CISA and Cyber Centre Issue Guidance on Managing AI Token Consumption



CISA, Canada's Cyber Centre and partners published joint guidance on managing token usage in third-party AI services, covering cost, availability and denial-of-wallet risks, and offering practical steps for resilience.

- The guidance is a joint release by CISA, the Canadian Centre for Cyber Security and partners.
- Token-based usage models in third-party AI services can affect cost, performance and access to critical services.
- Unmanaged consumption may lead to unpredictable costs, reduced availability, token exhaustion, and denial-of-tokens or denial-of-wallet attacks.
- Recommended practices include prioritizing token allocation, efficient workflows, better prompting, caching, data optimization and ongoing monitoring.

> - Critical infrastructure operators are encouraged to identify AI-dependent mission-critical functions and plan for reduced AI availability.

CISA and the Canadian Centre for Cyber Security (Cyber Centre), together with partners, have released joint guidance titled Managing AI Token Consumption for Operational Resilience. It helps organizations understand and manage token usage as they adopt third-party AI services.

Why token usage matters

Many third-party AI services are billed and limited by tokens. According to the guidance, this model can affect cost, performance and access to critical services. When consumption is not managed, organizations may face rising and unpredictable costs, reduced service availability, degraded performance and token exhaustion. The guidance also names denial-of-tokens and denial-of-wallet attacks as possible risks.

What the guidance recommends

The document outlines practical steps in several areas:

- Prioritize how tokens are allocated.
- Design efficient workflows.
- Improve how users write prompts.
- Use caching and reusable context.
- Optimize data before submitting it.
- Monitor token consumption over time.

Critical infrastructure owners and operators can use it to identify mission-critical functions that depend on AI, understand provider and operational dependencies, and plan for reduced AI availability during outages, incidents or periods of high demand. Cyber defenders can also use it when managing AI-enabled operations.

The guidance is not specific to any one sector or jurisdiction, but its themes of dependency mapping, availability and third-party risk are relevant to regulated institutions in the Kingdom that are adopting AI services.

What to do now

1. Inventory business functions that depend on third-party AI services and rank which are mission-critical.
2. Review provider contracts and technical limits for token quotas, rate limits and billing models.
3. Set token budgets, consumption monitoring and alerts to detect abnormal spikes that may signal denial-of-wallet abuse.
4. Include reduced or lost AI availability in business continuity and incident response plans.

5. Issue internal guidance on efficient prompting, caching and trimming data before submission.