



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

Table of Contents

Executive Summary	1
Barriers to Public Sector AI Adoption	1
Unlocking Public Sector AI with Purpose-Built Infrastructure ...	2
What is RAG?	3
The Supermicro RAG solution featuring the NVIDIA RTX PRO™ 6000 Blackwell Server Edition GPU	4
Step 1: Select The Right Supermicro and NVIDIA Infrastructure	4
Step 2: Accelerate your AI development with NVIDIA Enterprise	6
Step 3: Start Building a Public Sector RAG Pipeline With NVIDIA Blueprints	7
In Conclusion	9
Get Started with Your RAG Deployment Today	9

Executive Summary

Public sector agencies are moving rapidly from artificial intelligence (AI) experimentation to operational deployment. As they seek to harness large language models (LLMs) for mission impact, they face persistent challenges—including infrastructure complexity, staffing constraints, and the need to deliver accurate, policy-aligned, and secure AI outputs.

This solution brief spotlights Retrieval-Augmented Generation (RAG) as a key enabler of trustworthy, fast, and scalable AI in government. RAG connects LLMs directly to agency data, grounding responses in authoritative sources while maintaining strict data governance and privacy. Enabled by Supermicro's flexible, U.S.-designed systems and NVIDIA's industry-leading GPUs, the AI Factory solution provides a turnkey foundation for deploying advanced RAG pipelines.

With validated software blueprints, AI-optimized networking, and full-stack support, agencies can pilot in hours or days, scale quickly, and maintain compliance with executive orders on AI and with the Cybersecurity and Infrastructure Security Agency's zero-trust architecture. By combining pre-trained models with secure knowledge retrieval, agencies can accelerate adoption, ensure reliable outcomes, and realize the full potential of AI across mission-critical operations.

Barriers to Public Sector AI Adoption

Public sector organizations face a distinct set of challenges when adopting AI. Staffing limitations in both IT and data science, combined with constraints on power, cooling, and physical space, make it difficult to deploy and sustain the infrastructure required for advanced AI workloads. Many agencies are still building experience with AI hardware and operations, making it harder to design, implement, and manage complex systems without significant support. As a result, there is growing interest in validated, turnkey solutions that reduce technical burden and accelerate implementation.



In addition to infrastructure challenges, pre-trained AI models introduce new considerations. While they can speed up adoption, these models may produce inaccurate or misleading responses that do not align with an agency's mission or policy context. Data protection and compliance remain critical concerns—especially when using open-source or cloud-based models—since public entities must safeguard sensitive information and adhere to strict governance and privacy regulations. Training large foundation models on agency data can be prohibitively costly, leading many to focus on fine-tuning smaller models. However, fine-tuning alone may not consistently deliver the level of accuracy, contextual awareness, and security needed for government decision-making and public service delivery.

These technical and security challenges contribute directly to the gap between AI investment and successful deployment. A McKinsey report highlights that 92% of organizations 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 spending less than 10% on AI initiatives. Governments are not exempt from these trends—public sector agencies often encounter similar barriers, from limited budgets to resource and expertise constraints. These figures underscore the urgent need for secure, scalable, and validated AI solutions that help organizations translate investment into successful, high-impact deployments.¹

¹ [AI in the workplace: A report for 2025 | McKinsey](#)

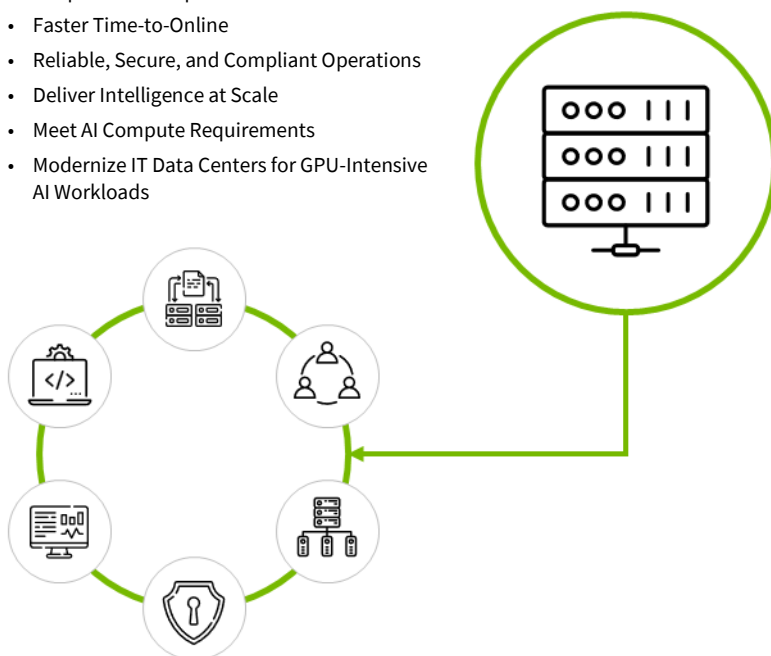
Unlocking Public Sector AI with Purpose-Built Infrastructure

AI is transforming how public agencies deliver services, manage operations, and achieve mission goals. As adoption grows, agencies need secure, scalable infrastructure to deploy advanced AI—not just for innovation, but to improve outcomes, boost efficiency, and serve communities more effectively.

AI Factories—purpose-built, scalable, and modular infrastructures—are emerging as the foundation for this transformation. Designed to meet public sector requirements, these systems provide a turnkey approach to deploying, managing, and scaling AI workloads safely and efficiently. Leveraging advanced hardware and flexible configurations, AI Factories help agencies accelerate AI adoption while maintaining security, compliance, and operational reliability.

Key Reasons AI Factories Matter to Public Sector Agencies

- Simplified AI Adoption
- Faster Time-to-Online
- Reliable, Secure, and Compliant Operations
- Deliver Intelligence at Scale
- Meet AI Compute Requirements
- Modernize IT Data Centers for GPU-Intensive AI Workloads



While AI Factories offer a compelling path forward, realizing their full potential requires overcoming the practical and technical challenges that many public sector agencies still face today.

From Pre-Trained Models to Context-Aware Intelligence

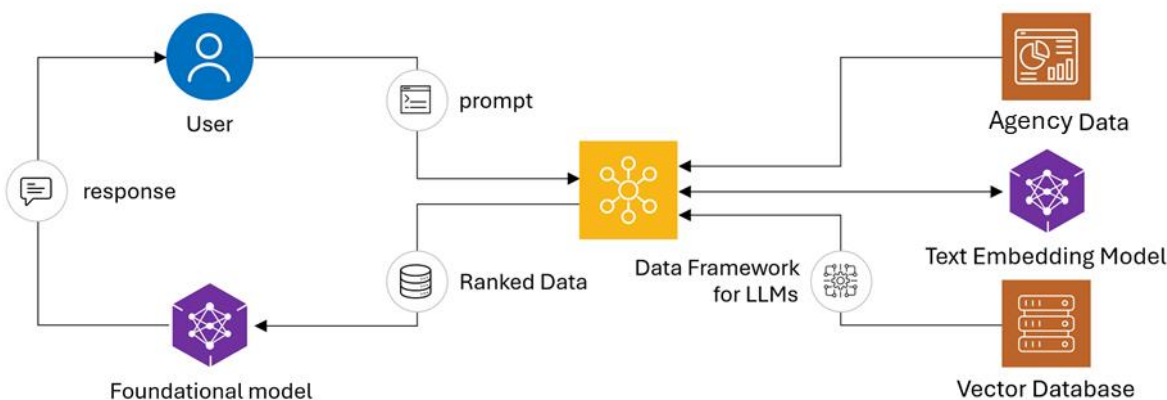
Agencies can post-train models using open-source AI and their own organizational data. Retrieval-Augmented Generation (RAG) provides a practical way to accomplish this, combining external knowledge with AI reasoning to deliver accurate, relevant, and secure results tailored to an agency's mission and public service objectives.

What is RAG?

Retrieval-Augmented Generation (RAG) connects large language models (LLMs) to an agency's own data, enabling AI to generate responses that are accurate, current, and aligned with mission context. By retrieving relevant information from internal sources—such as policy documents, case files, or knowledge bases—RAG grounds model outputs in trusted agency content. This reduces hallucinations and ensures sensitive data remains secure within government environments. RAG helps agencies deliver precise, policy-compliant 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.



Why RAG for the Public Sector, Now

RAG is ideally suited for domains where correctness, auditability, and current policy matter as much as fluency and speed—such as citizen services, records and case research, defense and civilian field operations, fraud detection, and scientific analysis. It pairs a powerful language model with a retrieval layer that pulls relevant, authorized content from your agency's own data to deliver answers you can trust.

Agencies benefit from faster time-to-value, lower costs, and greater trust because answers trace back to specific documents. Updates to a knowledge base are reflected immediately in future queries. For example, if the Department of Veterans Affairs institutes a benefits change, a traditional model requires expensive retraining. With RAG, the agency simply updates the source document, and the system responds based on the revised policy—no retraining required.

RAG systems can also be strictly confined to a specific knowledge domain. If a question falls outside the scope of the provided documents—such as asking an IRS chatbot about airport security—the system will recognize the gap and decline to answer, preserving accuracy and trust.

From Strategy to Solution

Recognizing the limitations of generic AI deployments, public sector agencies are increasingly adopting post-training strategies that combine open-source models with their own institutional data. This approach enables agencies to tailor AI outputs to their specific mission context while maintaining control over data privacy, relevance, and compliance.

Retrieval-Augmented Generation (RAG) plays a central role in this strategy, allowing pre-trained models to dynamically access and incorporate agency knowledge at runtime. However, unlocking the full potential of RAG requires infrastructure that is powerful, scalable, and ready to deploy—without adding operational 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 agencies 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 public sector RAG deployments, talk about leveraging [NVIDIA 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 in enabling RAG across your agency is selecting infrastructure that meets your performance, scalability, and workload requirements. Supermicro offers a range of NVIDIA-Certified Systems optimized for RAG—from GPU count and network bandwidth to form factor—featuring the latest NVIDIA RTX PRO 6000 Blackwell Server Edition.







Supermicro's building block approach allows agencies to tailor systems to mission environments, accommodating power, thermal, and space constraints across data centers and edge sites. U.S.-based design and manufacturing support federal acquisition requirements, while secure supply chains and compliance with standards like IPv6, CMMC 2.0, FIPS 140-2/140-3, and Section 508 streamline approvals for sensitive workloads.

This alignment with federal expectations helps agency leaders deploy and scale AI confidently and efficiently.

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 Enterprise software. This enables agencies to simplify the development and deployment of production-grade AI while meeting performance, reliability, and compliance requirements.

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
<div>2-2-3</div> <div></div> <div></div>	<div>2-4-3</div> <div></div> <div></div>	<div>2-8-5</div> <div><div><div>Intel</div><div></div></div><div><div>AMD</div><div></div></div></div>			
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 2x 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.			
<div>2-2-3 architecture:</div> <div><ul style="list-style-type: none">• 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</div>	<div>2-4-3 architecture:</div> <div><ul style="list-style-type: none">• 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</div>	<div>2-8-5 architecture:</div> <div><ul style="list-style-type: none">• 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</div>			

[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 agencies 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

agencies looking to run a range of public sector 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 accelerates generative AI and high-performance computing (HPC) workloads with breakthrough performance and memory capabilities. It is well-suited for agencies training foundational models or deploying large-scale inference (models exceeding 70B parameters), supporting mission-critical applications that demand speed, scale, and reliability.



- 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 delivering up to 30x more AI Factory output than the previous generation, NVIDIA Blackwell Ultra-based HGX systems are built for the most demanding generative AI, data analytics, and HPC workloads. Designed for agencies with intensive AI training and inference needs, this is NVIDIA's most powerful GPU for mission-critical model development and deployment.



- 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 Enterprise

NVIDIA Enterprise is a cloud-native suite of software tools, libraries, and frameworks—including NVIDIA NIMs and NeMo microservices—that accelerates and simplifies the development, deployment, and scaling of AI applications. Agencies of all sizes can deploy agentic AI systems across clouds, data centers, or edge environments, leveraging a broad partner ecosystem. By delivering high-performance solutions, NVIDIA empowers the public sector to redefine citizen services, strengthen cybersecurity, advance geospatial intelligence, and address society's most complex challenges. NVIDIA Enterprise helps reduce infrastructure costs, accelerate time to mission impact, and ensure reliable, secure, and scalable AI operations.

Environment Overview

NVIDIA 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 mission-critical workloads. The cloud-native design integrates with leading orchestration platforms, allowing AI applications to run seamlessly across clouds, on-premises infrastructure, or edge environments. Extended-life software branches, proactive security updates, and tested deployment guidance make it government-ready, while the robust partner ecosystem supports hardware, software, and system integration aligned with federal standards and lifecycle assurance.

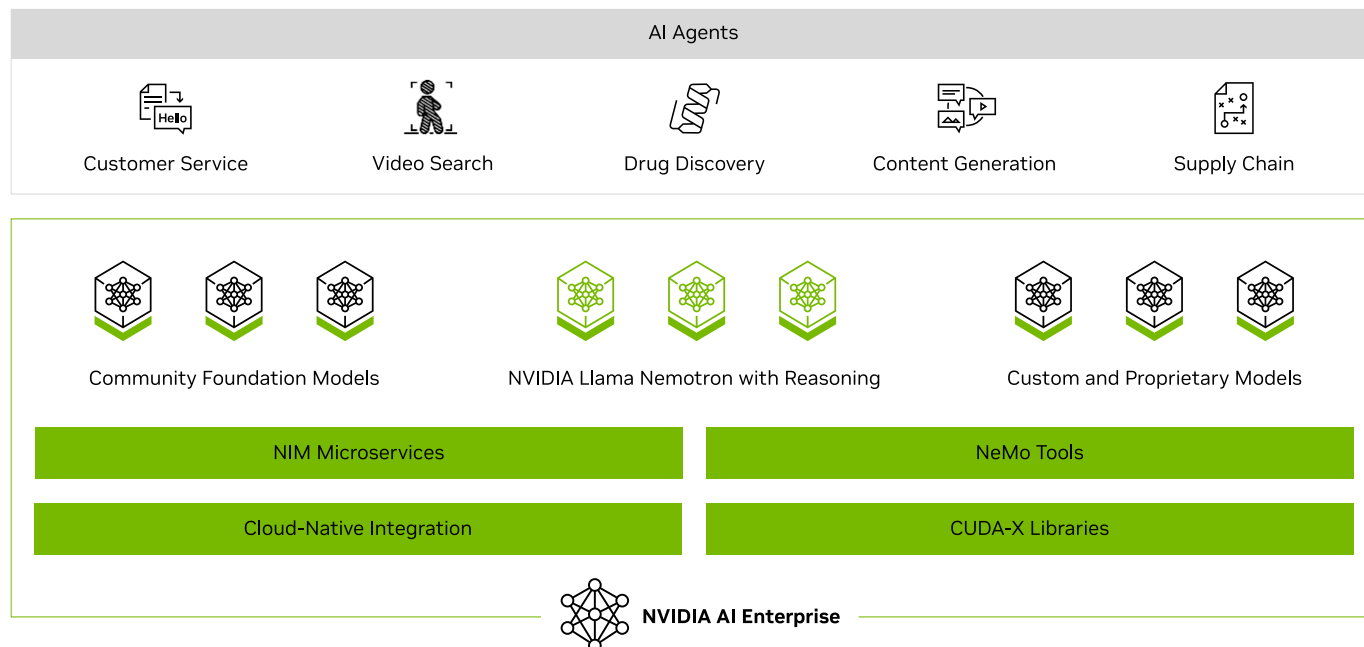
Learn more: [NVIDIA Enterprise | Cloud-native Software Platform | NVIDIA](#), or get started with NVIDIA Enterprise: [Get Started With NVIDIA Enterprise | NVIDIA](#).

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

Production-Ready Software for Agentic AI



Step 3: Start Building a Public Sector RAG Pipeline With NVIDIA Blueprints

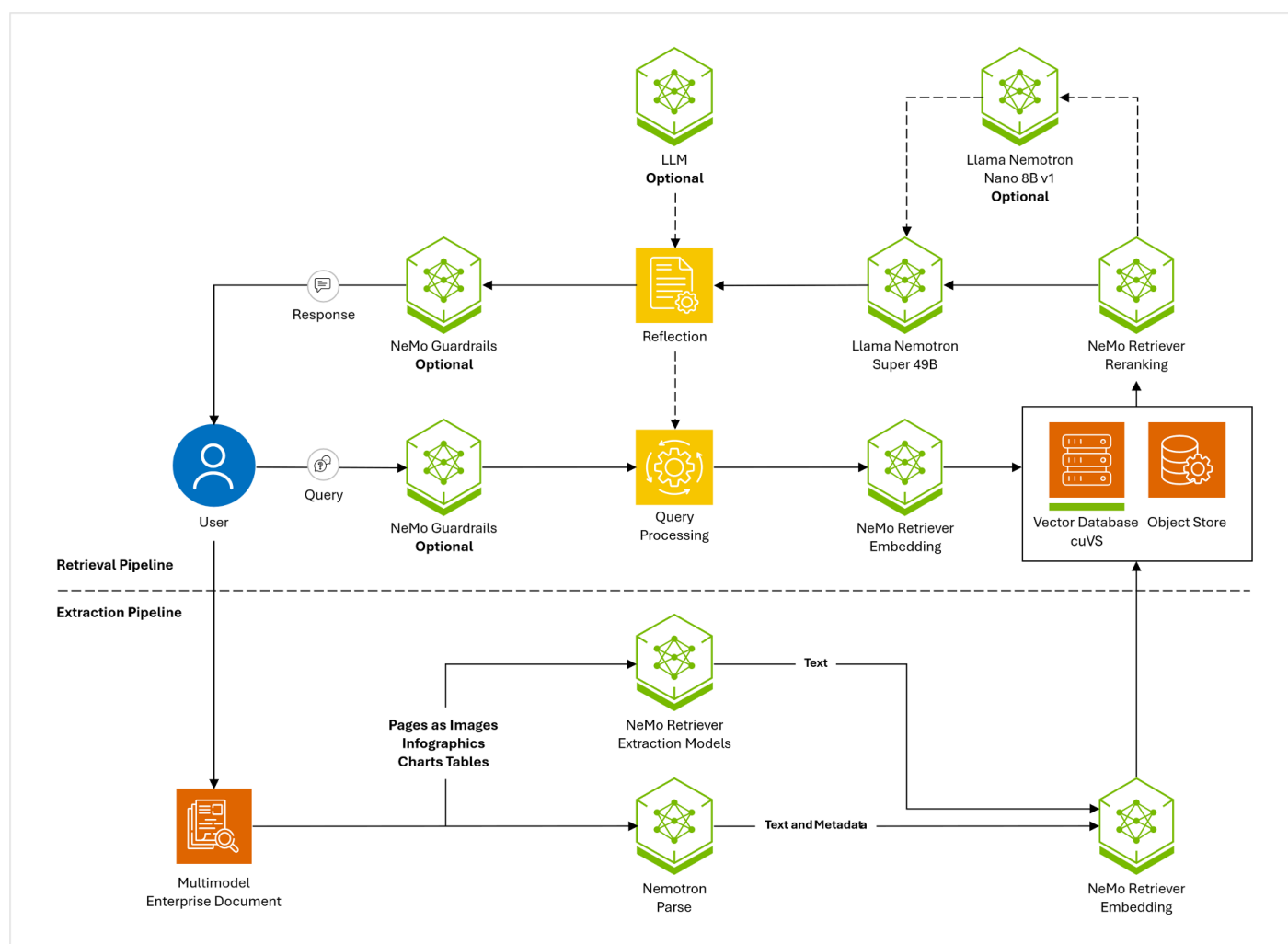
Begin by visiting [Build.nvidia.com](#) and selecting the [NVIDIA AI Blueprint for RAG](#). This NVIDIA Blueprint provides developers with a foundational starting point for building scalable, customizable data extraction and retrieval pipelines using NVIDIA NeMo Retriever models. Use the NVIDIA Blueprint to connect LLMs to extensive multimodal agency data—including text, tables, charts,

and millions of PDFs—to deliver context-aware responses. Agencies can unlock actionable insights with 15x faster multimodal PDF data extraction and 50% fewer incorrect answers, driving productivity and mission impact at scale.

The NVIDIA AI Blueprint for RAG provides a production-ready workflow for building agency-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 scalable throughput, enabling agencies 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 internal data and policy. 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 internal 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, agencies can begin deployment either on-premises, following the documented steps, or in the [cloud](#).

In Conclusion

With the right foundations in place, your agency is ready to move from AI experimentation to operational deployment. Retrieval-Augmented Generation (RAG) offers a practical path forward—enabling fast, trustworthy, and policy-aligned AI by grounding responses in agency data and keeping sensitive information within secure environments.

We've explored the challenges of scaling AI in government, the importance of purpose-built infrastructure, and how Supermicro and NVIDIA's integrated platforms—backed by validated RAG Blueprints—make deployment simpler, faster, and more secure. Their U.S.-designed, full-stack solutions combine servers, software, and AI-optimized networking to help agencies pilot in hours or days, scale confidently, and maintain compliance with executive orders on AI and with the Cybersecurity and Infrastructure Security Agency's zero-trust architecture.

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 their trusted partners—offer the expertise and infrastructure to tailor solutions to your mission requirements. This collaborative approach ensures high-impact deployments that align with agency goals, accelerate time-to-value, and support long-term lifecycle assurance.

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 a Retrieval-Augmented Generation pipeline tailored to your agency's mission 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>.