ACCELERATE DEVELOPMENT. UNLOCK ENTERPRISE AI. SUPERMICRO AND NVIDIA: RAG-READY INFRASTRUCTURE

Table of Contents

Executive Summary

The rapid evolution of artificial intelligence (AI) is driving enterprises to seek more secure and effective ways to harness large language models (LLMs) for business impact. Yet, organizations face persistent challenges—including infrastructure complexity, staffing shortages, and the need to deliver accurate, contextually relevant, and secure AI outputs.

This solution brief spotlights Retrieval-Augmented Generation (RAG) as the key to unlocking enterprise-ready AI—connecting LLMs directly to your organization's data for precise, business-specific results while maintaining strict data governance and privacy. Enabled by Supermicro's flexible systems and NVIDIA's industry-leading GPUs, the AI Factory solution provides a turnkey, scalable foundation for deploying advanced RAG pipelines. By combining pre-trained models with enterprise knowledge retrieval and robust security controls, businesses can accelerate adoption, ensure trustworthy outcomes, and confidently protect sensitive information as they realize the full potential of AI.

Barriers to Enterprise AI Adoption

Enterprises face a range of practical and technical challenges when deploying AI. Staffing shortages for routine IT and AI tasks, combined with constraints on power, cooling, and space, make it difficult to build and maintain the infrastructure needed for today's demanding workloads. Many organizations are also new to AI hardware and lack the in-house expertise to design, deploy, and manage complex systems, increasing the need for turnkey, validated solutions.

Beyond infrastructure, pre-trained AI models bring additional obstacles. While they offer a fast path to adoption, these models can generate inaccurate or "hallucinated" responses and often fail to provide business-relevant answers. Data security and safety are major concerns, particularly when using open-source models or cloud solutions, as enterprises still need to protect sensitive information and meet governance requirements. Training foundation models on enterprise-specific data is often prohibitively expensive, leading most organizations to rely on fine-tuning—but this alone may not deliver fully accurate, contextual, and secure results.

These technical and security challenges contribute directly to the gap between AI investment and successful deployment. A McKinsey report highlights that 92% of companies plan to increase AI investments over the next three years, yet only 1% of leaders consider their organizations mature in AI deployment. Many enterprises also allocate a small fraction of their digital budgets to AI, with 58% reporting less than 10% of their spend directed toward AI initiatives. These figures underscore the urgent need for secure, scalable, and validated AI solutions that help enterprises translate investment into successful, high-impact deployments.[1]

1 AI in the workplace: A report for 2025 | McKinsey

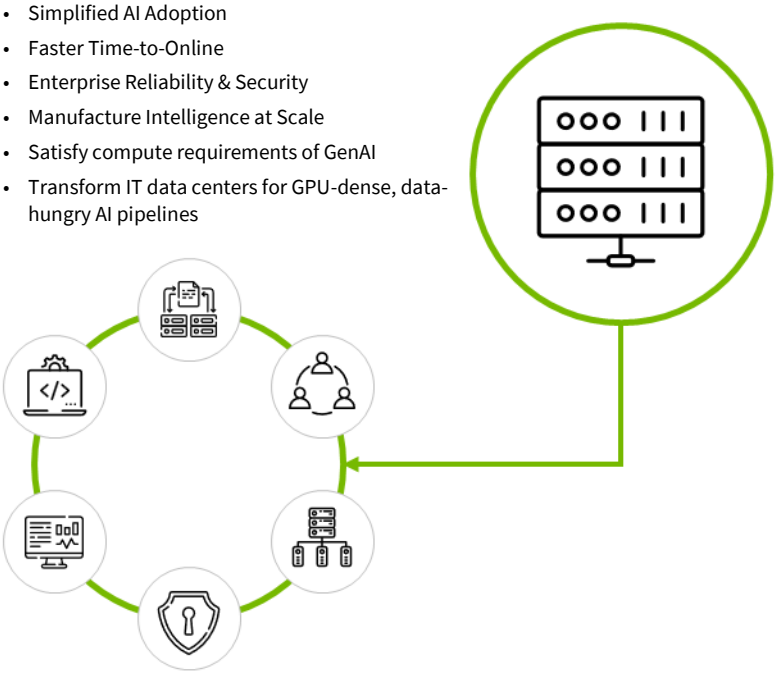
Unlocking Enterprise AI with Purpose-Built Infrastructure

The rapid evolution of artificial intelligence is driving a fundamental shift in enterprise IT, enabling organizations across every industry to unlock new value, streamline operations, and maintain a competitive edge. As AI adoption accelerates, enterprises are seeking ways to harness advanced AI capabilities—not just for innovation, but as a core driver of business transformation.

AI Factories—purpose-built, scalable, and modular infrastructures—are emerging as the foundation for this transformation. Powered by the latest NVIDIA GPUs and delivered through Supermicro's flexible, enterprise-ready systems, AI Factories provide a turnkey approach to deploying, managing, and scaling AI workloads across the enterprise.

Key Reasons AI Factories Matter to Enterprises

- Simplified AI Adoption
- Faster Time-to-Online
- Enterprise Reliability & Security
- Manufacture Intelligence at Scale
- Satisfy compute requirements of GenAI
- Transform IT data centers for GPU-dense, data-hungry AI pipelines



While AI Factories offer a compelling path forward, realizing their full potential requires overcoming a set of practical and technical challenges that many enterprises still face today.

From Pre-Trained Models to Context-Aware Intelligence

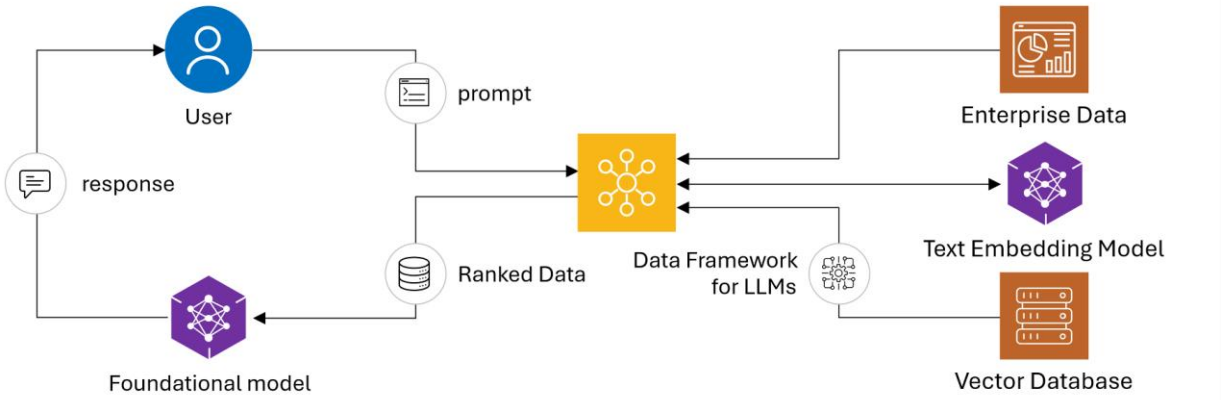
To overcome the limitations of pre-trained AI models, enterprises can post-train models using open-source AI and their own organizational data. Retrieval-Augmented Generation (RAG) provides a practical way to accomplish this, combining external knowledge with AI reasoning to deliver accurate, relevant, and secure results tailored to your business context.

What is RAG?

Retrieval-Augmented Generation (RAG) connects large language models (LLMs) to your organization's own data, allowing AI to generate responses that are accurate, current, and aligned with business context. By retrieving relevant information from internal sources—like knowledge bases or document repositories—RAG grounds model outputs in real enterprise content, reducing hallucinations and keeping sensitive data secure within your environment. This approach helps enterprises get more precise, trustworthy results without the cost or complexity of training large models from scratch.

Typical RAG requirements include:

- **GPUs:** For accelerated model inference and retrieval tasks.
- **Storage:** To host indexed enterprise documents and embeddings.
- **Software stack:** A pre-trained LLM, a retrieval system (such as a vector database), and orchestration tools to connect them.



From Strategy to Solution

Recognizing the constraints of generic AI deployments, enterprises are increasingly adopting post-training strategies that combine open-source models with their own proprietary data. This approach enables organizations to tailor AI outputs to their specific business context while maintaining control over data privacy and relevance.

Retrieval-Augmented Generation (RAG) plays a central role in this strategy, allowing pre-trained models to dynamically access and incorporate enterprise knowledge at runtime. However, unlocking the full potential of RAG requires infrastructure that is powerful, scalable, and ready to deploy—without adding complexity or delay.

The Supermicro RAG solution featuring the NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU

That's where Supermicro and NVIDIA step in. Together, they provide a complete, validated solution designed to support RAG workloads from day one. NVIDIA's RAG Blueprint offers a proven framework for building and scaling RAG pipelines, helping organizations deploy with confidence. To complement this NVIDIA Blueprint, Supermicro delivers a portfolio of NVIDIA-Certified Systems™ engineered for RAG workloads. Optimized for NVIDIA's latest GPUs—including the NVIDIA RTX PRO 6000 Blackwell Server Edition—these systems provide the performance, scalability, and efficiency required for real-time inference, retrieval, and multimodal AI applications.

In the following sections, we'll highlight the Supermicro systems and NVIDIA GPUs optimized for enterprise RAG deployments, talk about leveraging NVIDIA AI Enterprise software to accelerate and simplify AI development and scaling, and access and get started with the NVIDIA AI Blueprint for RAG.

Step 1: Select The Right Supermicro and NVIDIA Infrastructure

The first step on the road to enabling RAG across your enterprise is selecting the right infrastructure to support it. Successful deployment depends on choosing systems that align with your performance, scalability, and workload requirements. Supermicro offers a range of NVIDIA-Certified Systems designed and validated to meet the demands of RAG—from GPU count and network bandwidth to form factor. Now, let's explore the Supermicro system and NVIDIA GPUs that are optimized for your RAG deployment strategy.

Supermicro System Configurations for the NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU

The following Supermicro system configurations are optimized and certified for the NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU, ensuring compatibility and support for NVIDIA AI Enterprise software to simplify the development and deployment of production-grade AI.

PCIe-Optimized Systems Supporting NVIDIA RTX PRO 6000 Blackwell Server Edition

2U CloudDC / SYS-222C-TN **2-GPU**	**4U MGX /** SYS-422GL-NR **4-GPU**	**5U GPU** SuperServer / SYS-522GA-NRT or AS -5126GS-TNRT2 **8-GPU**
2-2-3 	**2-4-3** 	**2-8-5** 
Dual Xeon 6700 CPUs, 6 PCIe 5.0 x16/x8 slots, 2 kW CRPS PSU. Up to 24 NVMe front drives in a DC-MHS chassis.	Eight GPU-ready slots; ship with four RTX PRO 6000s for a 2.4 kW draw. Plenty of NIC slots for 2× 200 GbE or IB.	13 PCIe 5 x16 slots & 24 front U.2/U.3 NVMe bays. Taller 5U plenum + 8.6kW usable power keeps GPUs & DPUs cool together.
2-2-3 architecture: • **2 (Dual)** Intel Xeon 6700 series CPUs • **2 GPU** PCIe per system (Up to 2-GPU) • **3 NIC:** E/W: 2x BF3 B3140H, N/S: BF3 B3220 • **Alternate NIC:** CX7 2x200G	**2-4-3 architecture:** • **2 (Dual)** Intel Xeon 6900 series CPUs • **4 GPU** PCIe per system (Up to 8-GPU) • **3 NIC:** E/W: 2x BF3 B3140H, N/S: BF3 B3220 • **Alternate NIC:** CX7 2x200G	**2-8-5 architecture:** • **2 (Dual)** Intel Xeon 6th / AMD EPYC 4th CPUs • **8 GPU** PCIe per system (Up to 8-GPU) • **5 NIC:** E/W: 4x BF3 B3140H, N/S: BF3 B3220 • **Alternate NIC:** CX7 2x200G

Intel **AMD**

2-GPU RTX PRO Server – SYS-222C-TN

4-GPU RTX PRO Server – SYS-422GL-NR

8-GPU GPU-optimized solution – AS -5126GS-TNRT2 or SYS-522GA-NRT

NVIDIA GPUs Optimized for AI Workloads

The following NVIDIA GPUs are well suited to accelerate RAG workloads at scale. Each GPU—NVIDIA RTX PRO™ 6000 Blackwell Server Edition, NVIDIA H200 NVL, and NVIDIA HGX™ B200 and B300—offers unique combinations of memory, performance, and interconnect capabilities, enabling enterprises to match GPU resources to the size and complexity of their RAG pipelines while maintaining low-latency, high-accuracy inference.

NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU

Built on the groundbreaking NVIDIA Blackwell architecture, the NVIDIA RTX PRO™ 6000 Blackwell Server Edition delivers a powerful combination of AI and visual computing capabilities to accelerate enterprise data center workloads. Suited for enterprises looking to run a range of enterprise workloads—in addition to generative AI—using small-model inferencing and fine-tuning (models less than 70B).

- Passive heatsink taps chassis airflow so **450–600 W** cards pack densely, boosting rack efficiency and lowering PUE.
- Most powerful PCIe Gen5 upgrade for L40/L40S/A40 slots—built for air-cooled racks, & up to **8-GPU scale-out servers**.
- Powers a range of enterprise workloads including enterprise inference, model fine-tuning, HPC, virtual desktops, and container-scale distributed graphics.

The NVIDIA H200 NVL GPU

The NVIDIA H200 Tensor Core GPU supercharges generative AI and high-performance computing (HPC) workloads with game-changing performance and memory capabilities. Ideal for customers training foundational models or using large models for inferencing (models greater than 70B).

- 141 GB HBM3e at 4.8 TB/s per GPU and 900 GB/s NVLink pair form a 282 GB logical device for huge models.
- Full-scale LLM training, high-batch GenAI inference, large-graph recommenders, and memory-bound HPC (CFD/weather).
- Maximizes per-node footprint to avoid sharding—cutting latency and energy per token while boosting throughput.

NVIDIA HGX Platform (HGX B300 / HGX B200)

As a premier accelerated scale-up platform with up to 30x more AI Factory output than the previous generation, NVIDIA Blackwell Ultra-based HGX systems are designed for the most demanding generative AI, data analytics, and HPC workloads. Designed for enterprises with intensive AI training and inference workloads, this is the most powerful NVIDIA GPU available for training and inference.

- Use when you need maximum per-node performance and 1.8 TB/s NVLink across 8 GPUs for large-model training and high-throughput inference.
- Full LLM training, AI reasoning, large-batch GenAI, and FP64/HPC that benefit from fast multi-GPU collectives.
- Facility ready for higher TDP (≈700–1,200 W/GPU, often liquid-cooled) to sustain clocks and performance density.

Step 2: Accelerate your AI development with NVIDIA AI Enterprise

NVIDIA AI Enterprise is a cloud-native suite of software tools, libraries, and frameworks, including NVIDIA NIMs and NeMo microservices, that accelerate and simplify the development, deployment, and scaling of AI applications. Organizations of all sizes can deploy agentic AI systems anywhere—across clouds, data centers, or at the edge—leveraging the extensive partner ecosystem. NVIDIA AI Enterprise helps accelerate time to market and reduce infrastructure costs while ensuring reliable, secure, and scalable AI operations.

Environment Overview

NVIDIA AI Enterprise provides a performance-optimized, modular environment for building and running AI applications. Its libraries, tools, and containers accelerate the full AI pipeline, from model orchestration to real-time inference, enabling faster deployment of enterprise workloads. Cloud-native design integrates with industry-leading orchestration platforms, allowing AI applications to run seamlessly in the cloud, on-premises, or at the edge. Extended-life software branches, proactive security updates, and tested deployment guidance make it enterprise-ready, while the rich partner ecosystem provides support for hardware, software, and system integration.

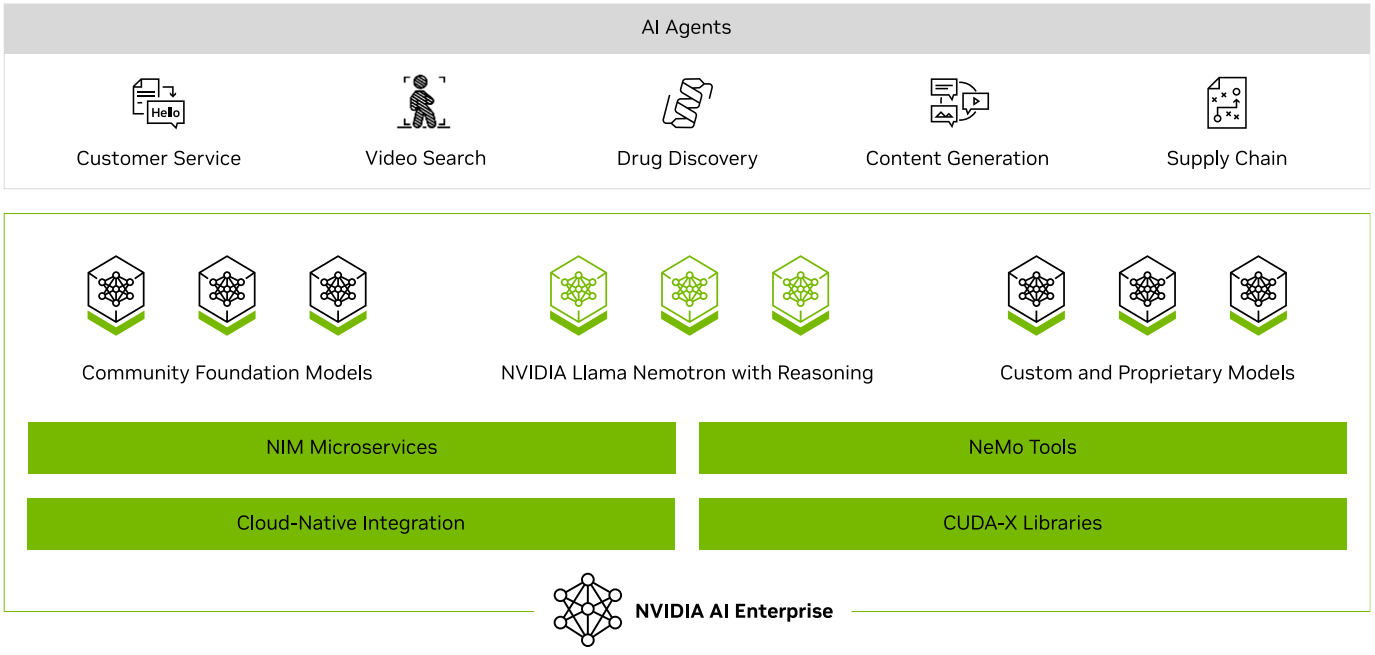
Learn more: NVIDIA AI Enterprise | Cloud-native Software Platform | NVIDIA, or get started with NVIDIA AI Enterprise: Get Started With NVIDIA AI Enterprise | NVIDIA.

Key Building Blocks:

1. NVIDIA NIM microservices – Provides prebuilt, optimized inference microservices for rapidly deploying the latest AI models on any NVIDIA-accelerated infrastructure—cloud, data center, workstation, and edge.
2. NVIDIA NeMo – an end-to-end platform for developing custom generative AI—including large language models (LLMs), vision language models (VLMs), video models, and speech AI—anywhere.
3. NVIDIA Blueprints – A collection of validated AI workflow templates to accelerate development and deployment.

With the NVIDIA AI Enterprise environment and its building blocks in place, the next step is to leverage NVIDIA Blueprints to start building and scaling your enterprise RAG pipeline.

Production-Ready Software for Agentic AI



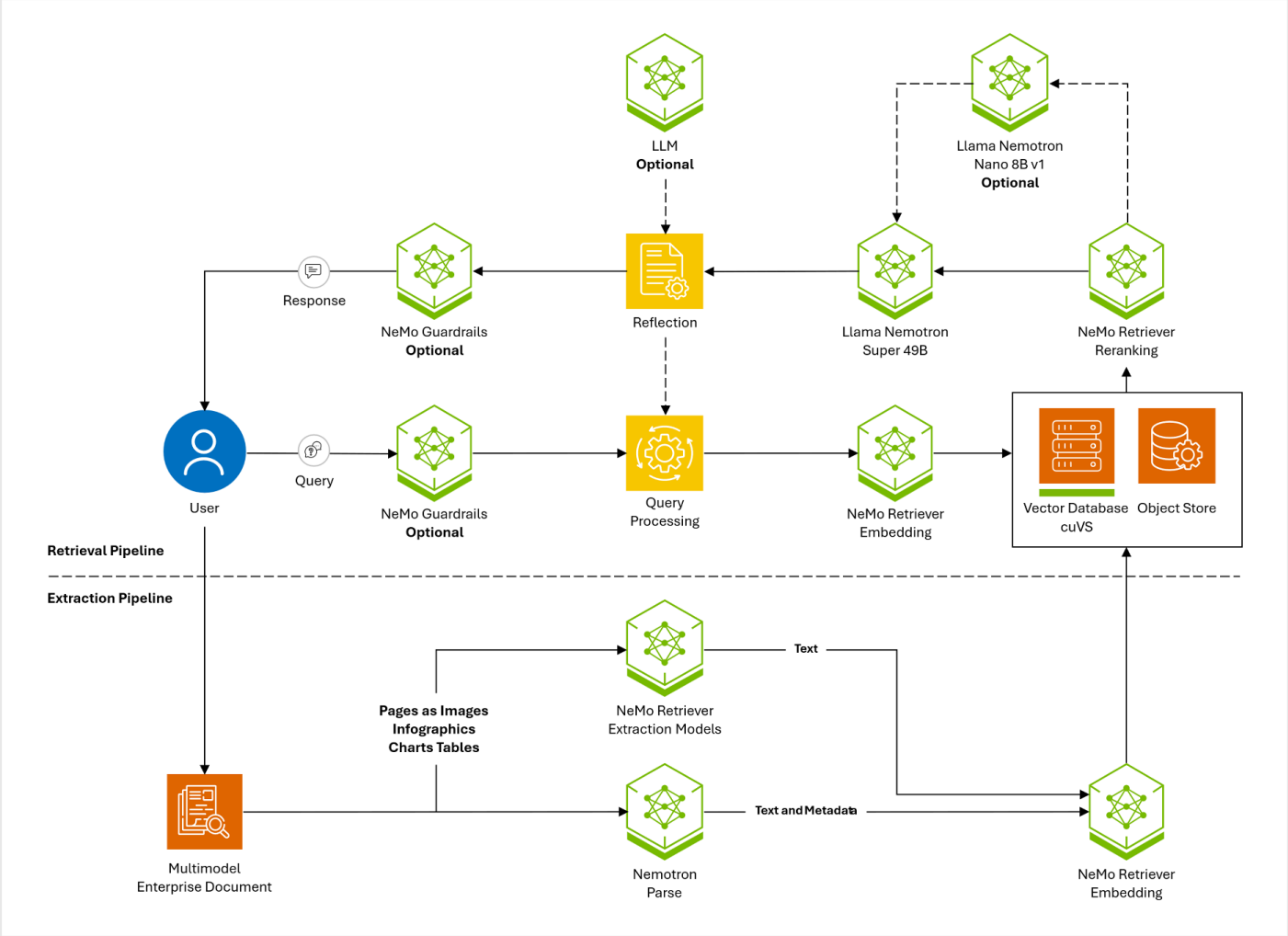
Step 3: Start Building an Enterprise RAG Pipeline With NVIDIA Blueprints

Begin by visiting **Build.nvidia.com** and selecting the **NVIDIA AI Blueprint for RAG**. This NVIDIA Blueprint provides developers with a foundational starting point for building scalable, customizable data extraction and retrieval pipelines using NVIDIA NeMo Retriever models. Use the NVIDIA Blueprint to connect LLMs to extensive multimodal enterprise data—including text, tables, charts, and millions of PDFs—to deliver context-aware responses. Enterprises can unlock actionable insights with 15x faster multimodal PDF data extraction and 50% fewer incorrect answers, driving productivity at scale.

The NVIDIA AI Blueprint for RAG provides a production-ready workflow for building enterprise-scale AI solutions that combine pre-trained LLMs with targeted data retrieval. Powered by NVIDIA NeMo Retriever and Llama Nemotron models, it delivers high accuracy, strong reasoning, and enterprise-scale throughput, enabling organizations to move from prototype to production in weeks rather than months.[2] Advanced retrieval, reranking, and reflection techniques reduce hallucinations and ensure outputs align with your internal data and policies. The blueprint also includes governance, observability, and safety features to protect sensitive information, while GPU acceleration ensures reliable, resilient performance at scale. Flexible plug-ins and customization allow teams to adapt the solution for enterprise search, knowledge assistants, generative copilots, or vertical AI workflows—standalone or integrated into more advanced agentic applications.

2 What Is Agentic AI? | NVIDIA Blog

The image below represents the architecture and workflow.



This modular design ensures efficient query processing, accurate retrieval of information, and easy customization.

We recommend reviewing the GitHub documentation to get a better understanding of the process, key features, and system requirements. From there, you can begin deployment either on-premises following the documented steps or in the cloud.

In Conclusion

With the right foundations in place, your organization is ready to accelerate its AI journey with enterprise-ready RAG solutions. We've highlighted the challenges of enterprise AI adoption, the power of purpose-built AI Factory infrastructure, and how Supermicro and NVIDIA's integrated platforms—backed by validated RAG Blueprints—make advanced AI deployments simpler, faster, and more scalable. Supermicro and NVIDIA are leading the AI transformation, delivering the tools and platforms enterprises need to turn AI potential into real-world impact.

Whether you're optimizing for model type and size, expanding your retrieval library, managing concurrent users, or tuning input/output context windows, Supermicro and NVIDIA—and our trusted partners—offer the expertise and infrastructure to tailor solutions to your unique requirements. This collaborative approach helps ensure a smooth, high-impact deployment that aligns with your business goals.

Get Started with Your RAG Deployment Today

- **Schedule a RAG Readiness Consultation** with **Supermicro** to tailor infrastructure, GPUs, and software for your specific workload.
- **Explore AI Blueprints at Build.nvidia.com** to discover proven architectures and deployment strategies.
- **Access the NVIDIA AI Blueprint for RAG** and start building your Retrieval-Augmented Generation pipeline today.

SUPERMICRO

As a global leader in high-performance, high-efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs and enables us to build and deliver application-optimized solutions based upon your requirements. See www.supermicro.com.

NVIDIA

NVIDIA accelerated computing platforms power the new era of computing, performing exponentially more work in less time with much lower energy consumption than traditional CPU-based computing. Accelerated computing revolutionizes energy efficiency across industries by harnessing NVIDIA GPUs, CPUs, and networking, all optimized through NVIDIA enterprise software solutions.
More information at https://nvidianews.nvidia.com.