

Solution brief

The new economics of AI for government

Managing the cost, security, and governance demands of agentic AI with Intel® CPUs.



Executive summary

Token consumption is becoming a defining factor in the economics of government AI.¹ As agencies adopt advanced and autonomous AI systems, overall costs can continue to rise even as model pricing declines.

Success will depend not only on model performance, but on how AI workloads are designed, deployed, and governed. This solution brief explores how architectural decisions like workload placement and local-first execution can help agencies improve token efficiency, strengthen data governance, and scale AI across edge, data center, and cloud environments.

For government leaders, the question is no longer whether AI works—it's how to scale AI across government operations while keeping control over cost, security, compliance, and public trust.

The future of
government AI will
be defined by how
AI workloads are
designed, deployed,
and governed.



Lower token prices don't guarantee lower AI costs

As agencies scale AI, the cost-per-query model has expanded into a broader cost tied to interactions, workflows, compliance reviews, and repeated executions.

Although token prices are expected to decline, overall costs can continue to rise when mission demand, citizen-service use cases, cybersecurity workloads, and agentic task complexity outpace efficiency gains. Every new AI-enabled workflow can increase interaction volume, and with it, token consumption.

Unlike traditional generative AI queries with one-to-one token usage per user, agentic AI systems run as ongoing loops that often involve dozens of queries. They reason, plan, retrieve data, call tools, and refine outputs repeatedly, and each step consumes tokens.

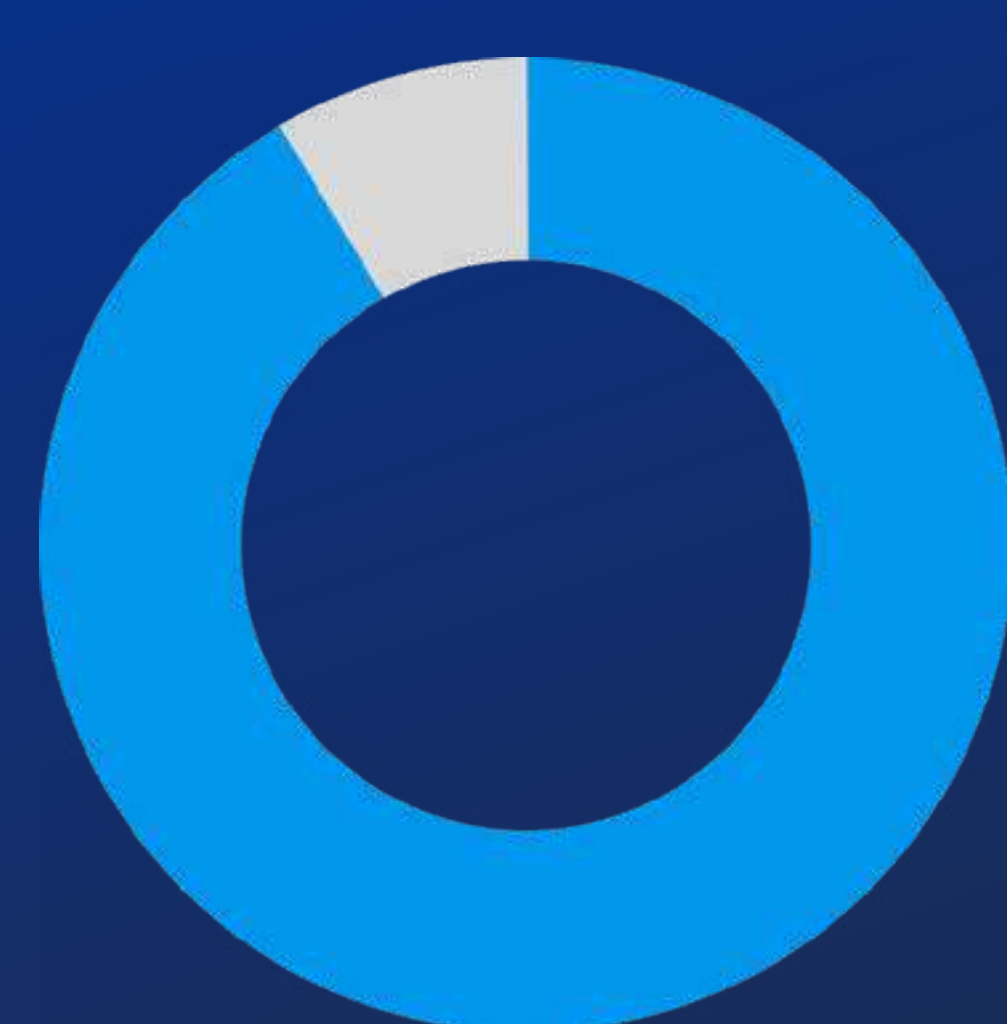


The result is a compounding effect on token usage rather than a linear increase. Workflows that appear simple at the business level might involve explosive growth in intermediate actions in cloud-based environments where every token is billed.

This creates a new type of cost volatility and can drastically reduce budgetary predictability. Organizations can improve productivity with AI, but without architectural optimization, they may also inadvertently escalate operating expenses.

Instead of evaluating cost per inference, government leaders must consider cost per successful mission task: the total cost needed to complete a workflow from initiation to outcome, including intermediate steps, retries, data processing, security controls, and evidence needed for auditability.

Design decisions affect how tokens are consumed and how sustainable, accountable, and defensible government AI adoption becomes.



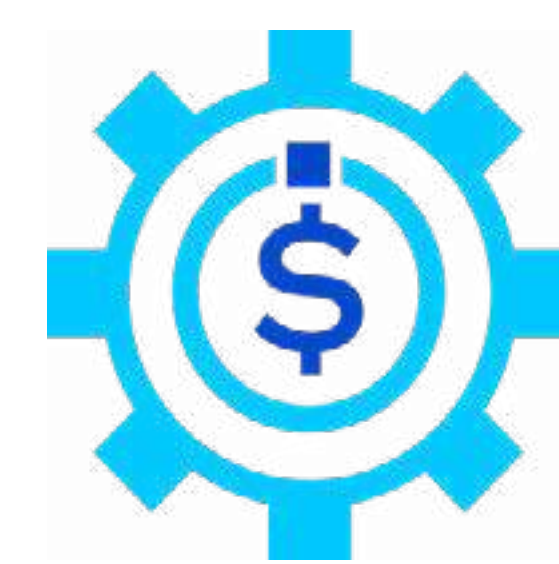
Over
90%
less inference cost
by 2030 vs 2025 for a
one-trillion-parameter
large language model²

The wrong architecture can make AI difficult to scale, govern, and afford

As token consumption grows, government leaders face decisions around how to balance performance, cost, scalability, compliance, and data security across complex AI systems.

Many agencies rely on cloud-based inference because it offers simplicity and speed to deployment. While effective for pilots, this model can become difficult to scale as token-related costs compound and sensitive data governance requirements increase.

Government leaders must decide which workloads should run locally, in agency data centers, at the edge, or in the cloud. They need to minimize unnecessary token generation and reduce reliance on high-cost inference environments without compromising security, compliance, and continuity.



Architecture is the primary lever for controlling cost, scale, compliance, and trust.

Key architectural decisions include:

- Where workloads should run.
- How to reduce unnecessary token generation.
- How to balance cost and mission quality.
- How to keep security and compliance at scale.



Reduce token usage with local-first, hybrid AI execution

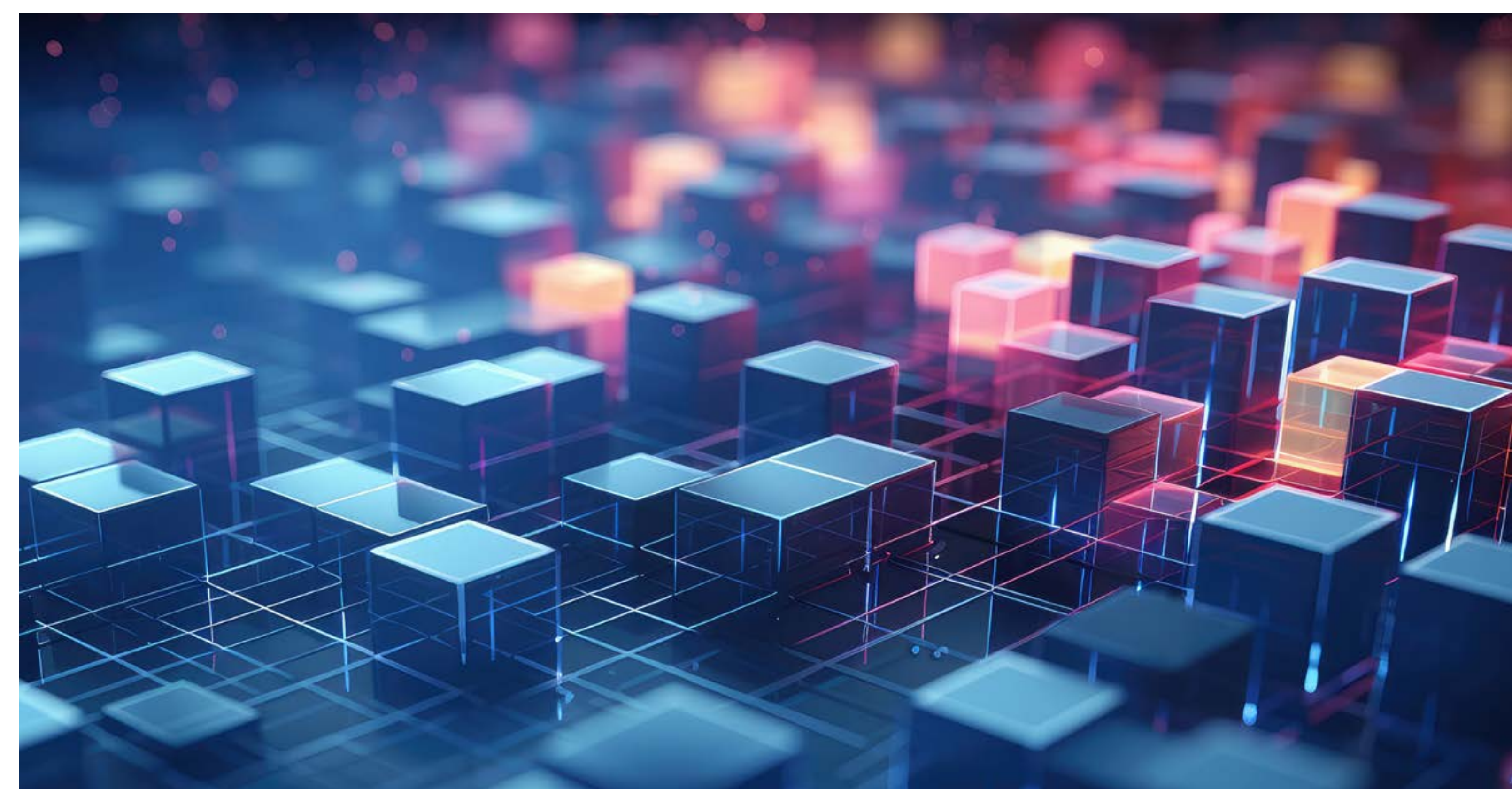
Intel's approach to AI infrastructure focuses on a critical source of government AI cost growth: inefficient system design and suboptimal workload placement across agency environments.

Rather than optimizing models alone, Intel helps agencies improve how, where, and when AI workloads run so leaders can align AI investments to mission outcomes, budget realities, and governance requirements.

This architectural shift can lower token consumption and address the broader agentic AI trilemma for government: cost, mission scale, and data security.

A hybrid AI execution model is particularly relevant for agentic systems because many repetitive operations can be handled locally before invoking cloud AI models. File access, document parsing, context preparation, planning, orchestration, case-file summarization, retrieval, tool routing, content generation, and intermediate reasoning often account for a large share of token consumption in agentic AI workloads.

By running high-frequency AI tasks on client devices, at the edge or in agency data centers, organizations can reduce unnecessary cloud interactions while also keeping sensitive information under tighter local control.



Intel® SuperClaw outcomes:

- Reduced dependency
- Improved system efficiency
- Stronger data governance
- Greater deployment flexibility

[Intel® SuperClaw](#) is a hybrid agentic AI solution built around this practical division of labor: sensitive and high-frequency tasks run locally on AI PCs, edge systems, or private infrastructure, while cloud models are used only when advanced reasoning or external information is needed. The result is a highly private, scalable AI solution for coding, research, email, and workflows.

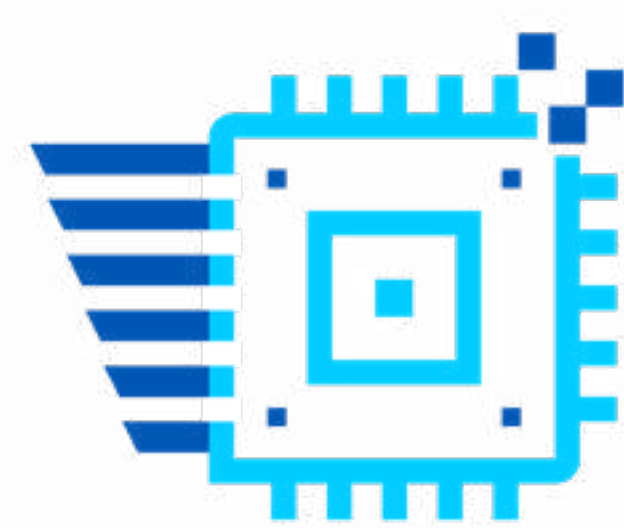
Intel® Xeon® 6 processors also support agencies by combining selected workloads onto fewer, more efficient systems, depending on workload, configuration, and baseline environment. This can reduce the number of servers, cores, licenses, and operational resources needed for certain AI environments.



Right-size workloads with CPU-based execution

Many government organizations assume AI workloads require GPU-based environments. In practice, some agentic AI workflows like orchestration, retrieval-augmented generation, document processing, case summarization, and smaller model inference can run on CPUs, depending on model size, latency requirements, software stack, and deployment configuration.

Intel Xeon 6 processors include built-in AI acceleration through Intel® Advanced Matrix Extensions (Intel® AMX), which can support inference on selected workloads without requiring specialized accelerators for every use case.



2x increase in throughput with Intel AMX enabled eliminating the need for dedicated GPU deployments in many use cases⁴



Up to
10:1
server consolidation⁵

Intel Xeon 6 processors can improve AI inference performance and concurrency for selected CPU-based workloads, depending on model, software stack, configuration, and SLA requirements.

For agencies, the result can be a balanced and cost-efficient architecture. GPUs can be reserved for workloads where they add the most value, while CPUs can support orchestration and selected inference workloads across environments in which agencies already operate.



Optimizing AI workload placement across environments

One of the most important levers for token efficiency and governance is workload placement.

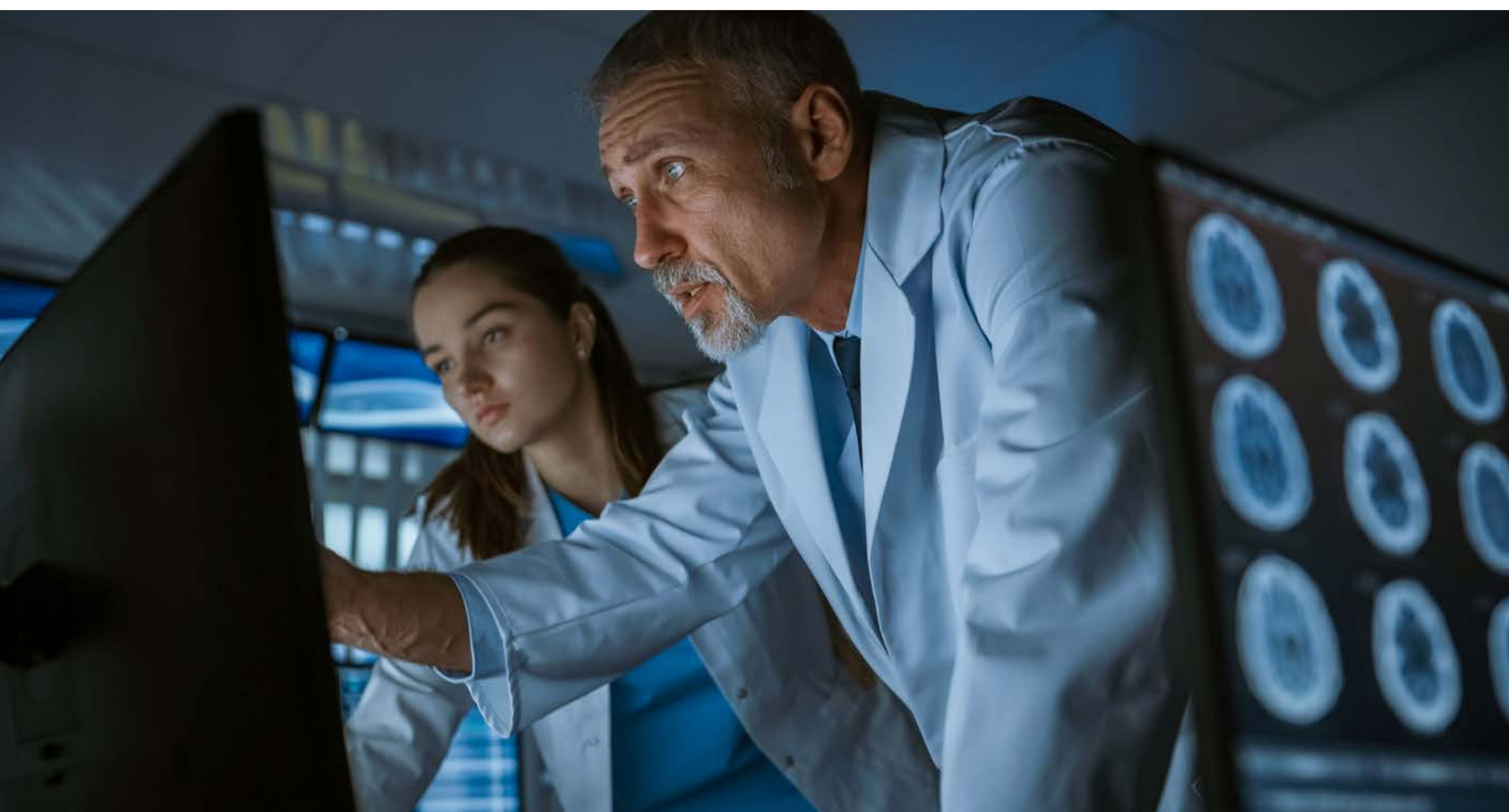
Agencies that optimize AI workflow across local systems, edge infrastructure, private data centers, sovereign cloud, and public cloud environments can reduce token usage and infrastructure cost while keeping sensitive data under control.

Intel® platforms can support a distributed architecture in which CPUs help coordinate workloads across environments, allowing agencies to execute selected workloads in locations that better align with cost, performance, latency, classification, compliance, and data governance requirements.

For government leaders, the result is reduced dependency on expensive cloud inference, improved system efficiency across agency environments, stronger governance over sensitive government data, and greater flexibility in scaling AI deployments without disrupting mission systems.

Agentic systems often introduce hidden inefficiencies through retries, redundant processing, and repeated retrieval across complex workflows. Each retry loop consumes more tokens, increasing cost without improving mission outcomes.

Intel platforms can provide predictable performance and integrated capabilities that may help reduce these inefficiencies, depending on workload design, deployment configuration, and governance requirements. The potential outcome is improved reliability, reduced unnecessary token consumption, and an accountable operating model for AI at scale.



Every retry increases token consumption without improving outcomes.

Hidden cost drivers:

- Retries
- Redundant processing
- Repeated retrieval

Throughput alone doesn't solve government AI challenges

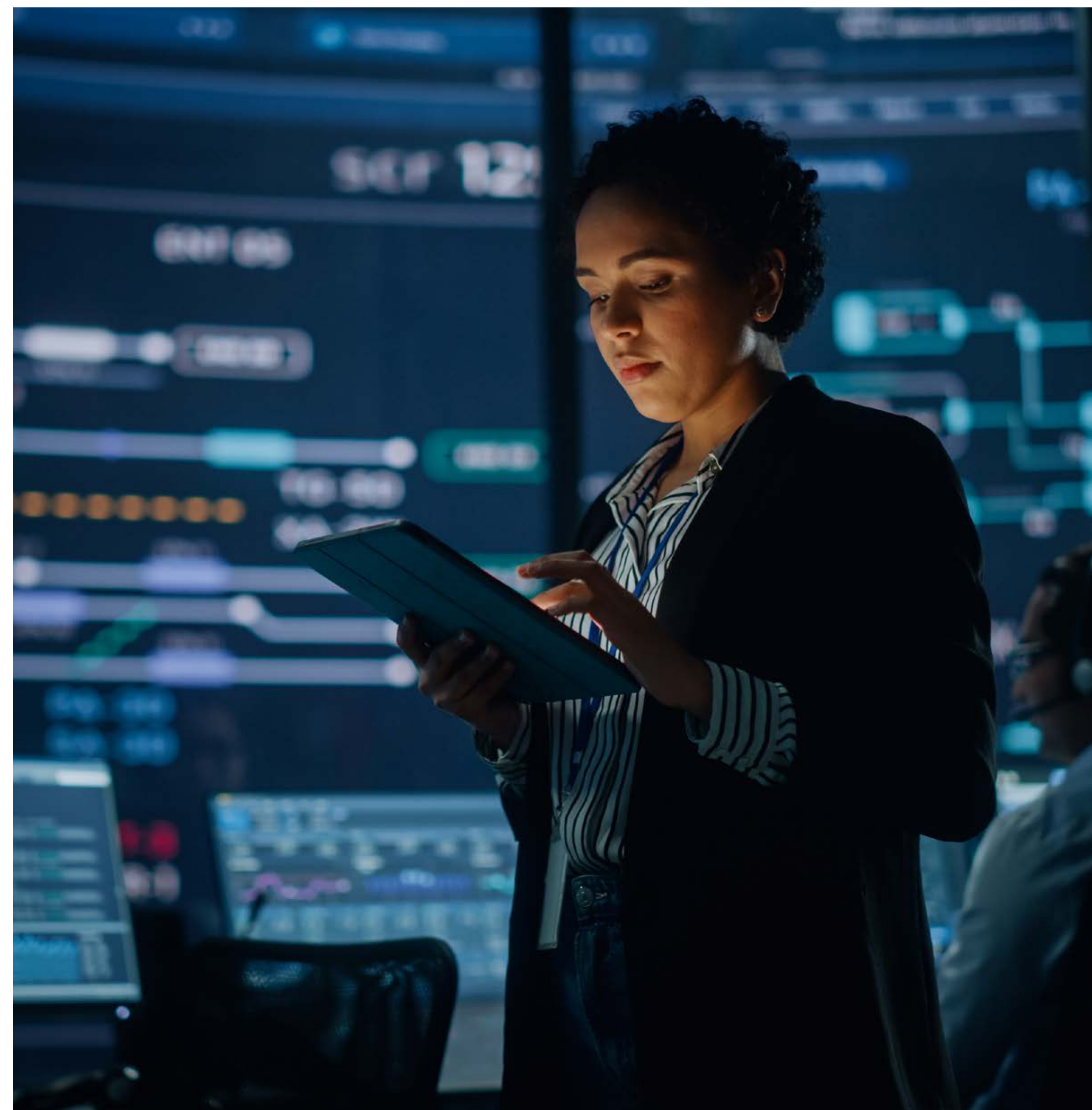
Some accelerated computing strategies emphasize large-scale inference throughput, including tokens per second and centralized AI factory architectures. These approaches can be effective for parallel workloads and large-scale model serving, but government leaders should also evaluate how they affect data placement, auditability, procurement flexibility, and long-term operating cost.

Intel emphasizes system-level efficiency and local-first execution. By supporting local execution, consolidation, CPU-based AI processing, and selective cloud use, Intel platforms can help agencies improve token efficiency, manage infrastructure cost, and keep sensitive data closer to the systems that own it.

For government leaders, the goal is clear: greater cost control, improved efficiency, stronger data governance, mission continuity, and a scalable path to responsible AI adoption.



Optimizing workload placement can help agencies reduce expensive cloud inference, improve system efficiency, strengthen data governance, and scale AI with more flexibility.



What comes next for government AI

The four shifts shaping the next phase of government AI:

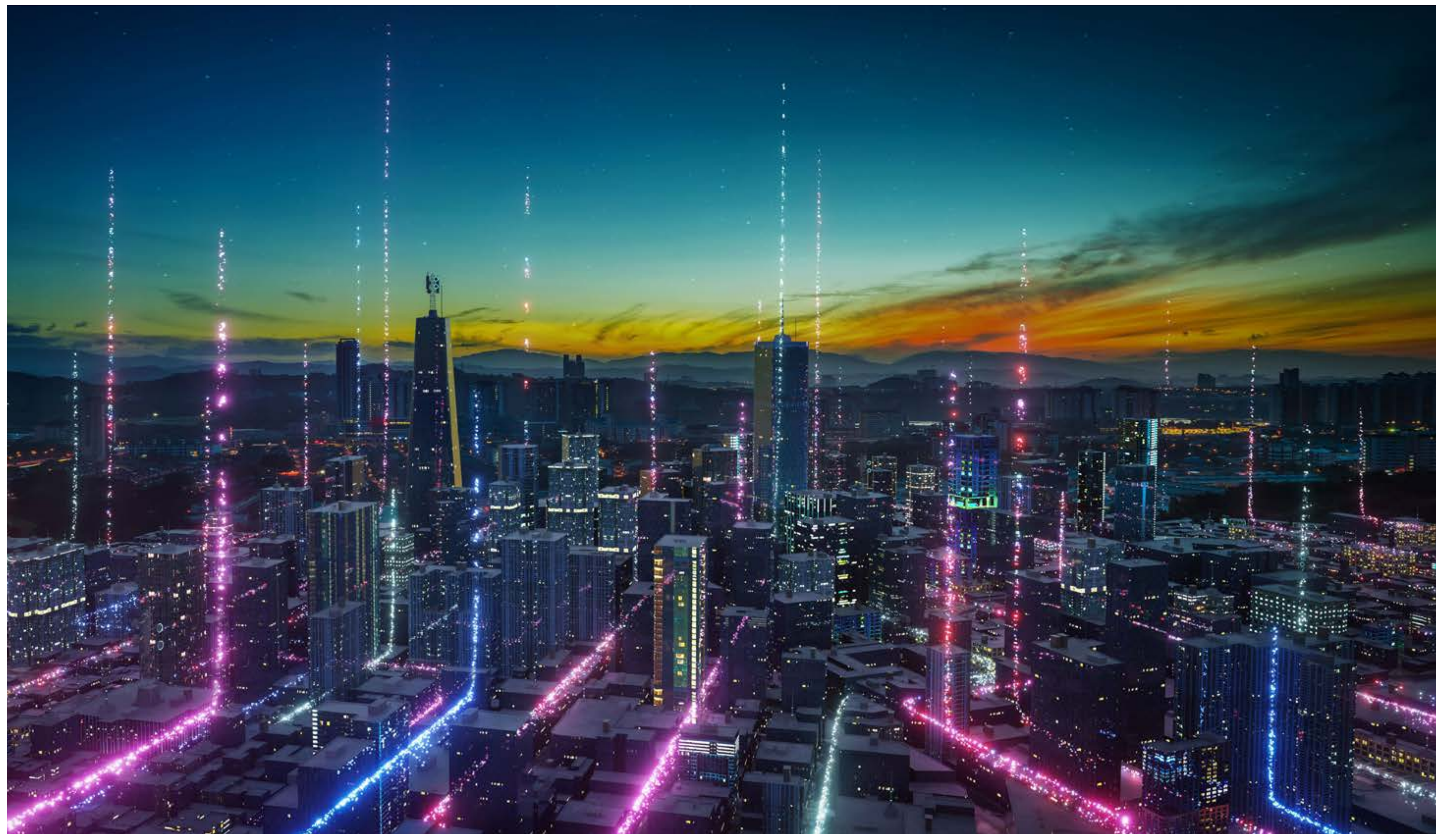
1. AI moves from assistants to mission orchestration.
2. Smaller, specialized models move work closer to the data.
3. AI infrastructure becomes more distributed.
4. Governed and auditable AI becomes essential.

Government AI is expected to move beyond isolated assistants and productivity experiments toward orchestrated systems that run across agencies, programs, and mission functions. Agentic AI may become embedded in core workflows, helping coordinate tasks, access data, interact with tools, and support decision-making across operations, citizen services, cybersecurity, public safety, and knowledge work.

As this shift accelerates, agencies will need to govern, secure, audit, and scale AI systems. This makes architecture, observability, and workload placement strategic priorities rather than back-end implementation details.

Smaller, task-specific models can often deliver strong results for use cases such as classification, summarization, extraction, routing, document review, and domain-specific analysis while running on CPUs, client devices, edge systems, and private infrastructure. By matching the model to the task and running work closer to where data is created or stored, agencies can reduce latency, protect sensitive information, lower cloud dependency, and control costs.

At the same time, AI infrastructure is becoming more distributed. Agencies will execute workloads across AI PCs, edge systems, data centers, private clouds, sovereign clouds, and public clouds based on performance, compliance, data sensitivity, and cost requirements. Effective AI architectures will determine where workloads run, which models are used, what data remains local, and when cloud resources are justified. CPUs play a critical role as a flexible coordination and execution layer across this infrastructure.



Conclusion

AI economics is changing for government as token-based consumption becomes a key driver of cost.

As AI systems gain autonomy and the number of models grows, success will depend on deploying the right AI workloads at mission scale.

Agencies need visibility into where data resides, which models are being used, and how outputs are generated. By matching each task to the appropriate model, routing workloads to the right environment, and keeping sensitive data within trusted boundaries, organizations can improve efficiency while strengthening governance, accountability, and data control.

Leading agencies will adapt to this new economic reality by optimizing infrastructures that deliver outcomes across local, edge, data center, sovereign cloud, and public cloud environments. Intel helps agencies scale agentic AI while managing costs, making every token count toward mission outcomes.



Next steps

USAI is our commitment to American-based innovation and AI excellence. Explore these resources to learn how a smarter architecture can help government organizations scale AI while maintaining control over cost, security, and performance.

[USAI: AI enablement for government](#)

Learn how Intel helps government organizations assess, pilot, and scale AI solutions while maintaining governance, security, and operational control.

[Intel AI Builder: SuperClaw](#)

Discover Intel's hybrid agentic AI approach, where sensitive tasks run locally while cloud models are reserved for advanced reasoning.

[AI with Intel Xeon](#)

Explore interactive guidance, benchmarks, and deployment architectures that show how Intel Xeon supports AI workloads.



^{1,2} Based on “Gartner Predicts That by 2030, Performing Inference on an LLM With 1 Trillion Parameters Will Cost GenAI Providers Over 90% Less Than in 2025,” a report by Gartner, March 25, 2026. For more information, visit www.gartner.com/en/newsroom/press-releases/2026-03-25-gartner-predicts-that-by-2030-performing-inference-on-an-llm-with-1-trillion-parameters-will-cost-genai-providers-over-90-percent-less-than-in-2025.

³ Based on “Solving the Agentic AI Trilemma – Cost, Scale, and Data Security,” Intel Newsroom, May 21, 2026. For more information, visit newsroom.intel.com/opinion/solving-the-agentic-ai-trilemma-cost-scale-and-data-security.

⁴ Based on “Intel AMX Enables High-Efficiency CPU Inference for AI Workloads,” OpenMetal, Inc., April 7, 2025. For more information, visit openmetal.io/resources/blog/intel-amx-ai-inference-performance.

⁵ See [7T21] at www.intel.com/processorclaims: Intel Xeon 6. Results may vary.

⁶ See [7T221] at www.intel.com/processorclaims: Intel Xeon 6. Results may vary.

SuperClaw is built based on the OpenCode framework, with additional hybrid AI capabilities, privacy controls, local context management, model routing, governance, and platform optimization developed by Intel.

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Other names and brands may be claimed as the property of others.

Intel is committed to the continued development of more sustainable products, processes, and supply chain as we strive to prioritize greenhouse gas reduction and improve our global environmental impact. Where applicable, environmental attributes of a product family or specific SKU will be stated with specificity. Refer to [Intel Corporate Responsibility Report](#) or visit www.intel.com/2030goals for further information.

© Intel Corporation. Intel, the Intel logo, Xeon, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.