WHY OPEN-SOURCE AI MODELS OFFER A SMARTER FUTURE FOR AGENCIES

Federal agencies are discovering that smaller, open-source AI models trained on agency data deliver greater transparency and more reliable results at a fraction of the cost of large language models.

A Scoop News Group Report

Federal agencies are at a critical juncture. Recent White House initiatives aimed at accelerating the adoption of artificial intelligence (AI), combined with the explosive proliferation of AI options, have federal officials scrambling to assess where and how to capitalize on what many regard as the most transformative technology in decades.

Why it matters: AI promises to revolutionize everything from citizen services and regulatory analysis to national defense. Yet for many government CIOs, CISOs, and program leaders, the path to AI adoption is fraught with enormous costs and deep-seated security risks at a time when budgeting and staffing assumptions are under intense pressure and uncertainty.

Unprepared for the cost: AI implementation experts warn that most organizations still underestimate the expenses involved with AI. At issue: Most organizations still focus primarily on the AI solutions without fully accounting for infrastructure upgrades, data improvements, integration and other navigational barriers associated with implementing AI.

According to various estimates:

- **Data preparation costs** typically represent 20-30% of overall AI project costs and often run higher for agencies aggregating data from siloed, legacy systems.
- **Regulatory compliance requirements** add to AI's cost burden. On average, financial services firms incur 25-40% higher AI implementation costs than less-regulated industries. For healthcare organizations, it's more like 30-50%. Those premiums are indicative of what federal agencies are likely to encounter.
- **Infrastructure upgrades and integration** can add another 15-25% to initial implementation costs, and substantially more for agencies still transitioning to the cloud.

The emerging paradox: Adding confusion to agency assessments is the fact that the AI models generating the most attention are often the least suited for the government's highly regulated work:

- **The prevailing assumption:** Since ChatGPT and the power of large language models (LLMs) broke into public view in November 2022, the common wisdom is that "bigger is always better." However, big-name LLMs are built primarily on massive infusions of internet-based content that are helpful for general business purposes, but fall short of mission needs.

- **Black box dilemma:** Even when agencies can run these AI models behind their own firewalls, using their own data, questions remain about the weighting of algorithms operating in opaque "black boxes." The lack of transparency and control continues to breed uncertainty about the reliability and efficacy of AI-generated outcomes.

FROM BLACK BOX TO GLASS BOX: THE OPEN-SOURCE ADVANTAGE

What's in your AI? "If someone shows up on your doorstep with an AI black box, and you don't know what data went into creating that or how it was valued… then you can run into some really interesting problems," warns **Adam Clater**, Chief Architect in Red Hat's CTO organization.

- **Built-in bias:** Some AI models developed in other countries are legally required to reflect party doctrine—illustrating how hidden biases can be embedded in specific systems. Even U.S.-based platforms, however, must continually refine their algorithms to reduce built-in biases that often remain invisible to enterprise users.

- **Sovereign AI:** Around the world, governments are pursuing sovereign AI strategies to ensure that the development and deployment of AI align with national interests, laws and values. While these efforts aim to strengthen security and trust in AI systems by emphasizing local data, infrastructure and oversight, they also highlight how differing regulatory priorities and cultural norms can shape how models perform and what data they prioritize. This creates new considerations for organizations operating across borders.

The open-source alternative: As federal leaders consider how to harness AI's power without losing control of their data, risking security, or exceeding their budgets, one growing solution focuses not on larger models but on smarter, more targeted open-source

AI models, according to a growing chorus of technology experts.

- **Speedier innovation:** For the past two decades, open-source communities have become the place where innovation happens at speed," says Clater. This dynamic has been responsible for foundational technologies like Linux and Kubernetes. Today, it is fueling a revolution in AI.

- **Powering AI:** "The reality is that all of the biggest and best implementations of AI right now are being built on open-source technologies, but also the models themselves are being open-sourced," Clater adds.

- **Smaller vs. bigger:** "Well-designed and well-built small models can be just as good at solving problems as big models or good enough. If I can get to 98 or 99% (performance) but I use only 20% of the GPU hardware, that's really valuable," he says.

Open source AI, however, cannot be viewed the same way as open source software. As a Red Hat blog on open source AI stated:

- **"Unlike software,** AI models principally consist of model weights, which are numerical parameters that determine how a model processes inputs, as well as the connections it makes between various data points."

- **"The majority of improvements** and enhancements to AI models now taking place in the community do not involve access to or manipulation of the original training data. Rather, they are the result of modifications to model weights or a process of fine-tuning, which can also serve to adjust model performance."

- **Freedom to improve models** to meet agency needs "requires that the weights be released with all the permissions users receive under open-source licenses."

Power of transparency: Clater adds, "By using open-source models, you're clearly eliminating many of the risks associated with black box AI. For agencies handling everything from classified intelligence and taxpayer information to sensitive healthcare records, losing that control isn't an option." Transparency also fosters greater trust. Unlike proprietary black boxes, open-source AI allows agencies to genuinely inspect, understand, and trust the tools they are deploying for the first time.

FEDERAL ACTION PLAN: THE PUSH FOR TRANSPARENT, MODIFIABLE MODELS

Accelerating AI adoption: The need for federal agencies to "adopt a forward-leaning and pro-innovation approach" to AI took on new urgency last April with a White House memorandum aimed at fostering "innovation, governance and public trust." A related memorandum instructed agencies to:

- **Protect American privacy:** Agencies must now ensure the AI systems they acquire comply with existing privacy and IP laws, and prevent vendors from processing such data for training, fine-tuning, or developing AI systems.

The biggest and best implementations of AI right now are being built on open-source technologies...If I can get to 98 or 99% (performance) but I use only 20% of the GPU hardware, that's really valuable."

— Adam Clater, Chief Architect, CTO Organization, Red Hat

- **Ensure cost-effective procurement:** Agencies must also ensure procurement contracts protect against vendor lock-in, preserve data and model portability, and leverage performance-based contracting provisions.

- **Assess AI lifecycle risks:** Among other requirements, agencies must be able to monitor and evaluate the performance, risks, and effectiveness of an AI system or service. Contracts also must comply with the minimum federal risk management practices for high-impact AI use cases.

Emphasis on open systems: The plan tasks government agencies with developing customized solutions that avoid closed platforms by promoting transparent, adaptable, and open-weight AI models. This is especially important for:

- **Handling sensitive data** that cannot be sent to commercial black-box models.

- **Encouraging interoperability,** fostering greater trust and faster adoption.

WHY A DICTIONARY CAN WORK BETTER THAN AN ENCYCLOPEDIA

The technical advantage of smaller models: The open-source movement is challenging

the "bigger-is-better" myth by championing a new class of smaller, specialized AI models, according to Clater. These models prove that precision and efficiency can deliver superior results for specific, mission-oriented tasks.

Clater offers a powerful analogy: "If I said I want you to memorize Webster's Dictionary, or I want you to memorize the World Book Encyclopedia, you'd say, well, 'I'd rather just do the dictionary because it will be easier to take in all that information.' You won't have the same depth of knowledge, but you'll be able to access data much faster. And tests have shown that smaller models can give you a high percentile completeness level without memorizing the entirety of the encyclopedia."

Proof in practice: Smaller, specialized models are proving highly effective in highly regulated environments like financial services and healthcare, where specific requirements dictate outcomes and where a wealth of policy documents and procedural knowledge is readily accessible.

Federal agencies, for instance, are discovering that smaller, more fine-tuned AI models, trained on decades of internal case files and regulatory text, can provide more reliable results and detect sophisticated fraud with far greater accuracy than more mainstream AI models, according to Clater.

In another example, Red Hat partnered with Guidehouse and researchers from Rush University to launch an initiative using data-driven innovation to help prevent veteran suicides. Red Hat contributed its open, scalable AI/ML infrastructure—leveraging Red Hat OpenShift, OpenShift Data Science, OpenShift Streams for Kafka and API management tools—to build a modular, continuous learning system capable of securely integrating veterans' electronic health records and operationalizing predictive models

Models trained on the corpus of the entire internet, in contrast, have little or no innate understanding of the domain-specific knowledge of these agencies, Clater explains.

STRATEGIC DEPLOYMENT: BRINGING AI TO WHERE THE DATA HAPPENS

Flipping the paradigm: This new, more purpose-driven approach fundamentally flips the old model. Instead of taking sensitive agency data to a massive, third-party AI model in the cloud, agencies can now bring smaller, more efficient AI models to their data, wherever it resides.

- **Ensuring data sovereignty:** "Bringing your AI implementation to your data in your data center makes a lot of sense," Clater states. "This ensures data sovereignty, security and control."

- **Bringing AI to where the data is:** But the true transformation happens when AI moves beyond the data center to the tactical edge, he explains. "That means we can take our AI to where our data happens: In the field, in hospitals, in the middle of a fire, at a

flood. That's very different from taking our AI to where our data is at rest. Being able to take our AI directly to where the mission is happening in real time is going to bring tremendous value."

- **Leveraging AI consistently:** There's another advantage, says Clater. "You can use that same AI model in the cloud, your data center or at the edge. You begin having consistency in the technologies you're using wherever you need to use that data, instead of only being able to use a model in one place

Practical applications: This is not a futuristic concept; it's an active area of development, says Clater. He paints two scenarios where AI models in the field can power decision-making in real time:

- **Imagine incident commanders** responding to fast-moving wildfires: They want data on soil conditions and the density of forest growth; they want to know where the fire is and where it's likely to go; they need to understand the weather conditions around them and the available resources. "An AI model running at the edge can synthesize this information to make life-saving decisions."

- **Defense and military leaders** similarly can expand their ability to capture emergent changes on the battlefield, where field intelligence from drones can accelerate tactical decision-making. "These changes are going to inform the war-fighter capability of the future, and it's all going to be built on some sort of AI capability."

ENGINEERING THE OPEN AI FUTURE FOR GOVERNMENT

The open road ahead: The future of AI is open source in large part because it is an approach that is accessible to more users, unlocks value faster, and ultimately is more powerful. Among other reasons:

- **Small models accelerate adoption:** They are more efficient to train and deploy, and offer significant advantages in customization and adaptability. IBM's Granite family of open source-licensed models, for example, can run anywhere from a laptop to standard GPU servers.

- **Small models are easier to optimize:** Leveraging techniques like sparsification (trimming the number of active parameters or weights in a model) and quantization (reducing the precision of weighting) results in lower costs, faster inference and the ability to run AI workloads on a wider range of hardware.

- **Small models are easier to train:** Small, open-source AI models offer greater flexibility to train AI using relevant datasets where they occur, on readily available CPUs, without sacrificing performance.

The power of Neural Magic: Red Hat's recent acquisition of Neural Magic enables agencies to align smaller, optimized AI models—including open-source-licensed ones—with their data, wherever it resides across the hybrid cloud, according to Hicks. Recognized as a leader in optimized LLM implementation, it enables IT organizations to run a virtual LLM inference server on less capable hardware, achieve faster model performance, and build an AI stack based on transparent and supported technologies.

Optimizing GenAI: Neural Magic supports Red Hat's vision to expand its AI offerings within the open source community, according to Clater, with capabilities like:

- **InstructLab** – an open-source AI community project created by Red Hat and IBM that allows anyone to help shape the future of generative AI by collaboratively improving open-source licensed Granite LLMs using InstructLab's fine-tuning technology.

Benefits of Red Hat OpenShift cloud services

Accelerate time to value: Quickly build, deploy, and run applications that scale as needed.

Focus on innovation: Simplify opperations so your teams can focus on innovastion, not managing infrastructure.

Enhance efficiency: Improve consistency, efficiency, and security with proactive management and support.

Rely on expertise: Red Hat Reliability Engineers (SREs) ensure platform reliability, making deploying applications easier and less error prone.

- Red Hat OpenShift AI – an AI platform that provides tools to rapidly develop, train, serve and monitor machine learning models across distributed Kubernetes environments, making it vastly easier for agency domain experts—not just data scientists—to contribute their knowledge to fine-tune and improve LLMs.

- Red Hat Enterprise Linux AI (RHEL AI) – an end-to-end platform for building, testing and managing open source LLM models for enterprise applications on Linux server deployments.

Clater concludes: Red Hat's integrated approach reflects what has "always been the mission of Red Hat: to give our customers choice about where things run, whether on-premises, in any cloud, or on any edge device. The goal is to provide a consistent, more secure, and powerful platform."

The value for agencies: For federal leaders feeling overwhelmed by the AI landscape, he adds, the message is one of empowerment. The path forward does not require a blank check or a leap of faith into a black box. It requires a strategic pivot to a more secure, efficient, and transparent paradigm.

Find out more about Red Hat's vision for AI's open future.

This article was produced by Scoop News Group, for FedScoop and sponsored by Red Hat.

SCOOP NEWS GROUP Red Hat