Improving accuracy, efficiency and trust in generative AI

CTG FEDERAL

NVIDIA

RAG and AI-ready infrastructure can break down barriers to federal AI adoption

With vast volumes of enterprise data and readily accessible large language models (LLMs), now is a seminal moment for generative AI in federal government. More than ever, AI has enormous potential to help agencies harness their data to enable employees, enhance citizen services and achieve more with less. But as organizations consider adopting these systems for their own needs, decision-makers face one looming question: Can LLMs be trusted in mission-critical, high-stakes scenarios?

Traditional LLM Limitations

LLMs excel in pattern recognition and generating human-like text—but they haven't always been reliable. Traditional LLMs are typically less accurate and efficient, with limited memory and context length. They are trained on static data from the past, but nothing after that—which makes retrieving real-time information impossible. Conventionally, LLMs also use probabilistic modeling, which produces responses based on likely word sequences, but with no concept of factual correctness. This can cause AI models to generate answers that sound plausible but are not always true—a phenomenon known as "hallucinating." For obvious reasons, agencies can't afford to take this risk in missions that rely on timely, detailed, factual information.

To compound the issue, implementing enterprise-scale AI can be costly and complex—especially without the right infrastructure. The last thing agencies need is a slow and expensive gen AI deployment that produces unreliable information.

When facing these challenges, government agencies need generative AI solutions that can:

Provide accurate answers from verified sources

Generate responses using continuously up-to-date information

Streamline AI deployment across the enteprise

Quickly refine and scale gen AI models

Together, emerging solutions like retrieval augmented generation (RAG) and accelerated infrastructure can address these issues and meet government requirement in trust, cost and performance.

Building Trust and Accuracy with Retrieval Augmented Generation (RAG)

RAG is a more controlled and verifiable version of AI that improves the quality and credibility of LLM outputs. It uses internal enterprise data to answer questions with information that's current, reliable and in context. RAG puts controls around data retrieval to ensure information is pulled only from vetted sources, with limited exposure to external data. It also adds context by deploying guardrails specific to a given use case, ensuring AI responses relate only the use case in question. This guides the LLM to produce answers that are accurate and situationally relevant.

Rather than relying solely on pre-trained data, RAG retrieves information dynamically from the agency's current internal knowledge base—including domain-specific knowledge the LLM wasn't trained on. This allows it to present the most current information available at the moment of query and continuously update its responses as enterprise data evolves. Additionally, RAG responses come with source attributions, so every answer is grounded in a verifiable and tracible source within the organization.

When implemented properly, RAG can help agencies apply an LLM to all of the structured and unstructured data in their enterprise—and turn it into a trusted source of truth.

CTG Federal/NVIDIA Solutions for Accelerated RAG Deployment

CTG Federal is a small business that serves dozens of federal defense, intelligence and civilian organizations with IT expertise and solutions. We've partnered with NVIDIA, a world leader in AI computing, to offer federal agencies a faster path to deploying powerful, scalable AI infrastructures.

As an Authorized NVIDIA Partner, CTG Federal can help organizations harness NVIDIA to efficiently deploy RAG-powered LLMs on an enterprise scale. efficiently deploy RAG-powered LLMs on an enterprise scale.

SOLUTION OVERVIEW

As the world's most advanced platform for generative AI, NVIDIA AI makes RAG-powered AI easy to deploy, manage and refine as needed. Innovations are built into every layer of the stack—including accelerated computing, essential end to-end AI software, pretrained AI models and foundries that enable users to quickly build, customize and run gen AI models for any application, anywhere.

NVIDIA INFERENCE MICROSERVICES (NIM)

This set of easy-to-use microservices is designed to accelerate the deployment of gen AI models across clouds, data centers and workstations. As part of NVIDIA AI Enterprise, NIM supports a wide range of AI models, including NVIDIA AI foundation and custom models. It ensures seamless, scalable AI inferencing, on-premises or in the cloud, leveraging industry-standard APIs.

AL MODELS AND FOUNDRIES

With NVIDIA pretrained models and AI foundries, developers can build and customize gen AI models for any application, anywhere in the organization. This lets developers tailor a model to a given use case and customize it with proprietary data, so it's tuned to the agency's mission, terminology, customers, domain expertise and security standards.

NVIDIA AI foundries are equipped with generative model architectures, tools and workflows, and they run on NVIDIA accelerated infrastructure for training, customizing, optimizing and deploying gen AI. Foundation models can be used out of the box or fine-tuned on proprietary datasets to return even more precise responses and lower operating costs by minimizing token usage.

- **NVIDIA NeMo** (with NeMo Retriever/NIM) empowers employees with continuously up-to-date LLMs. It grounds AI assistants in approved repositories so employees can ask natural-language questions and receive permission-aware, citation-backed answers. The system retrieves relevant passages from sources like SharePoint, wikis, tickets and policy PDFs, then generates accurate responses that stay current without the need to retrain.

- **NVIDIA Nemotron** is a family of foundation models for building production-ready generative AI, including applications for multilingual information retrieval.

NVIDIA AI ENTERPRISE

This enterprise software platform accelerates the development and deployment of production-grade generative AI. To help public sector developers get started, NVIDIA AI Enterprise offers access to 100+ NVIDIA AI frameworks, libraries, pretrained models and open source tools. Developers can then deploy these resources using the AI Enterprise platform, which also offers enterprise-grade support, security, stability and manageability. Additionally, it's compatible with industry-leading infrastructure solutions.

ACCELERATED INFRASTRUCTURE

The **NVIDIA DGX platform** incorporates the best of NVIDIA software, infrastructure and expertise into a unified AI development and training solution. Every aspect of the DGX platform is infused with NVIDIA AI expertise—featuring world-class software, record-breaking NVIDIA accelerated infrastructure and direct access to NVIDIA DGXperts. By accelerating development, the DGX platform helps agencies realize value faster throughout their enterprise.

NVIDIA-CERTIFIED™ SYSTEMS

These accelerated systems from leading NVIDIA partners are certified to deliver maximum performance, reliability and scale. They're optimized to run generative AI solutions in the cloud, on workstations, and in on-prem data centers. Consulting and service delivery are available via NVIDIA's worldwide, vibrant partner ecosystem that puts solutions together for organizations in concert with their in-house personnel. NVIDIA partners like CTG Federal know the NVIDIA AI stack well—allowing them to provide implementation services and support that best leverage accelerated infrastructure anywhere.

NVIDIA RAPIDS

NVIDIA RAPIDS accelerates data analytics and AI workflows for government agencies, enabling them to process vast amounts of data at high speeds and lower costs. By harnessing the power of GPUs, RAPIDS helps organizations perform data-intensive tasks more efficiently, achieve results faster and get more done with the same investment.

NVIDIA RIVA

Riva is a collection of GPU-accelerated microservices for building real-time, customizable speech AI applications. It allows agencies to deliver world-class accuracy and lifelike voices with industry-leading automatic speech recognition (ASR), text-to-speech (TTS), and neural machine translation (NMT). They can all be customized for the organization and seamlessly integrated with LLMs and RAG. It can be deployed anywhere, from cloud to data center to edge.
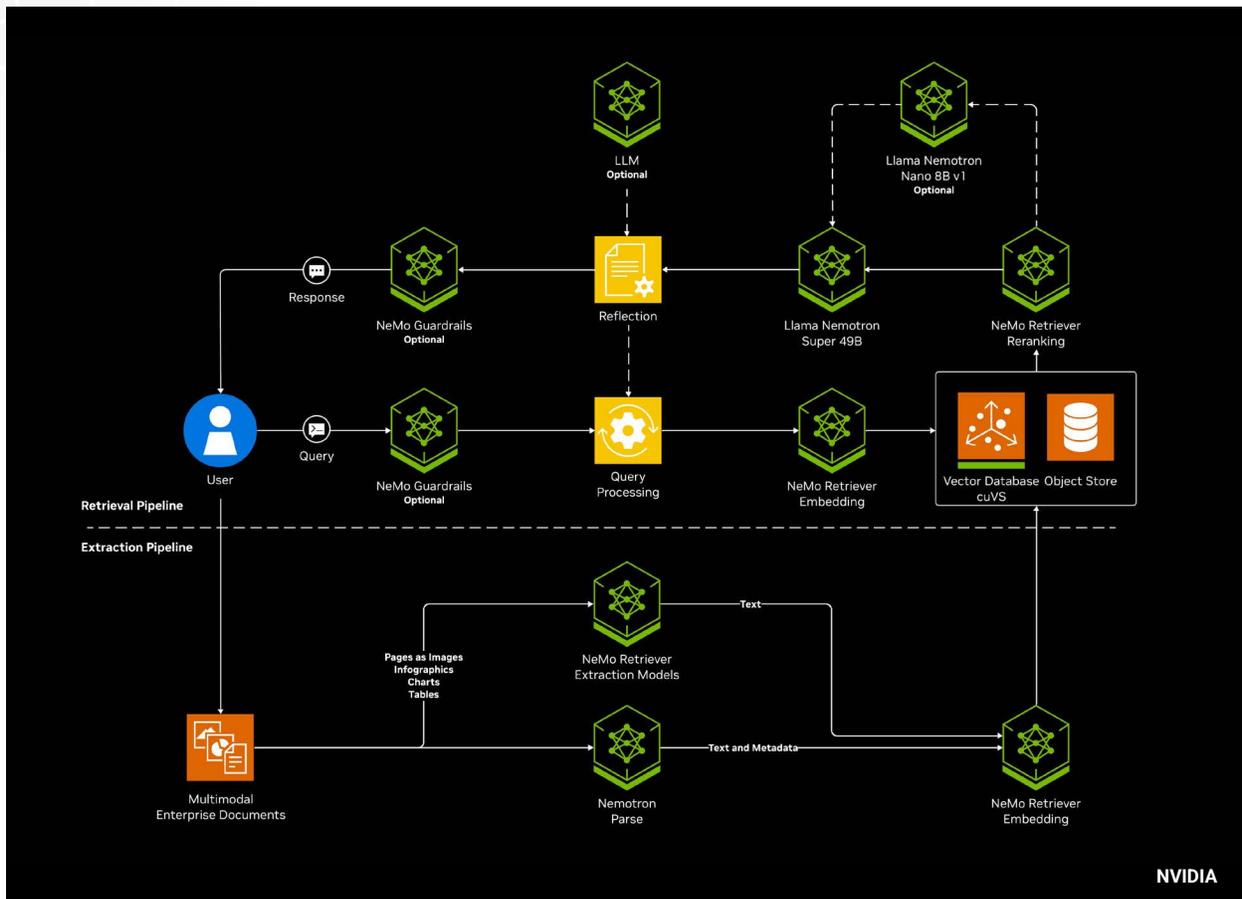
RAG Key Benefits

Accelerated AI deployments will allow agencies to speed time to value from their RAG systems. Once in place, RAG can quickly help them…

- **Improve factual accuracy and reduce hallucinations.** By pulling information directly from authoritative sources, RAG dramatically decreases the risk of false information. The system can also cite the exact source documents for its answers, so users can feel confident in the response and verify it independently if desired.

- **Offer up-to-date information.** Government laws and policies change frequently. RAG systems can access and incorporate the latest information from regularly maintained government data sources, ensuring citizens receive timely and relevant guidance. This also ensures employees can easily access the most current policies and requirements related to their jobs.

- **Empower employees.** Internal RAG tools can help staff quickly extract precise answers from vast, siloed documentation. This increases operational efficiency by reducing the time employees spend on manual searches and allows domain experts to focus on more strategic tasks. It also streamlines and improves knowledge management across the organization.

- **Simplify complex information for citizens.** RAG-powered chatbots and virtual assistants can help citizens navigate complicated policies and processes, such as filing taxes or determining benefits eligibility. By providing clear, accurate answers grounded in official documents, RAG can demystify bureaucracy and increase public understanding.

- **Efficiently respond to public queries.** A RAG-based system can significantly reduce the time citizens spend manually searching for information across government websites. Faster resolution, along with source attribution, can contribute to higher user satisfaction and trust.

- **Ensure traceability and accountability.** A RAG system can explain how it arrived at a particular answer by referencing the specific documents used. This builds credibility for the agency and demonstrates accountability to citizens.

Solution Key Benefits

This CTG Federal/NVIDIA generative AI solution delivers numerous significant benefits to federal agencies:

- **Faster time to solution**— Quickly build custom enterprise-grade models grounded in your own data and domain expertise.

- **Ease of use** — Simplify development with a suite of model-making services, pretrained models, cutting-edge frameworks and APIs.

- **Proper protections** — Safeguard data, privacy and intellectual property with enterprise-grade security and limited exposure to external data. With Sovereign AI, your organization can produce AI models using domestic infrastructure, data and networks—ensuring your data remains in a controlled, compliant environment. NVIDIA can support deployments in the private cloud and on-premises infrastructures, so PII and sensitive government information stay contained within your enterprise.

- **Reduced costs** — NVIDIA's full-stack architectural approach ensures AI-enabled applications deploy with optimal performance, fewer servers and less power computation—resulting in faster insights and dramatically lower costs.

- **Flexible deployment options** — NVIDIA supports AI applications across public and private clouds, on-premises data centers, workstations and the network edge—so agencies can run workloads wherever they need to.

- Rapid implementation — **CTG Federal provides turnkey solutions with bundled services to stand up the infrastructure** and help customers quickly start have validated solutions that integrate natively with NVIDIA solutions. In many cases, vendors also provide validated reference architectures.

Use Cases

CTG Federal and NVIDIA can help agencies redefine citizen services, employee productivity and more. Together, we make RAG-powered AI simple to adopt in a wide range of use cases:

Accelerating scientific discovery

Researchers across scientific organizations manage petabytes of unstructured data, experiment logs, telemetry and simulations. Using NVIDIA NeMo, teams can deploy RAG assistants that allow scientists to query years of research in natural language, correlate findings across disciplines and accelerate innovation.

Intelligent patent search and review

Patent examiners evaluate millions of existing patents and publications. A RAG-powered knowledge system built on NVIDIA NIMs can index this entire corpus, enabling semantic search and question answering. Examiners can now ask questions like, "Has a similar mechanism for optical stabilization been filed before 2018?" and receive cited results with document links.

Technical manual search and maintenance support

Maintenance and engineering personnel often work on equipment spanning decades of design changes and multiple manufacturers. A RAG assistant running on NVIDIA DGX infrastructure can instantly search across scanned manuals, maintenance logs and parts catalogs. Workers can now ask questions like, "What is the torque specification for the radio chassis mount?" and receive answers directly from the correct page. NVIDIA Riva can also transcribe hands-free voice queries from the shop floor, improving safety and efficiency.

Legal discovery and research

Legal professionals manage massive document sets spanning depositions, filings and prior rulings. RAG systems powered by NVIDIA NeMo and NIM can streamline e-discovery and legal research by summarizing relevant rulings and providing source citations.

Knowledge access for staff and citizens

Personnel across large organizations must navigate extensive procedural and policy documentation. A RAG-enabled chatbot can retrieve up-to-date guidance from verified internal databases, while NVIDIA Riva enables multilingual, voice-enabled and accessible support for diverse audiences—including individuals with visual, auditory or language-based accessibility needs. This improves responsiveness, inclusivity and accessibility for all users, reduces administrative backlogs and maintains strict data protection.

Conclusion

To derive the most value from their enterprise data, federal agencies need generative AI models that are accurate and trustworthy, with an accelerated infrastructure to make deployments fast and easy. As an authorized NVIDIA Partner and value-added reseller in the federal marketplace, CTG Federal can help agencies harness NVIDIA solutions to efficiently deploy RAG-powered LLMs on an enterprise scale.

CTG Federal: An Indispensable Agency Partner for NVIDIA Solutions

CTG Federal, a Cohesive Technology Group company, delivers unrivaled expertise and value to federal agencies looking to optimize their AI and HPC infrastructures with NVIDIA solutions. We are one of only a handful of value-added resellers that carry DGX AI Compute Systems Competency. As a member of the NVIDIA Partner Network, CTG Federal has access to privileged product roadmap information, and product and corporate engineering.

In addition, CTG Federal offers:

- A sales and engineering staff that is fully certified and NVIDIA-accredited on the full portfolio of NVIDIA solutions and products.

- Best practices on the design, architecture and sizing of NVIDIA solutions.

- NVIDIA product updates and NDA (non-disclosure agreement) roadmaps to help agencies future-proof their modernization plans.

- Promotional pricing, discount programs and value-added price breaks.

- Speed, efficiency and accuracy in booking, shipping and tracking customer purchases.

- Dedicated and knowledgeable post-sales and logistical support.

- NVIDIA-certified and accredited engineering staff for pre-sales support and post-sales implementation and migration work.

- U.S. citizen, U.S.-based, DoD TS- and DoE Q-cleared personnel at all levels of sales and engineering.

About CTG Federal

CTG Federal, a Cohesive Technology Group company, is a small business that excels in servicing dozens of federal defense, intelligence, and civilian organizations with best-in-class information technology. Our experienced team of sales and engineering professionals design and deliver IT hardware and software solutions that save time and money for our customers. Headquartered in Virginia, we have dedicated resources in all regions across the continental United States.

Contracts

DUN & Bradstreet: 080932836
UEI: G2D4Q7UKR5P5
CAGE Code: 7ZHE9
NAICS Code(s): 541519

NASA SEWP V
Contract Number: NNG15SD12B
Group: Group B_Small Business

GSA Multiple Award Schedule (MAS)
Contract Number: 47QTCA25D003P

DOE ICPT
Contract Number: ICPT CISCO BOA 4I-30062-0008A

Dell Technologies ICPT Agreement
Contract Number: 4I-31841 Small Business

Horizon ELA
Contract Number: W519TC-25-D-A005

ITES-4H
Contract Number: W519TC25DA013

Smart | Secure | Scalable

(703) 278-3885
contact@ctgfederal.com
1818 Library Street, Suite 500
Reston, VA 20190

www.ctgfederal.com