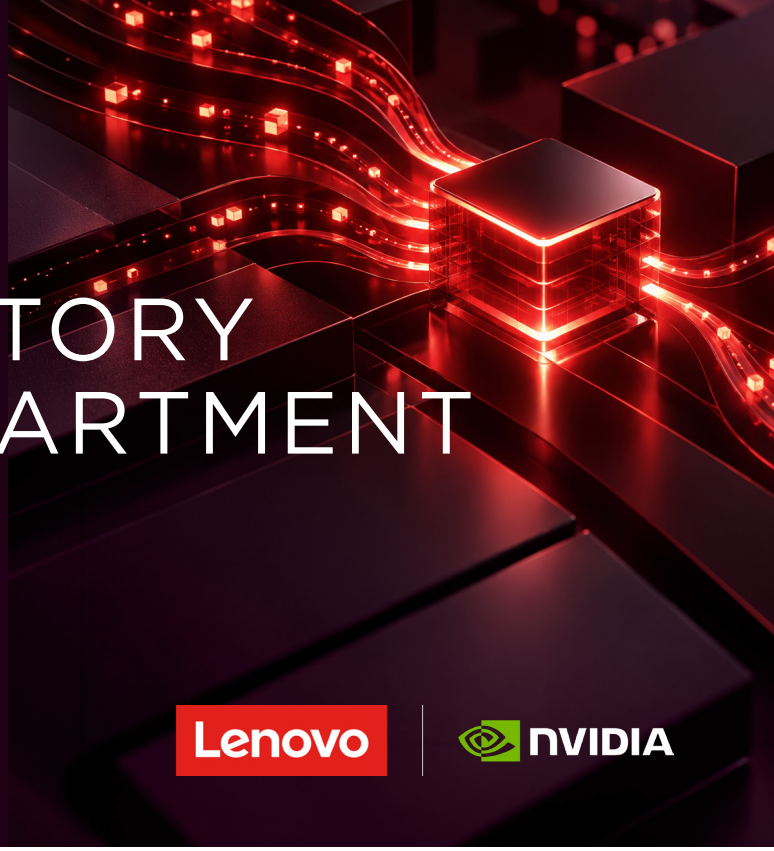
Lenovo Hybrid AI Factory with NVIDIA and Canonical

LENOVO HYBRID AI FACTORY FOR YOUR DEPARTMENT



TRUSTED, PRIVATE AI. READY IN HOURS.

Turn your internal knowledge into a governed AI assistant on a single Lenovo server. Built with NVIDIA® and Canonical, the Lenovo Hybrid AI Factory gives you a Lenovo validated starting point and keeps inference local, responsive, and scalable as users, models, and workloads grow.

TRUSTED ANSWERS

Get answers grounded in your approved enterprise knowledge. In Lenovo testing, responses scored 0.86 for groundedness.[1]

CONTROLLED DATA

Keep sensitive data and AI workloads in the environment you control.

EFFICIENT INFERENCE

Run inference locally on right-sized infrastructure for fast responses and better cost per token.

WHY TRUSTED ENTERPRISE ASSISTANTS?

KEY CHALLENGES

Enterprise knowledge is growing across policies, reports, documents, databases, images, and PDFs. Internal teams, customer support, legal and compliance, product engineering, and regulated industries all depend on it, yet access is often fragmented, expensive, and difficult to govern.

Enterprises want to begin with a right-sized infrastructure footprint and avoid a premature build-out. They need a small deployment that still looks and behaves like the platform they plan to scale in the future.

84%

of organizations leverage on-premises or edge deployments for AI workloads and applications.[2]

SOLUTION OVERVIEW

Lenovo Hybrid AI Factory with NVIDIA combines accelerated computing, networking, storage, orchestration, AI software, open models, applications, and services in Lenovo validated architectures.

This Lenovo Hybrid AI Factory with NVIDIA offering provides a focused single-node starting point for enterprise AI assistants based on the NVIDIA Blueprint for Retrieval-Augmented Generation (RAG), leveraging Canonical solutions and NVIDIA NeMo Retriever™. Running inference locally delivers fast, predictable responses while reducing cost per token as usage grows. The same validated architecture can scale seamlessly to support additional users, models, and AI workloads.

Typical use cases include: customer support, AI assistants, enterprise search, sales assistants, and smaller AI agent workflows. Marketing generates campaigns. Sales gets proposal copilots. Finance summarizes reports.

Start with enterprise RAG, then use the same Lenovo platform for agentic AI, enterprise model serving, computer vision, and real-time analysis as your needs grow.

DEPLOY TRUSTED ENTERPRISE ASSISTANTS.

Lenovo gives you a practical path from pilot to production. Start with a single Lenovo server powered by NVIDIA accelerated computing and NVIDIA AI Enterprise, with Canonical, then scale on the same Lenovo validated foundation.

Lenovo Hybrid AI Factory for your department with NVIDIA and Canonical

ENTERPRISE KNOWLEDGE

- Documents & policies
- Research & reports
- Images & tables
- Databases & systems

INGEST & PROCESS

- Doc Parsing (OCR, tables, layout)
- Text chunking & enrichment
- Embeddings generation
- Metadata extraction

RETRIEVE & AUGMENT

OpenSearch

Vector & keyword search

Hybrid retrieval & reranking

GENERATE

NVIDIA AI Enterprise

NVIDIA NeMo Retriever™

- LLM inferencing
- Reranking
- Embeddings
- Advanced parsing

ENTERPRISE AI ASSISTANT

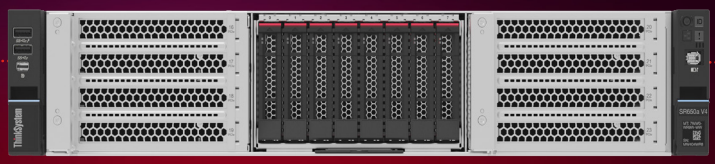
- Accurate, grounded answers
- Citations & source references
- Secure & private
- Built for the enterprise

ENTERPRISE-GRADE FOUNDATION

Canonical

Lenovo AI Inferencing Servers

Single scalable node • Production ready • Built for AI



Lenovo ThinkSystem SR650a V4

OPTIONAL ENTERPRISE-GRADE FOUNDATION

Red Hat

START AND DEPLOY

Single-node, two GPUs

Lenovo ThinkSystem SR650a V4 • 2x NVIDIA RTX PRO™ 6000 Blackwell Server Edition • NVIDIA and Canonical

SCALE

Multi-node scalable units

Lenovo ThinkSystem SR675 V3 • NVIDIA HGX™ B200 or H200 NVL • More users, higher concurrency, greater throughput, larger workloads on the same Lenovo validated foundation

Lenovo Hybrid AI Factory with NVIDIA = accelerated computing + networking + storage + software + models + applications + services.

WHY LENOVO AND NVIDIA

More than 20 years of partnership, engineered into one Lenovo validated stack.

674
MLPerf™ world records across AI training and inference[3]

12
consecutive years ranked number one for server reliability[4]

#1
most secure and toughest to crack, ITIC 2025 server security survey[5]

#5
Gartner Supply Chain Top 25 for 2026, a Lenovo-owned supply chain[6]

Leader
in the Integrated AI Infrastructure Systems quadrant[7]

Lenovo | NVIDIA

Learn more — Lenovo Hybrid AI Factory with NVIDIA and Canonical

Learn more — Lenovo Hybrid AI Factory with NVIDIA and Red Hat

1 Lenovo Validated Design, Hybrid AI 221 Canonical Single-node-RAG, v1.0
2 IDC CIO Playbook 2026
3 MLPerf benchmark results, Lenovo Press LP1145
4 ITIC Global Server Hardware and Server OS Reliability Study
5 ITIC 2025 Server Security Survey
6 Gartner Supply Chain Top 25 for 2026
7 2026 ISG Provider Lens® global AI-ready Infrastructure Solutions report

© Lenovo 2026 All rights reserved. Lenovo, the Lenovo logo, Smarter Technology for All, ThinkCentre, ThinkPad, ThinkStation, and ThinkSystem 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. Red Hat and the Red Hat logo are trademarks or registered trademarks of Red Hat, LLC or its subsidiaries in the U.S. and other countries. Ubuntu and Canonical are registered trademarks of Canonical Ltd. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries.