in healthcare is predicted to reach \$11.6 billion by 2026, highlighting the industry's shift toward technology-based solutions.²²

Featured GenAI application story: Molecular simulation

The BioPharma sector utilizes GenAI for various applications to accelerate research and development. These include Molecular Simulation, Structural Biology, GenAI Bio Training/Model Development, GenAI Bio Inference/Design in Biomedical Imaging, and Real-World Evidence through Predictive Modeling.

Customized private GenAI analyzes large datasets to identify patterns, make predictions, and suggest potential drug development or patient treatment pathways. These models are particularly adept at managing the high volumes of complex and sensitive data typical in healthcare research.

Selected additional industry applications

Industry application	Description	Impact
AI-Powered Personalized Healthcare	AI-driven algorithms develop personalized treatment plans based on individual health data.	Enhances patient outcomes and improves treatment efficiency.
AI-Enhanced Medical Imaging	AI algorithms improve the quality and detail of medical images, aiding in diagnosis and treatment.	Accelerates diagnostic processes and enhances patient care.
AI-Assisted Genomic Analysis	AI analyzes genomic data to identify patterns and anomalies, aiding in research and personalized medicine.	Advances in genomics could lead to breakthroughs in various medical conditions.

Lenovo brings AI to data

Realize the potential of GenAI with Lenovo and NVIDIA

Lenovo and NVIDIA have joined forces to provide a wide range of solutions and services designed to drive the global adoption of GenAI and help customers realize its full potential. Enjoy a competitive advantage and an accelerating future with fast, secure, scalable, end-to-end solutions backed by industry-leading know-how and professional services.

In collaboration, Lenovo and NVIDIA offer the most comprehensive AI portfolio, proven expertise, and advisory support, unlocking the power of GenAI for every industry and helping customers work toward a smarter, faster future.

Increase revenue opportunities, optimize productivity and costs, and mitigate risks with a private GenAI model powered by Lenovo and NVIDIA's AI-ready portfolio of solutions and services:

- **Leverage innovative AI solutions:** Take advantage of Lenovo's \$100m AI commitment and achieve next-level results. With over 150 AI-ready solutions from almost 50 AI partners, Lenovo's AI Innovators program delivers the fastest time-to-solution for every industry.
- **Optimize AI infrastructure:** Deploy an industry-leading AI portfolio and deliver AI everywhere the business needs with the lowest TCO — from edge to cloud. Drive breakthrough performance with a private GenAI solution backed by accelerated servers, fast networking, reliable storage, and NVIDIA's AI software platform.
- **Enable AI discovery:** Work with Lenovo and NVIDIA's AI experts to get the most value while lowering project risks. Lenovo has been pushing boundaries at the forefront of AI for almost a decade. Benefit from the Lenovo AI Discover Lab, AI assessment workshops, and an AI committee driving AI adoption for customers on every continent.



Adopt AI faster with Lenovo NVIDIA Solutions

Innovating AI Solutions

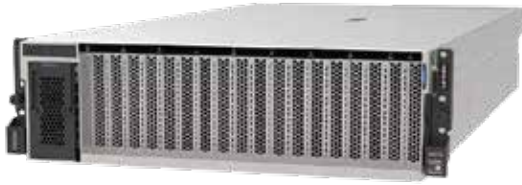
In partnership with AI innovators like DeepBrain — AI videos for sales and customer service, Chooch — AI computer vision for quality control and safety, and Edgebricks — AI infrastructure for faster AI deployment, Lenovo offers an ecosystem of customizable enterprise AI solutions through the AI Innovators program. The program provides AI for end-to-end operations, including audio recognition, prediction, security, and virtual assistants for every industry, including finance, retail, manufacturing, and healthcare.

The industry's most comprehensive AI-optimized portfolio

Lenovo has launched a portfolio of purpose-built AI infrastructure solutions that power high-performance edge-to-cloud and fulfill the growing market demands for GenAI. The Lenovo and NVIDIA accelerated computing stack enables every industry to tap into the power of AI, delivering the performance, scale, and efficiency levels needed for running the next wave of applications. It's a full-stack platform enabling innovation and creativity for solving the world's toughest challenges.



AI-optimized infrastructure



Data Center Servers

Lenovo ThinkSystem SR675 V3

The Lenovo ThinkSystem SR675 V3 is a versatile GPU-rich 3U rack server that supports eight double-wide GPUs, including the new NVIDIA H100 and L40S Tensor Core GPUs or the NVIDIA HGX H100 4-GPU offering with NVLink and Lenovo Neptune hybrid liquid-to-air cooling.

The server delivers optimal performance for AI, High Performance Computing (HPC), and graphical workloads, allowing users to extract greater insights and drive innovation utilizing machine learning and deep learning.



Data Center Servers

Lenovo ThinkSystem SR670 V2

Powered by the new NVIDIA H100 and L40S Tensor Core GPUs or the NVIDIA HGX H100 4-GPU, the Lenovo ThinkSystem SR670 V2 is optimized for AI, HPC, and graphic workloads for a wide variety of applications, including GenAI.

The SR670 V2 delivers an enterprise-grade solution, accelerating workloads in production and maximizing system performance. Retail, manufacturing, financial services, and healthcare industries leverage the SR670 V2 to enhance processes and drive innovation.



Edge Server

Lenovo ThinkEdge SE455 V3

The Lenovo ThinkEdge SE455 V3 server brings a new modular approach to edge computing and AI-ready processing power, storage, and network closer to where data is generated.

As Lenovo's flagship edge-optimized server, the SE455 V3 with NVIDIA L40 or L4 GPUs is ideal for large and demanding edge AI workloads, featuring best-in-class performance and built-in sustainability.

Supporting components

NVIDIA ConnectX-7

The NVIDIA ConnectX-7 SmartNIC is optimized to deliver accelerated networking for modern cloud, artificial intelligence, and traditional enterprise workloads. ConnectX-7 provides a broad set of software-defined, hardware-accelerated networking, storage, and security capabilities that enable organizations to modernize and secure their IT infrastructures.

Infrastructure as a Service (IaaS)

Lenovo TruScale for AI helps businesses achieve premium performance with a limited upfront capital expenditure. The IaaS model offers immediate access to AI deployment, and a one-stop connection to Lenovo's 150+ turnkey AI solutions, to accelerate intelligent transformation. TruScale delivers scalable, end-to-end services from deployment to management and refresh, providing customers with a predictable monthly payment model.

NVIDIA AI Inference Platform

The NVIDIA AI inference platform offers a complete end-to-end stack and suite of products, infrastructure, and services, including NVIDIA AI Enterprise, to deliver the performance, efficiency, and responsiveness that are critical to powering the next generation of AI inference — in the cloud, in the data center, at the network edge, or in embedded devices.

NVIDIA AI Enterprise

The enterprise-grade software that powers the NVIDIA AI platform, NVIDIA AI Enterprise accelerates data science. It streamlines developing and deploying production-ready generative AI, computer vision, speech AI, and more. Enterprises that run their businesses on AI rely on NVIDIA AI Enterprise to improve the productivity of AI teams and achieve business insights faster.

NVIDIA NeMo

Part of NVIDIA AI Enterprise, NVIDIA NeMo enables organizations to build custom large language models (LLMs) from scratch, customize pre-trained models, and deploy at scale. NeMo includes training and inferencing frameworks, guardrail toolkits, data curation tools, and pre-trained AI models.

Enabling AI Discovery

Adopt AI faster with AI experts, workshops, and best practices. The Lenovo AI Discover Lab provides access to Lenovo data scientists, AI architects, and engineers to help explore, deploy, and scale AI solutions. The service guides customers to the most appropriate software partners and AI-optimized infrastructure, sharing expertise developed from Lenovo and NVIDIA's combined AI investments and innovations.

Lenovo offers assessment workshops to drive AI adoption and responsible AI guidance to help organizations understand and address privacy, fair usage, diversity, equity, inclusion, and accessibility considerations through the Lenovo Responsible AI Committee.

In partnership, Lenovo and NVIDIA are helping customers harness the value of their data to deploy purpose-built AI solutions with speed, transforming organizations with more predictable outcomes.

Talk to the Lenovo and NVIDIA experts to start your AI journey faster.



Contact Lenovo's AI Discover Lab at AIDiscover@lenovo.com to schedule an appointment.



Lenovo and NVIDIA

In partnership with NVIDIA, Lenovo is developing world-changing technologies and sharing combined expertise through professional services to create a more efficient, connected, and digital society. By designing and engineering the world's most complete portfolio of innovative, AI-ready devices and infrastructure, and driving adoption through advisory and support services, Lenovo and NVIDIA are leading an Intelligent Transformation — to create better experiences and opportunities for millions of customers worldwide.

Accelerating AI relies on accelerated infrastructure and powerful software, and NVIDIA delivers acceleration everywhere needed — to data centers, desktops, laptops, and the world's fastest supercomputers. As companies are increasingly data-driven, the demand for AI technology grows. AI provides enterprise teams the power, tools, and algorithms to work effectively, from customer service chatbots to hyper-personalized communication and automated production optimization.

Lenovo and NVIDIA bring innovative solutions and intelligent infrastructures to solve the most significant challenges of today and tomorrow. Together, we equip data-centered researchers, pioneers, and visionaries across all industries with the tools to help them evolve, transform, and implement enterprise AI solutions to deliver Smarter Technology for All.

[Find out more](#)

Getting started



There has never been a better time to invest in AI

Get in touch today



Harness the expertise and power of Lenovo and NVIDIA

Schedule an AI business strategy workshop



Contact the team at Lenovo to kick-start your AI journey

Speak to the AI experts

Lenovo

 **NVIDIA**

References

- ¹ [Statista, 2021, Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025](#)
- ² [The University of Chicago Department of Computer Science, 2020, Globus Reaches One Exabyte Milestone in Research Data Management](#)
- ³ [Scality, What is an Exabyte anyway?](#)
- ⁴ [MarketsandMarkets™ Research Private Ltd., Artificial Intelligence \(AI\) Market by Offering \(Hardware, Software\), Technology \(ML \(Deep Learning \(LLM, Transformers \(GPT 1, 2, 3, 4\)\), NLP, Computer Vision\), Business Function, Vertical, and Region — Global Forecast to 2030](#)
- ⁵ [McKinsey & Company, 2023, What's the future of generative AI? An early view in 15 charts](#)
- ⁶ [Bloomberg, 2023, Generative AI to Become a \\$1.3 Trillion Market by 2032, Research Finds](#)
- ⁷ [McKinsey Digital, 2023, The economic potential of generative AI: The next productivity frontier](#)
- ⁸ [Lenovo, 2023, Reference Architecture for Generative AI Based on Large Language Models \(LLMs\)](#)
- ⁹ [Avenge, 2022, Artificial Intelligence \(AI\) for Credit Risk Management in Banking](#)
- ¹⁰ [Deloitte, 2023, 2024 banking and capital markets outlook](#)
- ¹¹ [McKinsey & Company, 2023, Five ways to drive experience-led growth in banking](#)
- ¹² [IDC, 2023, How Retailers and Brands are Taking Advantage of Generative AI](#)
- ¹³ [Accenture, 2023, 6 Pivotal Benefits of AI for Retail \(+ Use Cases from Top Brands\)](#)
- ¹⁴ [Statista, 2023, E-commerce as percentage of total retail sales worldwide from 2015 to 2027](#)
- ¹⁵ [IDC, 2023, How Generative AI is Impacting Industries](#)
- ¹⁶ [McKinsey & Company, 2021, The Internet of Things: Catching up on an accelerating opportunity](#)
- ¹⁷ [Lumen Technologies, 2021, Edge Computing: Services for Manufacturing](#)
- ¹⁸ [Deloitte, 2023, 2023 manufacturing industry outlook](#)
- ¹⁹ [Gartner, 2023, Beyond ChatGPT: The Future of Generative AI for Enterprises](#)
- ²⁰ [CNA, 2017, AI, Robots, AND SWARMS: ISSUES, QUESTIONS, AND RECOMMENDED STUDIES](#)
- ²¹ [OECD, 2021, The importance of increasing access to high-quality health data](#)
- ²² [Lenovo, 2023, Moving AI from Idea to Execution](#)

© 2023 Lenovo. © 2023 NVIDIA Corporation. All rights reserved.

Trademarks: Lenovo, the Lenovo logo, ThinkSystem and ThinkEdge are trademarks or registered trademarks of Lenovo. NVIDIA, the NVIDIA logo are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries.

