Red Hat

AI for the public sector

- Automation
- Cost reduction
- Transparency

Contents

Introduction

The public sector at an inflection point

AI has reshaped what citizens expect from private sector service providers—instant responses, predictive recommendations, and always-on channels have become baseline requirements. Public agencies now face the same pressure, but have never been funded or staffed to compete with digital-native, private sector companies; as a result, they are now racing to catch up. In fact, government investments in AI are estimated to reach a 19% compound average growth rate (CAGR) between 2022 and 2027.[1]

These public sector agencies are discovering an uncomfortable paradox: the very AI models that promise to slash manual effort can come with staggering compute and cloud subscription bills. Flexera's 2025 State of the Cloud snapshot found that 33% of organizations now spend more than US$12 million annually on public cloud resources, with 72% naming generative AI (gen AI) services as the primary focus.[2] In state and local governments, which have tight operating budgets, those costs have significant consequences; every extra graphics processing unit (GPU)-hour forces a trade-off with frontline programs.

In light of these challenges, open source AI solutions can give public sector agencies greater flexibility, transparency, and control. This allows teams to build, audit, and adapt models to meet their specific needs. At the heart of open source is community-powered innovation, which allows agencies to avoid vendor-lock-in, reduce costs, and accelerate development while maintaining compliance across security-focused, scalable environments.

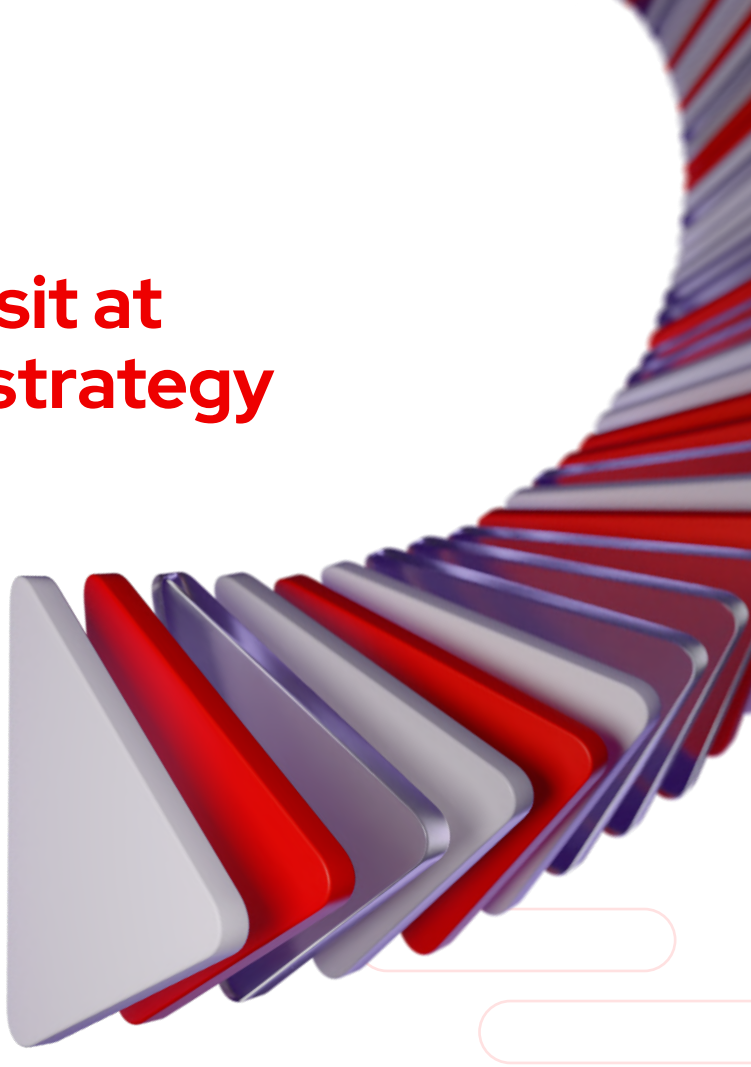
This e-book explores how the public sector can resolve this tension, using AI as an automation engine to mitigate costs while applying open source innovation to keep infrastructure spending in check. The following chapters will:

- Quantify the fiscal opportunity of automating repetitive processes.
- Examine the hidden costs that mount as pilot projects scale.
- Compare large, general-purpose models with task-tuned small language models (SLMs) and cost-saving techniques such as virtual large language model (vLLM) prompt caching.
- Show how open source communities such as Red Hat's new llm-d project are rewriting the economics of inference.
- Map the policy, privacy, and procurement considerations unique to government.
- Highlight Red Hat® technologies and partner solutions that operationalize AI—from security-focused supply chains, to AI-augmented IT automation with Red Hat Ansible® Lightspeed, to federated learning from partners such as Intel.

1 Gartner. "AI in Government Promises Automation and Better Decisions," accessed 1 Jun. 2025.
2 Flexera. "2025 State of the Cloud Report," accessed 12 May 2025.

Chapter 1

Why automation must sit at the center of every AI strategy



The oldest mandate in public administration is to do more with less. Automation is a key tool that aligns this mandate with 21st century workloads. Analysts predict that through 2026, 20% of organizations will use AI to increase the efficiency of their teams by eliminating more than half of current middle management.[3]

Each government function—including licensing, benefits adjudication, and procurement—has grown around paper-era controls designed to maintain transparency and prevent corruption. Every control requires a hand-off, and every hand-off demands salaried staff. Automation replaces those hand-offs with software logic that creates the same audit trail in milliseconds. According to Deloitte's 2024 analysis of more than 19,000 U.S. government tasks, large portions of public sector work—particularly routine, data-heavy activities—are amenable to automation by gen AI tools. The report also found that automation can save between 75% and 95% of the time spent on functions like drafting reports or routing documents.[4]

Automation also alters the cost curve in surprising ways. When a help desk chatbot is introduced, the number of support tickets may drop by as much as 60%, reducing personnel costs.[5] However, cloud costs can increase because each step in a chatbot conversation requires powerful computing to understand the question and generate a response—a process called inference. A midsize state agency that recently automated benefits-eligibility queries saw its monthly AI bill triple, consuming funds earmarked for service center closures.

Historically, agencies attempted piecemeal fixes such as digitizing individual forms or deploying a bot to triage emails, however this resulted in disconnected processes and interruptions when services moved within the agency. The next wave of AI-powered automation must therefore operate end to end, orchestrating document understanding, decision rules, workforce scheduling, and citizen self-service in the same security-focused workflow. Getting there requires a platform to manage, maintain costs, and allow for economies of scale.

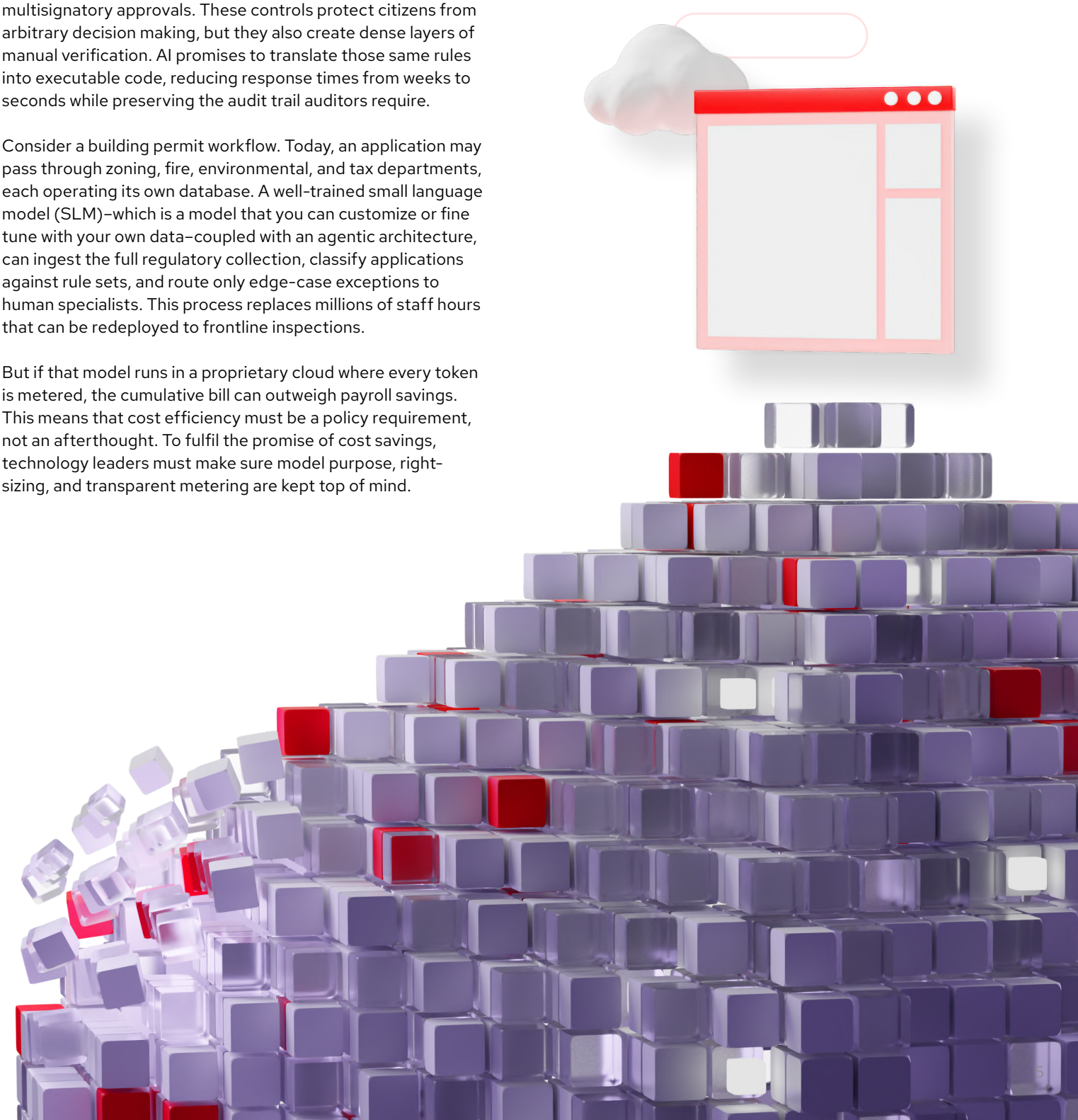
3 Gartner. "Gartner Unveils Top Predictions for IT Organizations and Users in 2025 and Beyond," accessed 22 Oct. 2024.
4 Austin, Tashs, Mariani, Joe, et al. "Generative AI and government work: An in-depth analysis of 19,000 tasks." Deloitte, 24 Apr. 2024.
5 Shark. Alan. "What the Rising Costs of AI Means for Government." StateTech, 17 Jan. 2025.

Policy, process, and the transparency imperative

Every public service is bound by painstakingly documented procedures: mandatory waiting periods, notarized forms, and multisignatory approvals. These controls protect citizens from arbitrary decision making, but they also create dense layers of manual verification. AI promises to translate those same rules into executable code, reducing response times from weeks to seconds while preserving the audit trail auditors require.

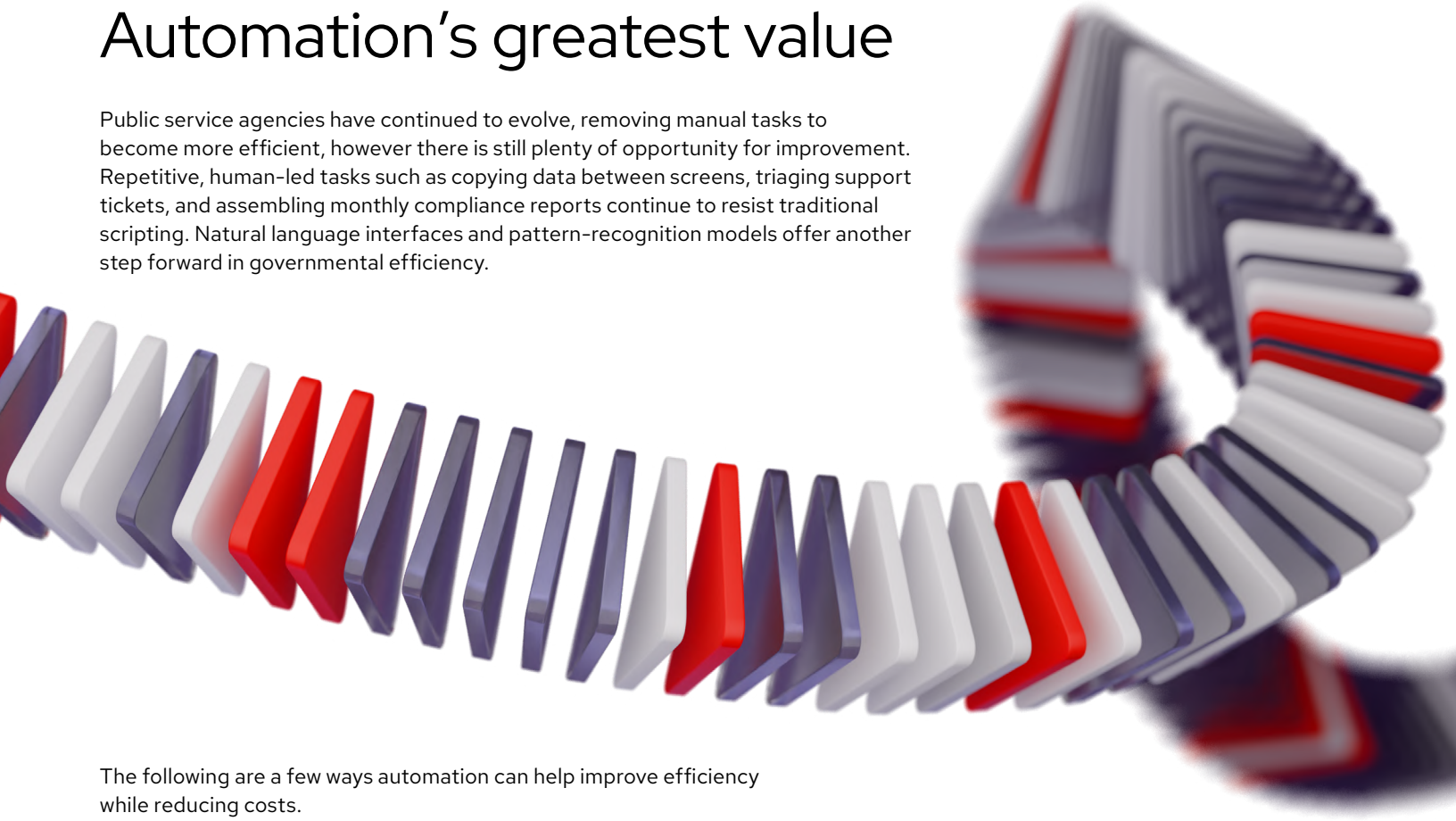
Consider a building permit workflow. Today, an application may pass through zoning, fire, environmental, and tax departments, each operating its own database. A well-trained small language model (SLM)–which is a model that you can customize or fine tune with your own data–coupled with an agentic architecture, can ingest the full regulatory collection, classify applications against rule sets, and route only edge-case exceptions to human specialists. This process replaces millions of staff hours that can be redeployed to frontline inspections.

But if that model runs in a proprietary cloud where every token is metered, the cumulative bill can outweigh payroll savings. This means that cost efficiency must be a policy requirement, not an afterthought. To fulfil the promise of cost savings, technology leaders must make sure model purpose, right-sizing, and transparent metering are kept top of mind.



From rote tasks to predictive insight: Automation's greatest value

Public service agencies have continued to evolve, removing manual tasks to become more efficient, however there is still plenty of opportunity for improvement. Repetitive, human-led tasks such as copying data between screens, triaging support tickets, and assembling monthly compliance reports continue to resist traditional scripting. Natural language interfaces and pattern-recognition models offer another step forward in governmental efficiency.



The following are a few ways automation can help improve efficiency while reducing costs.

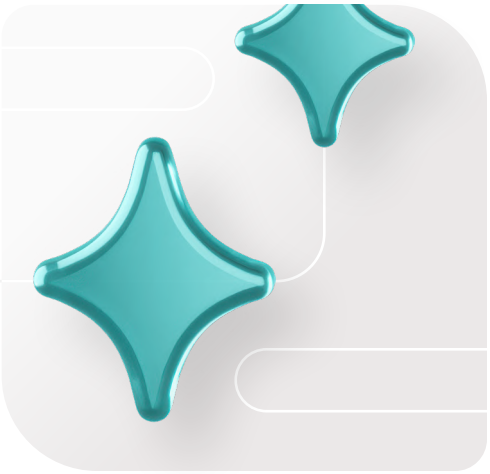
- **Streamline citizen support**
  Automated prompt caching allows for accelerated, repeatable responses to common inquiries such as "Where's my refund?" This can reduce strain on GPUs and improve service delivery speed.
- **Accelerate fair eligibility processing**
  SMLs trained on regulatory text help automate permit and benefit adjudication by identifying incomplete applications and drafting consistent eligibility decisions aligned to agency language.
- **Inform resource allocation**
  AI-powered forecasting models can predict service demand, such as emergency department volumes, allowing administrators to proactively reallocate staff and avoid bottlenecks.
- **Personalize citizen outreach**
  Recommender systems can automatically identify and engage at-risk individuals, such as veterans due for prescription refills, before small issues become larger health concerns.

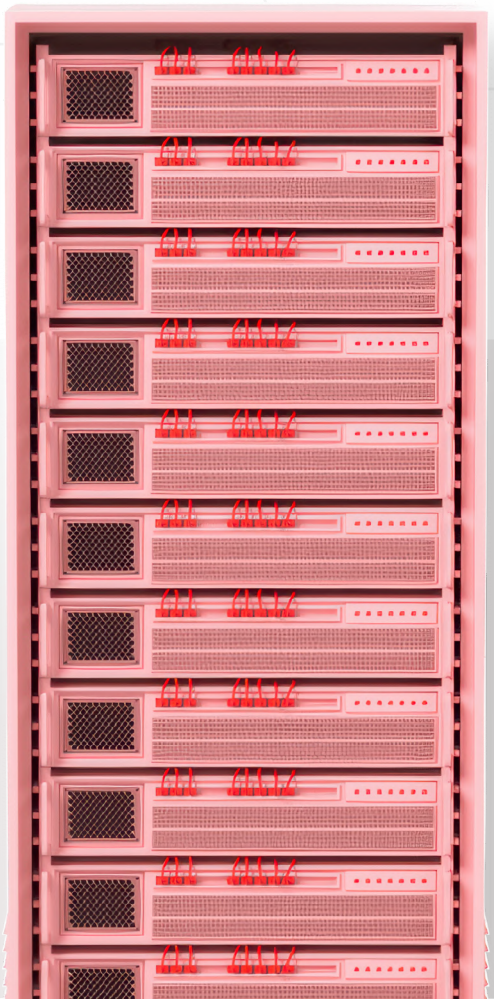
For example, a nationwide chatbot serving 20 million inquiries could process 3 trillion tokens, a figure that increases drastically without optimization using automation. To address this, agencies can serve a distilled SLM via llm-d, which breaks the model into smaller components and runs them across a Kubernetes cluster, keeping use high and costs low.

Chapter 2

Barriers on the road to automation

The promise of AI in the public sector is clear, but so are the roadblocks. From aging systems and disconnected data to rising cloud costs, governments face steep structural and operational challenges on the path to automation. Add mounting compliance obligations and a widening skills gap, and it becomes clear why many agencies are struggling to scale their AI strategies.





The following barriers to automation adoption are among the most persistent—and solvable—today:

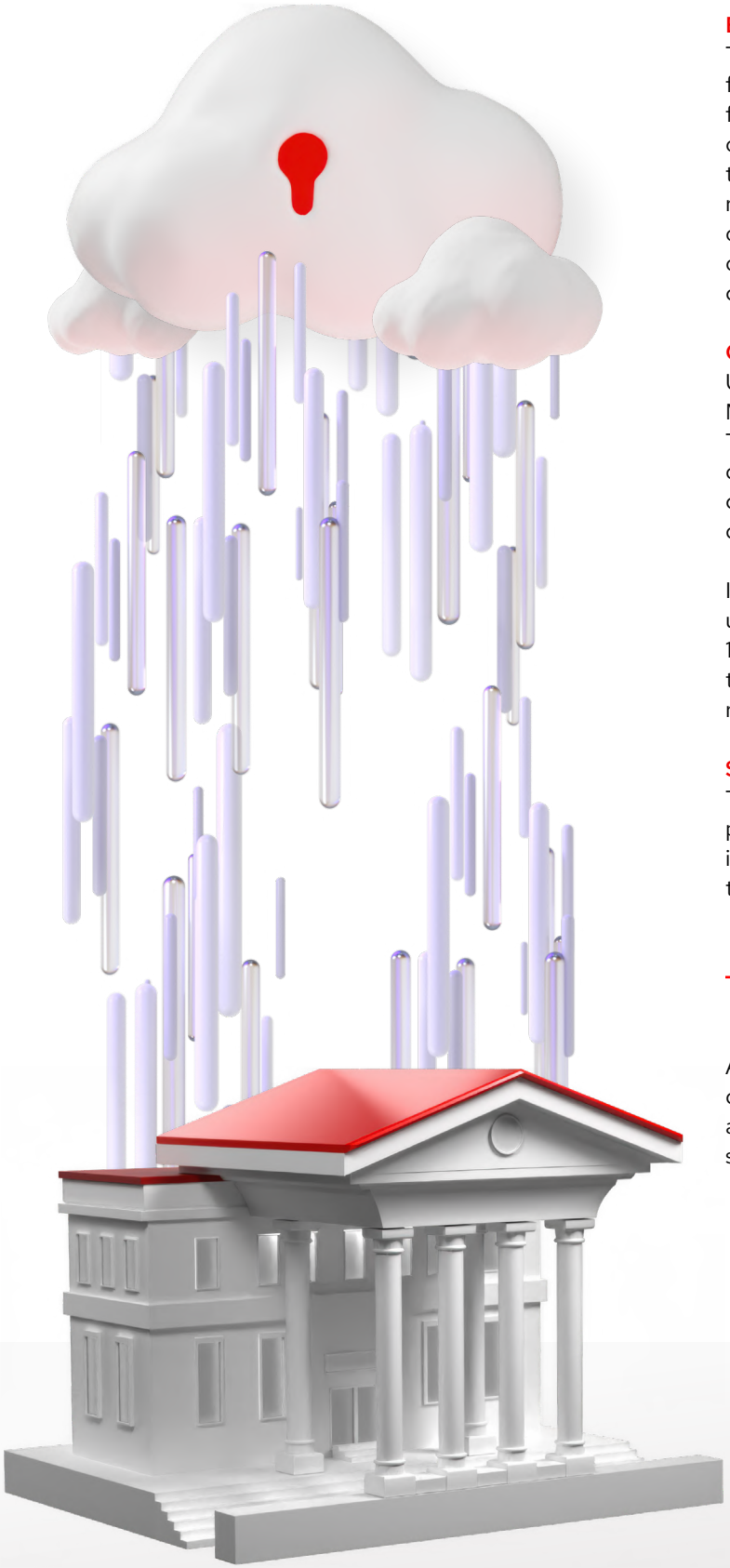
Aging infrastructure

Aging systems and IT infrastructure consume the majority of public sector IT budgets, slowing modernization efforts. The U.S. Government Accountability Office notes that about US$100 billion of the federal government's annual IT budget is allocated to operating and maintaining aging systems, leaving little room for innovation until traditional infrastructure is retired.[6]

Disconnected data

Data that can't be accessed by all parts of the enterprise prevents agencies from deploying automation at scale or training AI on complete, real-time information. Salesforce's 2024 Connectivity Report found that 81% of IT leaders point to disconnected data as the cause of hindered digital transformation efforts, and 62% admit their systems are not configured to use AI at all.[7]

6 "Information Technology: Agencies Need to Plan for Modernizing Critical Decades-Old Legacy Systems" U.S. Government Accountability Office, 17 Jul. 2025.
7 "85% of IT Leaders See AI Boosting Productivity, but Data Integration and Overwhelmed Teams Hinder Success." Salesforce, 23 Jan. 2024.



Escalating cloud spend

The computer demands of gen AI are pushing cloud budgets far beyond what public sector teams originally planned. In fact, enterprise cloud bills have jumped 30% year-over-year on average, and IT and finance leaders attribute the spike to gen AI workloads.[8] Flexera's 2025 State of the Cloud report underscores the strain, reporting that nearly a 3rd of organizations now spend more than US$12 million per year on public cloud resources to power AI, pushing overall cloud budgets 17% beyond plan.[9]

Compliance and ethical risk

U.S. Government agencies must meet the U.S. Office of Management and Budget (OMB) Memorandum M-24-10. This orders federal departments to inventory AI systems, document training data, and model decisions, and implement continuous bias testing with auditable records before any deployments that could affect rights or safety.[10]

In the U.K., the EU AI Act now classifies most public sector uses as high risk, and therefore requires agencies to maintain 10-year technical documentation, perform systematic bias tests, and preserve tamper-proof audit logs for every release cycle.[11]

Skills gap

The U.S. Department of Energy's 2025 reduction-in-force plan eliminates more than 3,500 positions and labels 43% of its roles "non-essential." This illustrates the federal imperative to automate routine work as budgets tighten.

At the root of each challenge is money: each barrier either drains scarce cash or hides excess spend. The flexibility and adaptability of open source tooling, backed by enterprise support, offers the clearest path forward.

8 Ashare, Matt. "AI adoption drives 'unmanageable' spike in cloud costs." CIO Dive, 7 Oct. 2024.
9 "New Flexera Report Finds that 84% of Organizations Struggle to Manage Cloud Spend." Flexera, 19 Mar. 2025.
10 "Advancing the Responsible Acquisition of Artificial Intelligence in Government."Office of Management and Budget, 24 Sept. 2024.
11 "The AI Act Explorer." EU Artificial Intelligence Act, accessed 10 Jun. 2025.

Addressing the AI skills gap

While AI promises to reduce headcount, governments are planning to reskill rather than lay off employees. The challenge is attracting top talent under public sector salary caps. As AI becomes more embedded in public services, the demand for skilled talent in AI, data science, and machine learning (ML) continues to outpace supply. Open source tools can help by lowering barriers to entry and supporting skills development within government teams, by offering:

Accessible AI tools

Many open source AI frameworks are free to use, easy to learn, and widely supported by active communities. They can help government teams to experiment, build, and deploy models without the cost or complexity of proprietary licensing.

Cross-functional collaboration

Encouraging knowledge sharing between IT staff, data scientists, and subject matter experts fosters a broader skill set and supports AI projects that are more aligned to mission outcomes.

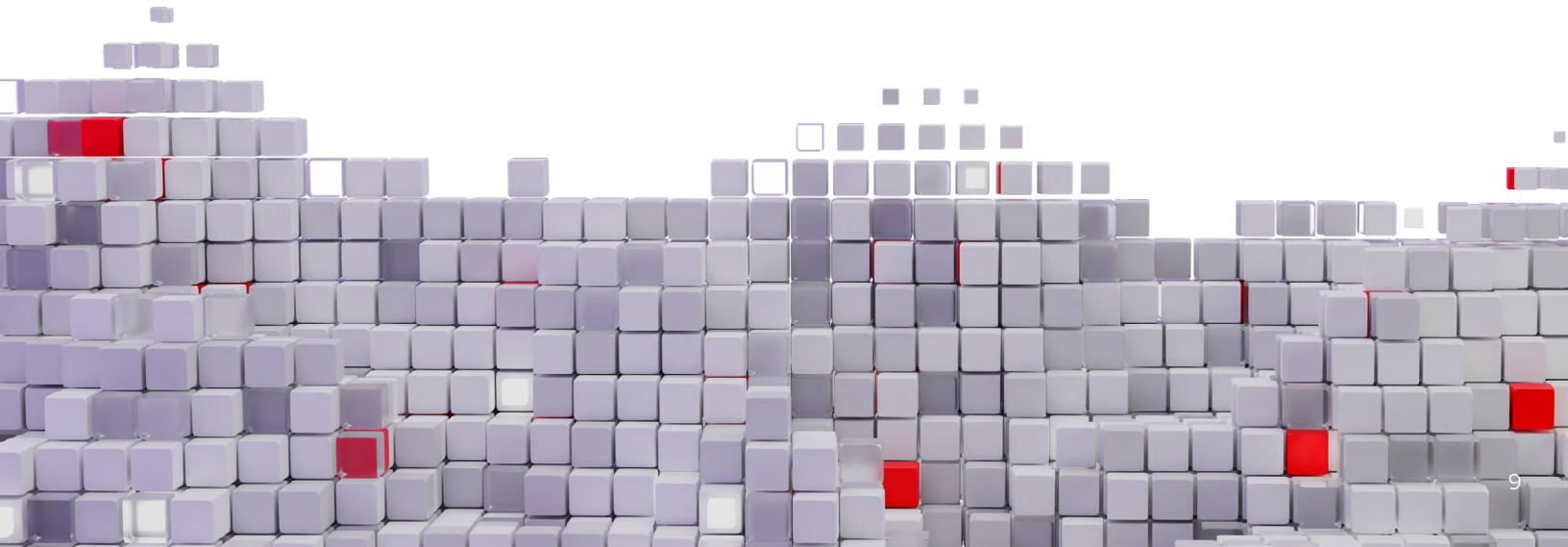
Access to external expertise

Working with open source communities and technology collaborators provides government teams with access to specialized knowledge that can accelerate AI adoption.

Unified AI Integration

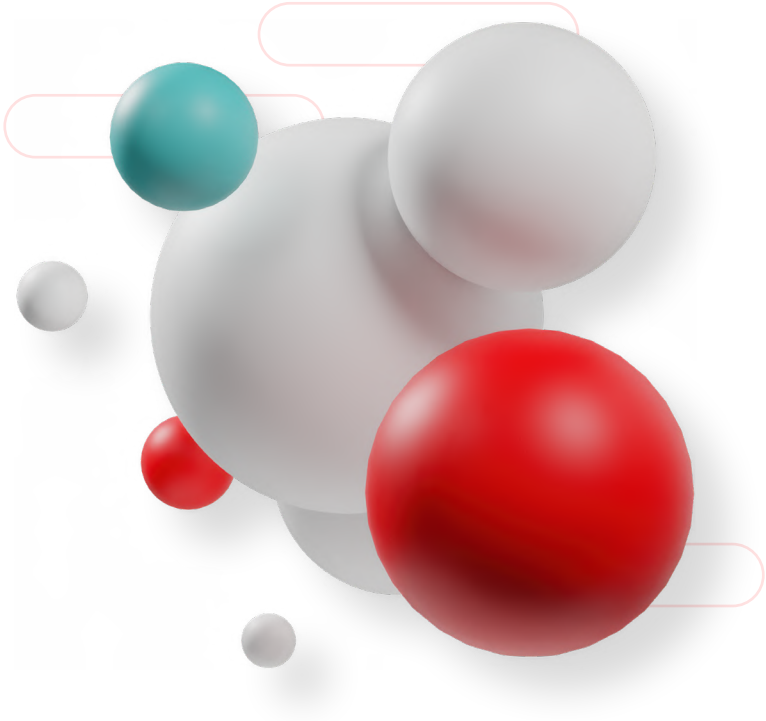
By standardizing how AI agents invoke external tools and data through a single, REST-style schema, the open-source Model Context Protocol (MCP) removes the need for bespoke integration code, helping ordinary developers—not scarce AI specialists—to plug models into existing systems and promote cross-agency reuse.

By embracing open source solutions, public sector organizations can adopt AI on their own terms, with a security-focused and open approach to innovation, at a pace that aligns with their mission. This approach not only helps protect citizen data, but also makes sure that the benefits of AI innovation are delivered responsibly and equitably.



Chapter 3

Open source as a cost-control and trust mechanism



An open source approach offers a natural fit for public sector priorities, lowering costs, increasing transparency, and accelerating innovation without vendor lock-in or reliance on proprietary tools. The following benefits show how open ecosystems align with government value and operational needs:

Innovation and choice

The fast-growing llm-d open-source community augments the vLLM codebase–a tool that helps serve large language models (LLMs) more efficiently. New features, such as ARM edge devices, tensor processing units (TPUs), and advanced inference techniques, are added quickly, without waiting for long procurement processes.

> vLLM allows for faster, more cost-effective AI model serving by maximizing GPU efficiency and reducing latency. This is critical for delivering real-time services at scale. Integrated with the open Llama Stack and Model Context Protocol (MCP), it supports more security-focused, portable, and interoperable gen AI workloads across hybrid cloud environments.

Vendor freedom

Kubernetes lets agencies run workloads consistently across on-premise systems and multiple cloud providers–avoiding vendor lock-in and unpredictable pricing models. Red Hat expands that flexibility by supporting any model, hardware, or cloud, so teams can choose the most cost-effective setup without extra coding or security trade-offs.

Transparency

Agencies need to know exactly what's in their AI models. Using tools such as Sigstore and software package data exchange (SPDX) software bills of materials (SBOMs), every component is tracked, verified, and logged, making it easier to enforce policies and spot tampering without relying on closed, third-party tools.

Democratized skills

Open, community-maintained frameworks such as TensorFlow, PyTorch, MCP, Llama Stack, and KNIME—paired with community support and peer reviews—make it easier and more cost effective for agency staff to build practical AI skills. Instead of costly training, teams can learn more quickly from free, trusted online resources, just like the 82% of programmers who already do.[12]

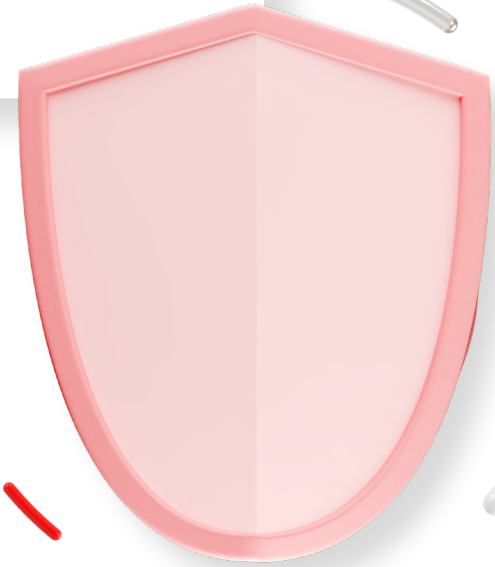
Independent analysts repeatedly find double-digit total cost of ownership (TCO) advantages for open source solutions—savings that accrue more quickly in government because operational costs are far greater than licence fees.

12 "2024 Developer Survey." Stack Overflow, accessed 10 Jun. 2025.

Safeguarding privacy and data sovereignty

While cost is a key consideration, public service agencies also face legal and technical barriers when trying to use data across regions or departments. Federated learning (FL) helps solve this by allowing models to be trained without moving the data. This makes sure that sensitive information stays where it is, addressing the jurisdictional and data sovereignty barriers that keep information trapped in disconnected locations.

The Linux Foundation's OpenFL project—created by Intel and now community-governed—gives data-science teams a Python SDK, secure gRPC channels, and attestation workflows for production-grade FL.



Agencies can work together on AI projects without compromising privacy or compliance by using a trusted combination of tools:

- **Differential privacy,** which adds mathematical noise to data, protecting individual identities while still allowing useful insights.
- **Homomorphic encryption,** to allow teams to analyze encrypted data without ever viewing the raw information.
- **Confidential compute** (also known as secure enclaves), for locking down the processing environment so even cloud providers can't access what's happening inside.

Collaborate without paying egress or risking privacy: Federated Learning OpenFL

The Linux Foundation's OpenFL project—created by Intel and now community governed—gives data science teams the tools they need, such as Python SDK and secure communication channels to build security focused, scalable FL solutions.

- **Transportation.** U.S. State Departments of Transportation (DOTs) used FL to predict sensor failures on bridges, improving accuracy together without exposing sensitive infrastructure data.
- **Public health.** Penn Medicine and Intel Labs first proved in 2018 that federated learning could match 99% of centralized-model accuracy, and the effort has since grown to more than double the 2020 site count across 6 continents while curating the largest glioblastoma dataset to date (5 TB, 6,314 patients). The latest consensus model delivers up to 33% better tumor-detection accuracy, 4.5x lower latency, and 2.3x less memory—compact enough for real-time use on clinic-edge devices.[13]

Summary:
Agencies can train locally using existing hardware. Because only model updates, not raw data, are shared, cloud costs are reduced and interagency data transfer fees can be avoided.

Protect privacy without sacrificing insights: differential privacy

Differential privacy adds mathematical noise to data before it's shared, protecting individual identities without changing the big-picture results. The U.S. Census Bureau used this approach in 2022 to meet strict confidentiality requirements while publishing nationwide data.[14]

- **Real-world use.** Agencies can apply the same Laplace-noise libraries (e.g., SmartNoise in OpenFL) to publish opioid-overdose heat maps or unemployment microdata without risking reidentification.

Summary:
It works during data preparation runs and doesn't slow down AI performance, so privacy is built in without additional costs at runtime.

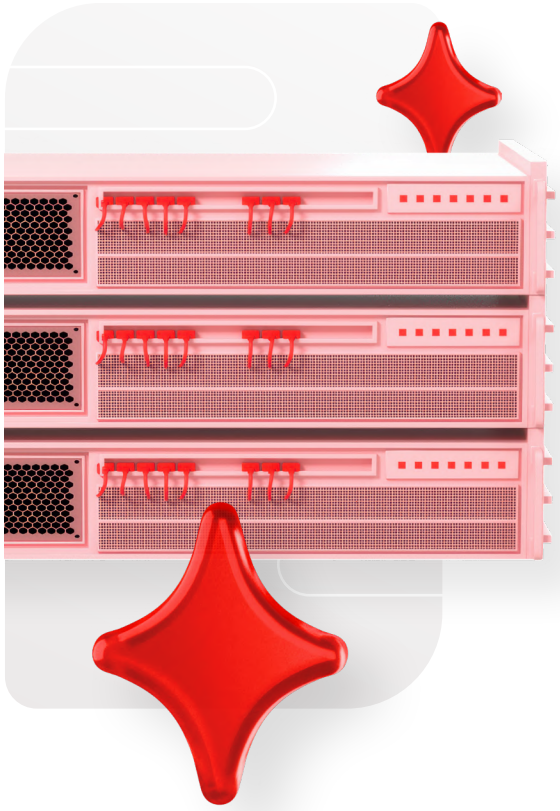
13 "Federated Learning: Protecting Data at the Source." Intel Labs, accessed 10 Jun. 2025.
14 "Understanding Differential Privacy." United Census Census Bureau, accessed 10 Jun. 2025.

Run AI on encrypted data: homomorphic encryption

Fully homomorphic encryption (HE) allows agencies to perform calculations on encrypted data–such as running models on encrypted tax records or legal cases–without ever seeing the data. Toolkits such as PALISADE integrate with OpenFL workflows, and Intel's upcoming Confidential AI hardware make this more practical.

- **Real-world use.** In Europe, courts are using HE to predict recidivism risk locally, feeding insights into a federated dashboard without exposing sensitive criminal records.

Summary:
While HE can increase CPU use, it avoids the costs and risk of managing complex data sharing agreements.



A security-focused AI runtime: confidential AI and secure enclaves

Red Hat OpenShift AI supports Intel Trust Domain Extensions (TDX) and AMD SEV that lock down AI models during training or inference, even in public cloud environments. These enclaves make sure no parties, not even the cloud provider, can see what's running inside.

- **Real-world use.** Hospitals sharing Health Insurance Portability and Accountability Act (HIPAA)-protected radiology data use enclave-based training with federated learning improving brain tumor detection by 33%–while still meeting strict privacy rules.[15]

Summary:
Agencies can safely use lower-cost shared cloud resources instead of paying more for isolated VMs, cutting infrastructure costs significantly.

By combining open-source tools such as OpenFL with Red Hat's trusted AI platform, public service agencies can work together on powerful AI solutions without moving sensitive data, compromising security, or straining their budgets.

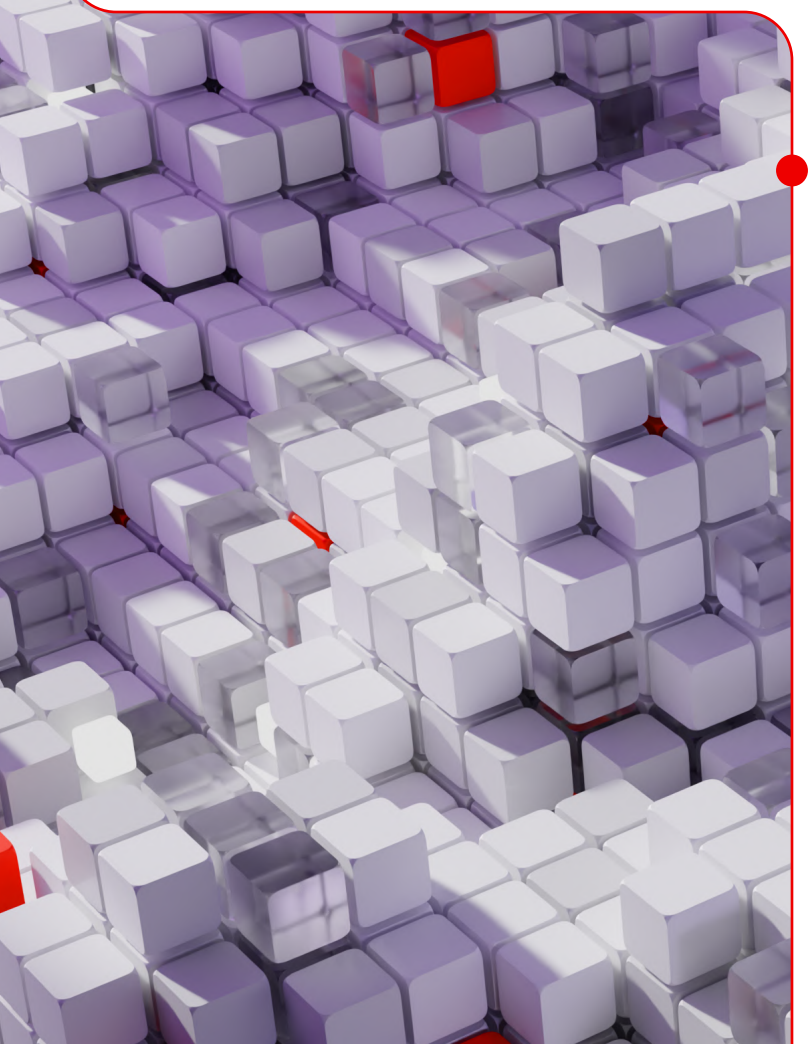
15 Intel press release. "Intel and Penn Medicine Announce Results of Largest Medical Federated Learning Study." accessed 10 July 2025.

Chapter 4

A trusted platform for cost-conscious automation

Red Hat AI is a flexible, open platform that helps organizations accelerate AI innovation while controlling cost and complexity. Red Hat AI, which comprises Red Hat Enterprise Linux® AI, Red Hat OpenShift® AI, and Red Hat AI Inference Server supports the full lifecycle of both predictive and gen AI models—from development to deployment to monitoring—across on-premise, cloud, and edge environments.

Powered by open source technologies and a trusted partner ecosystem, Red Hat AI helps public sector agencies:



Gain a more consistent way to build and run LLMs

Red Hat Enterprise Linux AI provides a complete, open source foundation model platform that supports model development, testing, and deployment across hybrid cloud environments. In addition to the Granite LLM family, it includes InstructLab tools that allow customers to align models with domain-specific data through a collaborative, community-focused approach.

Red Hat Enterprise Linux allows AI workloads to run on security-focused, policy-compliant hosts from the datacenter to the tactical edge. A curated model registry ships with pre-vetted SLMs and LLMs complete with provenance, signatures, and SBOMs that are continuously verified, giving agencies trusted building blocks. Teams can then use the open source InstructLab workflow to align or fine tune those models on domain data without needing hyperscale GPU fleets or external APIs.

Accelerate AI innovation and streamline deployment across environments

Red Hat OpenShift AI provides an AI platform for managing the lifecycle of predictive and gen AI models at scale across hybrid cloud environments.

Public sector agencies gain an end-to-end, container-native workbench where data scientists can author and fine tune models using familiar notebooks or integrated development environments (IDEs) while IT teams provision GPU resources on demand. Built-in MLOps pipelines automate everything from data ingestion and feature engineering to govern continuous integration and continuous delivery (CI/CD) deployment, so each model travels through versioned stages with auditable approvals. During operation, event-driven data pipelines send new data from on-premise or cloud sources to AI services running on the same platform. This setup keeps predictions fast and up to date, without the need for complex integration.

Serve AI models more efficiently and at lower costs

Red Hat AI Inference Server is designed to speed up AI model performance while lowering costs across hybrid cloud environments.

It combines vLLM's advanced memory-sharing and dynamic batching with Neural Magic's model compression technology to help each GPU do more with less–without sacrificing accuracy. By trimming unnecessary model data and reusing repeated prompts, it reduces memory usage and can provide up to 3 times lower cost per token compared to traditional serving methods. The solution runs as a container on any Kubernetes environment, automatically scaling across different CPUs and GPUs, while tracking cost metrics in real time to help agencies stay on budget.

Modernize, reduce operational costs, and close skills gaps

Red Hat Lightspeed brings gen AI guidance directly into Ansible Automation Platform, Red Hat Enterprise Linux system roles, and Red Hat OpenShift GitOps pipelines, translating plain-language commands into production-ready Ansible playbooks, system configurations, and cluster resources in seconds. By simplifying complex workflows and eliminating manual scripting, Lightspeed accelerates time to value and lowers costs.

The assistant uses thousands of proven automation patterns from community-validated patterns to help teams work smarter. It can suggest the right tasks based on context, follow best practices automatically, and fill in system-specific variables, saving time on manual scripting, reducing errors, and helping teams with different skill levels to stay consistent across on-premise, cloud, and edge environments.

The result is a unified, AI-accelerated infrastructure automation that lets hybrid cloud teams provision, patch, and scale environments more quickly while preserving the same open, auditable workflows that make Red Hat tooling trusted in the public sector.

Build and deploy security-focused, compliant applications in less time

Red Hat Trusted Application Pipeline verifies pipeline compliance is met by using an automated chain of trust to increase the security posture in build systems.

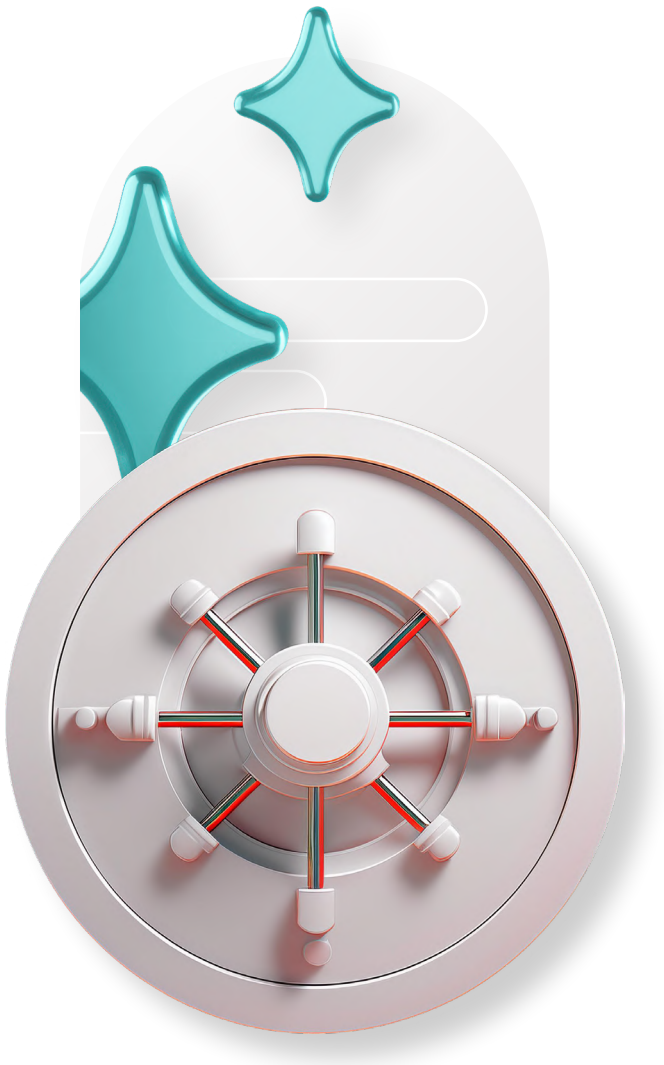
By embedding DevSecOps controls—including SLSA-aligned build policies, SBOM generation, signature attestation, and CVE scanning—directly into each CI/CD run, Red Hat Trusted Application Pipeline creates a hardened, repeatable path from source to container.

For AI projects, the same guardrails maintain security posture and compliance. Model components such as fine-tuned weights, training data, and inference images are automatically signed, verified, and only moved forward if policy checks are passed. Running natively on Red Hat OpenShift AI, Red Hat Trusted Application Pipeline lets data science teams go from development to security-focused production with a single code merge, accelerating results and built-in audit trails without relying on external tools.

Build cost-effective, agency-owned AI services

Building on the automation and security foundation already in place, public sector teams can now create their own gen AI services, tailored to their mission needs and run on infrastructure they control.

By combining Red Hat Inference Server with InstructLab's community workflow for low-cost alignment and fine-tuning of base models, agencies can fine tune models using existing hardware, without the need for expensive proprietary tools. When paired with lightweight agentic frameworks such as Agent Squad, Vantiq, or Coagent, these models can handle complex tasks such as answering policy questions or routing documents–automatically and efficiently. Because inference happens on premise or scales only when needed, agencies get faster results, more control, and lower costs to public cloud APIs.

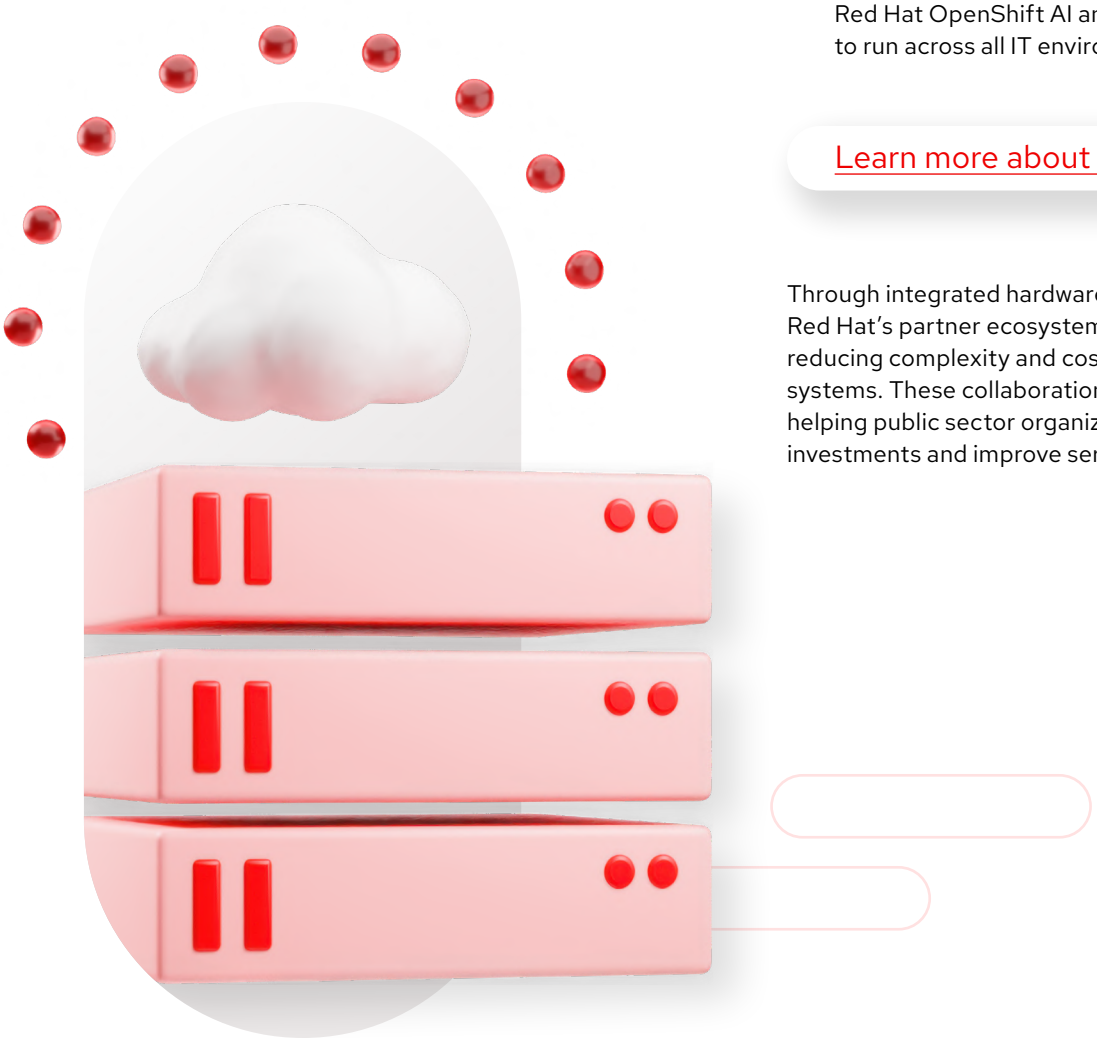


Ecosystem advantage: Intel and beyond

Red Hat's partner ecosystem brings together industry leaders to provide tested, interoperable AI solutions that accelerate innovation and reduce risk. Agencies benefit from shared standards, validated integrations, and the flexibility to choose the best tools for their needs.

For example, Intel hardware validated on Red Hat OpenShift AI gives agencies flexible deployment options across datacenter and edge environments. Shared cost models and coordinated vulnerability disclosures spread research and development and security overhead across thousands of deployments–bringing innovative capabilities to the public actor at lower costs.

Red Hat's open source platforms and Intel's AI tools help public sector organizations build, deploy, and scale AI models efficiently while maintaining compliance with strict data protection regulations across a myriad of environments.



- **In the datacenter.** Intel Gaudi AI accelerators, Intel Xeon processors with Intel Advanced Matrix Extensions (AMX), and Intel Infrastructure Processing Units (IPU) cater to a diverse set of AI use cases, including gen AI training, fine tuning, retrieval augmented generation (RAG), and inferencing, as well as confidential AI to protect data in use with Intel Trust Domain Extensions (TDX).
- **At the edge.** Intel supports local execution of LLMs on platforms based on Intel Core Ultra, Intel Xeon processors, and Intel Arc GPUs. Red Hat is collaborating with Intel to certify its hardware solutions on Red Hat OpenShift AI to create interoperability and comprehensive AI capabilities.
- **Across your IT environments.** Intel brings together a rich set of scalable open source and commercial software that come ready to use and can be integrated with Red Hat OpenShift AI and validated for interoperability to run across all IT environments.

Learn more about Intel tools

Through integrated hardware and software solutions, Red Hat's partner ecosystem accelerates AI deployment, reducing complexity and costs associated with proprietary systems. These collaborations offer comprehensive support, helping public sector organizations to maximize their AI investments and improve service delivery.

Learn more

Automate strategically, spend wisely, serve better

Government investment in AI is expected to accelerate more quickly than any other sector, while hyper-automation promises double-digit cost relief.[1] Taking advantage of the full value of AI depends on employing an automation-first approach, addressing and planning for costs, and adopting open source and a security-focused platform.

Take the 1st step in advancing AI for the public sector

1. **Pilot with purpose**
   Automate a backlog-plagued process in 90 days.
2. **Measure relentlessly**
   Expose GPU hours and response times alongside labor baselines.
3. **Scale what works**
   Use Red Hat OpenShift AI pipelines to roll successful models across departments and into production.

Ready to move from potential to production? Speak to a Red Hatter about how to advance your AI strategy, get hands-on experience with a 60-day trial subscription for Red Hat OpenShift AI, and start capturing the only 2 resources no budget can refund: **time and public trust.**