Solution Overview

NVIDIA

How Can Generative AI Transform the Public Sector?

Enhance citizen services and mission outcomes by streamlining generative AI deployment with foundation models and accelerated infrastructure.

Opportunities of Generative AI Adoption

Whether it's modernizing humanitarian assistance, improving public health and disaster response, powering multilingual chatbots for public servants and citizens, or streamlining reporting workflows, generative AI can boost cost savings and efficiency across mission-focused use cases in the public sector.

Rather than create a new generative AI model, public sector agencies can begin with a foundation model suited for their use case and customize it with proprietary data to make it savvy with their operations, vocabulary, customers, environment, and particular skills.

Retrieval-augmented generation (RAG), an AI approach that leverages internal enterprise data to answer inquiries, can empower government employees and citizens with up-to-date information instantly. By relying on controlled information retrieval from vetted sources and limiting exposure to external data, RAG also helps protect privacy and keeps digital operations secure.

Public sector organizations can take advantage of flexible deployment options, including private cloud and on-premises infrastructure, to safeguard personally identifiable information and sensitive government data.

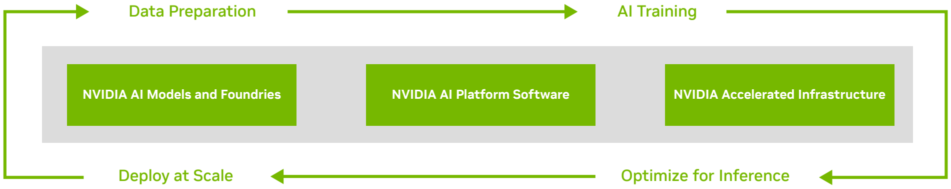
By deploying guardrails specific to the use case, agencies can ensure that generative AI addresses only the desired use case and avoids returning harmful or off-topic information.

As capabilities grow, public sector agencies can fine-tune and scale generative AI solutions by tapping into accelerated infrastructure for AI model training and inference.

With **NVIDIA AI**, enterprises get the complete platform they need to adopt and accelerate generative AI.

Introducing NVIDIA AI for Generative AI

NVIDIA AI is the world's most advanced platform for generative AI. Innovations are built into every layer of the stack—including accelerated computing, essential end-to-end AI software, pretrained models, AI models, and foundries.


Data Preparation
AI Training
NVIDIA AI Models and Foundries
NVIDIA AI Platform Software
NVIDIA Accelerated Infrastructure
Deploy at Scale
Optimize for Inference

Generative AI Challenges

- **Training data:** Large, clean datasets must be curated to train the large language model (LLM) to understand, predict, and respond accurately.
- **Data storage:** Training data must be collected, processed, and managed.
- **Large-scale compute:** The infrastructure required to train an LLM with billions of parameters can require significant compute infrastructure.
- **Training and inference:** Generative AI models need to be chosen and customized for the business, and then deployed for inference on distributed GPU-accelerated infrastructure.
- **AI expertise:** Generative AI requires an understanding of AI, machine learning, and data science principles. Businesses must add new technical expertise to their data science, AI, or application development teams.

NVIDIA AI Platform for Generative AI

AI Models and Foundries

- **NVIDIA NeMo™** lets organizations build, customize, and deploy

AI Models and Foundries

With **NVIDIA pretrained models** and AI foundries, you can build and customize generative AI models for any application, anywhere. 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 generative 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.

AI Platform Software

To get started building generative AI, public sector developers have access to **NVIDIA AI frameworks**, libraries, pretrained models, and tools. They can deploy these resources with NVIDIA AI Enterprise, which offers enterprise-grade support, security, stability, and manageability. It's also compatible with industry-leading infrastructure solutions.

Accelerated Infrastructure

The **NVIDIA DGX platform** incorporates the best of NVIDIA software, infrastructure, and expertise in a modern, 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 in clouds or on premises, and direct access to NVIDIA DGXperts to speed the ROI of AI for every enterprise.

NVIDIA-Certified Systems from the world's leading providers are optimized to run generative AI solutions in on-premises data centers, in clouds, and on workstations.

Consulting and service delivery is available via NVIDIA's worldwide, vibrant partner ecosystem that puts solutions together for organizations in concert with their in-house personnel. NVIDIA partners know the NVIDIA AI stack well and can provide implementation services and support that best leverage accelerated infrastructure anywhere.

Get Enterprise-Grade Support, Stability, Manageability, and Security With NVIDIA AI Enterprise

NVIDIA AI Enterprise is an enterprise-grade, end-to-end software platform that delivers the performance, efficiency, and responsiveness critical to powering the next generation of AI—in the cloud, in the data center, at the network edge, on workstations, and in embedded devices. This includes over 100 frameworks, pretrained models, and open-source tools.

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.

Ready to Get Started?

To learn more about NVIDIA AI, visit: **www.nvidia.com/ai**

To learn about NVIDIA public sector customers and solutions, visit: **www.nvidia.com/public-sector**

generative AI models anywhere. It includes training and inferencing frameworks, a guardrail toolkit, data curation tools, and pretrained models.

- **NVIDIA Nemotron-3 8B** is a family of foundation models for building production-ready generative AI, including applications for multilingual information retrieval.
- **NVIDIA Picasso** runs optimized models to generate image, video, 3D, and 360 HDRi content from text or image prompts.

AI Platform Software

- **NVIDIA AI Enterprise** is an enterprise software platform that accelerates the development and deployment of production-grade generative AI. It includes over 100 frameworks, pretrained models, and open-source tools.

Accelerated Infrastructure

- **NVIDIA DGX™** integrates AI software, purpose-built hardware, and expertise into a comprehensive solution for AI development that spans from the cloud to on-premises data centers.
- **NVIDIA-Certified Systems™** provide the performance, reliability, and scalability to deliver cutting-edge products and services while increasing operational efficiencies.

Key Benefits

- **Time to solution:** Quickly build custom enterprise-grade models with 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.
- **Production-ready:** Create enterprise-grade models that protect privacy, data security, and intellectual property.

Partner Logo

NVIDIA