WHITE PAPER

Operationalizing AI in Government:

RAG and Trusted Infrastructure as the Fastest Path to Mission Value

Executive summary

Public sector agencies are moving rapidly from artificial intelligence (AI) experimentation to operational deployment. Retrieval-augmented generation (RAG) enables trustworthy, fast, and policy-aligned AI by grounding responses in agency data and keeping information inside secure environments. Supermicro and NVIDIA deliver a U.S.-designed, full-stack platform: servers, validated software blueprints, and AI-optimized networking, so agencies can pilot in hours or days, scale quickly, and maintain compliance with executive orders on AI and the Cybersecurity and Infrastructure Security Agency zero trust architecture.

Table of contents

Why RAG for the public sector, now . . . 2

How RAG enables AI in sensitive environments . . . 3

Proof points: scale, reliability, and early RAG wins . . . 4

From pilot to “AI factory”: what changes with RAG . . . 4

Supermicro: Unique advantages for government RAG . . . 5

Partnerships that de-risk delivery . . . 6

Trusted infrastructure for secure AI . . . 7

Put AI to work on mission challenges . . . 9

Key takeaways for agency leaders

RAG for mission accuracy and control:

Provides grounded, cited answers while keeping sensitive data securely in-environment

Up-to-date policy alignment:

Updates to authorized documents instantly ensure responses reflect current policy

Compliant infrastructure:

U.S.-designed systems with transparent supply chains and federal readiness, including IPv6, CMMC, FIPS, and Section 508 compliance

Speed to value:

Validated NVIDIA RAG Blueprint on Supermicro systems enable pilot deployments in days and production in weeks

AI in government is moving rapidly from experimentation to operational impact. Public sector agencies need fast, accurate, and trusted answers, without the risk of sending sensitive data outside or the burden of retraining massive models.

From benefits processing to public safety operations, **retrieval-augmented generation (RAG)** is a practical way to meet these needs in secure enclaves, denied or disconnected environments (DDILs), and hybrid deployments. When queried, a RAG system retrieves the most relevant, authorized passages from agency repositories and drafts an answer that cites its sources. This approach keeps data fully contained within agency environments while delivering rapid, trusted results.

Supermicro and **NVIDIA** offer a full-stack, RAG-ready path, from U.S.-designed and manufactured systems to validated software blueprints, enabling agencies to pilot quickly, scale efficiently, and maintain security in air-gapped or hybrid environments.

WHY RAG FOR THE PUBLIC SECTOR, NOW

RAG 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. It is ideally suited for citizen services, records and case research, defense and civilian field operations, fraud detection, and scientific analysis – domains where correctness, auditability, and current policy matter as much as fluency and speed.

Agencies also benefit from time to value (no months-long model retraining), lower cost, and improved 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 to incorporate new policies. With RAG, the agency updates the source document, and the system immediately answers based on the revised policy – without retraining.

A RAG system can be strictly confined to a specific knowledge domain by grounding its answers exclusively in the provided documents. If a question is asked that cannot be answered using the information in its knowledge base – such as asking an Internal Revenue Service policy chatbot about airport security – the system will recognize that it has no relevant information to draw from and will decline to answer.

HOW RAG ENABLES AI IN SENSITIVE ENVIRONMENTS

RAG enables AI deployment in classified or disconnected environments by keeping all data inside the enclave. Only the minimal text fragments required for a response are processed, and each output is verifiably linked to its source document.

This enables AI in places where cloud retraining is not possible, including:

- Mission systems on SIPRNet and NIPRNet
- Disconnected networks where cloud retraining is not an option
- Regulated datasets with strict residency and dissemination controls that forbid exporting documents to external vendors
- Use cases requiring full source provenance for every answer to pass legal or policy review

It is a common misconception that most generative AI platforms automatically and continuously learn from user inputs. In reality, for critical privacy and security reasons, these models are typically static. Their knowledge is updated only through deliberate fine-tuning or retraining processes, not through casual interaction.

This creates a key security advantage of the RAG approach: Because the core AI model is never trained on sensitive information, the risk of it internalizing and later exposing that data in unrelated contexts is effectively eliminated.

However, assuming this removes the need for protection is a serious mistake. With RAG, the security focus shifts from safeguarding a "knowledgeable model" to rigorously protecting the document knowledge base and managing the retrieval process, which is now the direct source of any sensitive information the system can access and relay.

The most important controls include access controls on data, encryption at rest and in transit, and deployment on a secure, trusted hardware stack.

AI guardrails on RAG systems can be engineered to mitigate the risk of classified data aggregation by actively monitoring and controlling the information retrieved and synthesized.

These guardrails can enforce rules that prevent the system from combining information from different sources or of certain types in a single response, effectively blocking the formation of a sensitive "mosaic" and ensuring the output remains at the unclassified level of the individual source documents.

Enterprise RAG for federal missions: Unlocking knowledge at scale

A system that combines retrieval of internal data (documents, knowledge bases, etc.) with generative AI to provide accurate, context-rich answers within a secure environment and providing full control over data usage.

Why it matters

Improves decision-making

- Real-time access to agency-specific knowledge
- Reduces guesswork and manual search

Boosts productivity

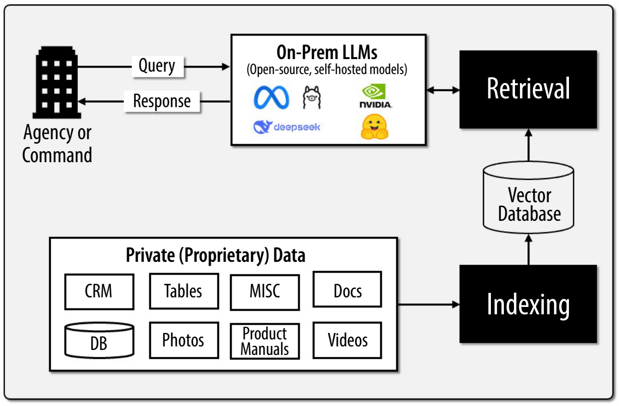
- 30-50% reduction in information retrieval time
- Automates complex Q&A and summarization tasks
- Frees up employees from repetitive document reading

Ensures accuracy & compliance

- Anchored in trusted enterprise data
- Supports version control and auditability

Scales knowledge across the organization

- Consistent access to tribal knowledge
- Breaks down information silos



PROOF POINTS: SCALE, RELIABILITY, AND EARLY RAG WINS

The Supermicro and NVIDIA combination underpins landmark federal high-performance computing and AI environments, including deployments at national labs and science agencies, demonstrating delivery of secure, high-performance infrastructure for demanding missions.

The same engineering discipline now shows up in smaller, RAG-specific builds for enterprise and agency workloads.

Early wins in RAG-style workloads include:

- Federal civilian e-discovery: A federal civilian agency deployed a RAG-based system to assist its legal team with e-discovery across millions of internal documents. The team reported a 60% reduction in time for initial case research, enabling a larger caseload with the same staffing.
- Defense technical analysis: A DOD component implemented a RAG pilot to synthesize large volumes of unstructured technical intelligence reports. Analysts can ask complex questions and get cited answers in minutes instead of days of manual reading and keyword search.

FROM PILOT TO “AI FACTORY”: WHAT CHANGES WITH RAG

An AI factory is a standardized, secure platform that transforms agency data into mission intelligence. This approach supports the pillars on innovation and infrastructure in America’s AI Action Plan, prioritizing faster deployment while building the secure compute, storage, and networking backbone agencies need to scale AI.

RAG is one of the most repeatable “products” of an AI factory: With a secure pipeline for data ingestion, indexing, retrieval, and generation, new use cases spin up by pointing the pipeline to a new dataset.

Pre-built, enterprise-grade components, such as the NVIDIA AI Enterprise 5.x RAG Blueprint (2025), ease the process of standing up a working RAG environment. Agencies can move from proof of concept to a production-ready, enterprise-supported service in weeks instead of months (and days for constrained pilots) because the blueprint covers orchestration, evaluation, guardrails, and operations patterns out of the box.

How to start small and scale fast

✓ Begin with a narrowly defined mission scenario and assemble a compact pilot on a small number of GPUs.

✓ Measure outcomes such as latency, grounded accuracy, citation coverage, and analyst throughput.

✓ Add capacity as demand grows – often faster than anticipated – by sliding in additional, pre-validated nodes without re-architecting the pipeline.

SUPERMICRO: UNIQUE ADVANTAGES FOR GOVERNMENT RAG

Building block approach

Supermicro's building block approach lets agencies tailor systems for their environment, accounting for component choices, power and thermal constraints, form factors for data centers or edge sites, and rack-level integration. This flexibility is essential for unique and varied government missions when space, power, and cooling capacity are limited.

This design- and engineering-led approach also supports designing for data precision: stable, low-latency input/output for vector search; storage tiers matched to data age and value; and predictable networking that keeps retrieval tight, so answers cite the most relevant passages instead of drifting to approximations.

U.S. design and manufacturing and a secure supply chain

Supermicro designs, engineers, manufactures, and tests systems in the United States, with traceability and on-shore validation that support federal acquisition requirements and lifecycle assurance. Country-of-origin documentation and Made-in-USA SKUs simplify risk reviews and accelerate approvals for sensitive workloads.

In addition, programs benefit from transparent pricing and documentation aligned to federal expectations, plus readiness for network and security requirements common in government environments. This includes IPv6 readiness, CMMC 2.0 compliance, alignment with FIPS 140-2/140-3 cryptographic module requirements, and Section 508 accessibility at the solution level in concert with federal systems integrator (FSI) partners and customer controls. This alignment tracks with the AI Action Plan's push to accelerate the adoption of AI systems across government, making it easier for mission owners to buy, deploy, and scale AI capabilities.

PARTNERSHIPS THAT DE-RISK DELIVERY

AI programs succeed when each partner brings its specialty to the mission: secure, compliant hardware; validated AI software and patterns; and integration to agency data and safeguards. That division of labor reduces risk and streamlines the journey from first pilot to an operational, enterprise-backed service. With Supermicro and NVIDIA, here are the partnership roles at a glance:

NVIDIA provides the accelerated computing platform, enterprise software, AI-optimized networking, and validated blueprints.

Supermicro delivers optimized, secure, compliant, and NVIDIA-certified servers that house the NVIDIA engine and come pre-validated to run the NVIDIA suite smoothly and efficiently.

FSIs customize the solution for the agency's unique mission data and requirements.

Practically, **FSIs** use a reference architecture to stand up a proof of concept on Supermicro hardware with the NVIDIA AI Enterprise 5.x RAG Blueprint (2025), capture metrics and user feedback, and then harden the design into a full, enterprise-supported application by adding constant security checks, automatic software updates, and forecasting for future needs so the system can handle growth. The transition from pilot to production follows a known playbook, reducing risk and staff burden.

Together, these roles advance the AI Action Plan's goal to establish American AI as the global standard by deploying an American full-stack – hardware, software, and models – into mission environments.

In addition, programs benefit from transparent pricing and documentation aligned to federal expectations, plus readiness for network and security requirements common in government environments. This includes IPv6 readiness, CMMC 2.0 compliance, alignment with FIPS 140-2/140-3 cryptographic module requirements, and Section 508 accessibility at the solution level in concert with federal systems integrator (FSI) partners and customer controls. This alignment tracks with the AI Action Plan's push to accelerate the adoption of AI systems across government, making it easier for mission owners to buy, deploy, and scale AI capabilities.

Supermicro advantage

Flexible

- Design/engineering-led approach
- Form factor options
- Low-latency I/O for vector search
- Storage tiers matched to data profile
- Predictable, high-performance networking

Fast

- First to market for rapid acceleration technologies for faster time to online

Secure

- U.S. manufacturing
- Trade Agreements Act compliance
- County of origin documentation
- IPv6 readiness
- CMMC 2.0 compliant
- FIPS 140-2/140-3 compliant

TRUSTED INFRASTRUCTURE FOR SECURE AI

Together, Supermicro and NVIDIA deliver a turnkey platform for government enterprise RAG, including Supermicro's servers, the NVIDIA Blueprint for RAG, a reference architecture/software accelerator, and an AI-optimized networking fabric.

Supermicro delivers this trusted stack with emphasis on a secure supply chain that provides assurance from the silicon up, including:

- **Host layer:** Supermicro platforms support UEFI Secure Boot and TPM 2.0 measured boot, enabling hardware-rooted attestation of the firmware and operating system before workloads run
- **GPU layer:** NVIDIA H100/HGX systems add a GPU hardware root of trust, verified GPU boot, and confidential computing with attestation so agencies can prove the accelerator and driver stack are in a known-good state
- **Supply chain:** Supermicro provides cryptographic attestation of components and firmware and offers U.S. manufacturing with traceability and country-of-origin documentation through the MITU program

A Defense Department component can point RAG at data regardless of classification, from the secret classifications to unclassified-but-sensitive documents such as technical reports and maintenance bulletins. For example, RAG can answer equipment readiness questions with cited passages, without exposing documents beyond its enclave.

The compliant-by-design solution simplifies the government procurement process with clear documentation, reducing administrative burden and putting mission-critical systems into operators' hands. Agency leaders are achieving faster time to pilot, more consistent and predictable performance, and simplified operations with a single integrated stack and a clear path to scale from a single server to a multi-node cluster without re-architecting.

The solution components include:

GPU platform: NVIDIA RTX™ PRO 6000 Blackwell Server Edition

Supermicro is the first to market with NVIDIA RTX PRO 6000 Blackwell Server Edition-based systems, with more than 20 optimized systems. For many RAG inference services, the RTX PRO 6000 Blackwell GPU provides an attractive performance, price, and power profile in PCIe systems for government – ideal for robust pilots and departmental rollouts where space and power are constrained, with a clean path to scale-out. It complements larger HGX-class systems where needed, giving programs a continuum from workstation to rack.

Supermicro SYS-522GA-NRT

Supermicro servers

Supermicro's latest servers are designed for high-performance NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, delivering peak performance in tough environments and supporting up to 35°C ambient temperature. These PCIe-optimized servers can hold up to eight NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs in a 4U or 5U form factor chassis.

Supermicro's high-ambient-temperature designs enable the operation of enterprise AI systems in hot environments without restricting GPU performance, which also helps reduce data center and cooling infrastructure costs. These Supermicro systems are optimized for maximum thermal efficiency and airflow, making it easier to deploy latest-generation GPUs in existing enterprise data centers.

Software: NVIDIA AI Enterprise & RAG Blueprint

NVIDIA's enterprise stack provides containerized microservices, orchestration, evaluation, guardrails, and model operations that agencies deploy on Supermicro-validated systems. The NVIDIA AI Enterprise 5.x RAG Blueprint (2025) accelerates stand-up and standardizes best practices so pilots come together in hours or days, not months, and transition cleanly to production. NVIDIA also has blueprints to help refine and optimize RAG solutions post-deployment, supporting users from pilot to deployment through continuous improvement.

Networking: NVIDIA Spectrum-X Ethernet

As services scale beyond a single server, predictable, low-latency networking keeps GPUs and retrieval layers fed, sustaining end-to-end responsiveness. Spectrum-X is engineered for AI traffic patterns and integrates with the overall stack to maintain consistent performance as concurrency grows.

Storage and data pipeline considerations

RAG's retrieval step demands fast, reliable access to indexes and source documents. Supermicro's storage portfolio and rack-integration expertise allow agencies to balance high-performance flash tiers for vector search with capacity tiers for large datasets, meeting availability targets and retrieval service level agreements.

Right-sizing RAG starts with three inputs

Agencies can scope any RAG deployment by addressing three practical requirements:

1. Which model family and size best fit the mission task?
2. How many concurrent users are expected and what latency targets are desired?
3. How many GPUs are needed to meet those targets?

From there, storage input/output operations per second/throughput and the network fabric are sized to keep retrieval and generation responsive. This method avoids both under- and over-provisioning and creates a clear "start small, scale fast" path.

Right-size the GPU tier: Start with **RTX PRO 6000 Blackwell** for pilots and departmental RAG services where power and space are constrained. Move to **HGX-class nodes** when concurrency, model size, or training needs require NVLink-class scaling and rack-level efficiency.

PUT AI TO WORK ON MISSION CHALLENGES

RAG, combined with an AI factory model, delivers a practical, secure, scalable, and cost-effective path for agencies to apply AI to their own data without retraining models or compromising control.

With Supermicro's U.S.-based manufacturing and customization and NVIDIA's accelerated computing and software, public sector agencies gain a best-in-class, full-stack RAG solution that delivers measurable mission impact with AI faster, more accurately, and with confidence.

Challenge	RAG solution	Supermicro & NVIDIA value
Sensitive data environments	RAG keeps data in secure enclaves and retrieves only relevant snippets	U.S.-designed and manufactured systems ensure compliance and controlled processing
Slow, costly model retraining	RAG updates knowledge via source documents; no retraining required	Validated NVIDIA Blueprint accelerates pilots and scaling
Compliance and security mandates	RAG aligns with executive orders on AI and zero trust pillars	Integrated hardware-software stack simplifies audits and procurement

Super Micro Computer, Inc.
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 on your requirements.
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.
www.nvidia.com

Printed in USA ♻ Please Recycle

SUPERMICRO® NVIDIA