

Managing AI token consumption for operational resilience



Communications Security
Establishment Canada
Canadian Centre
for Cyber Security

Centre de la sécurité des
télécommunications Canada
Centre canadien
pour la cybersécurité

Canada



Partnering agencies



Communications Security
Establishment Canada

**Canadian Centre
for Cyber Security**

Centre de la sécurité des
télécommunications Canada

**Centre canadien
pour la cybersécurité**



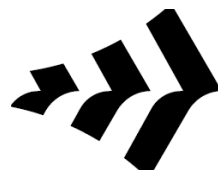
Australian Government

Australian Signals Directorate

ASD AUSTRALIAN
SIGNALS
DIRECTORATE

ACSC Australian
Cyber Security
Centre



 **National Cyber
Security Centre**
NEW ZEALAND



UK Government



**National Cyber
Security Centre**
a part of GCHQ



Table of contents

Demystifying AI tokens	3
How AI systems use tokens	4
Risks related to AI token usage	4
Rising and unpredictable costs	5
Reduced access to critical services	5
Limited visibility and oversight	5
Overall performance	5
Token exhaustion	5
Denial-of-tokens and denial-of-wallet attacks	6
Managing AI token usage	9
Prioritize token allocation	9
Design efficient workflows	9
Improve user practices and prompting	10
Leverage caching and reusable context	10
Optimize data before submission	10
Harness AI effectively	10
Optimizing AI token usage	11
Right-sizing the model	11
Governance and oversight	11
Resiliency and operational management	11
Organizational awareness and responsible use	12
Monitoring and continuous improvement	12
Learn more	13



Demystifying AI tokens

Many organizations are adopting third-party artificial intelligence (AI) services that use large language models (LLMs) to generate content, summarize information, answer questions, analyze files, and automate tasks.

These services operate under different usage and licencing policies than traditional software. Instead of paying per user or licence, organizations pay for AI services based on usage, which is measured in tokens. While many cloud and software services already use consumption-based pricing models (based on metrics such as throughput, transactions or compute usage), AI services differ as they commonly charge based on the amount and type of input and output, as well as the processing performed.

This pricing can be more difficult for organizations to predict and manage, as consumption varies according to the length of prompts and responses, the context provided to the model and the specific AI capabilities being used.

AI tokens are units of data (text, characters, phrases) that AI models use to read, understand and process language.

There are two different types of tokens:

- Input tokens are those that a user sends or inputs into the model
 - A subset of input tokens is referred to as cached tokens
 - Cached tokens allow the AI model to reuse input from previously processed prompts
- Output tokens are those that the AI tool generates in the response provided to the user

The AI model breaks text into small units to better analyze the language and provide an appropriate response. This process is known as tokenization.

Tokenization converts raw inputs (text, images, audio, video) into numerical sequences a model can learn from. It is a mechanism for measuring usage and the underlying representation the model uses to work with language.

Tokens are the fundamental units through which AI models process and generate language. Every input and output is transformed into a sequence of tokens that the model processes.

Token usage can change depending on the complexity of the task. Simple tasks, such as answering a basic question, may use hundreds of tokens. More advanced AI workflows, such as analyzing large documents or completing multi-step automated processes, can consume significantly more tokens, ranging in the tens of thousands. Actual usage varies depending on the AI service, model and amount of information processed.



This publication explains how organizations can better understand, manage and reduce AI token usage to support reliable and cost-effective operations. Organizations, despite size or sector, should use this guidance to assist in their understanding of token usage and pricing models prior to engaging with vendors and signing service level agreements or contracts.

Cyber defenders are encouraged to review this guidance given the growing use of AI-enabled tools and the operational dependence on AI.

Critical infrastructure owners and operators are also encouraged to review this guidance to assist in identifying mission-critical functions, understanding AI dependencies, and deciding what resilience, continuity planning, and resource prioritization are needed if AI services are degraded or unavailable.

How AI systems use tokens

AI token use depends on how systems are designed and used. Understanding these factors can help organizations manage consumption more effectively.

Token usage and the associated cost increases when:

- larger inputs and outputs, such as lengthy text and images, require the AI tool to process and generate more content
- large reference datasets and connector payloads are not cached or optimized
- additional information, data and conversation history is included with requests
- multi-step workflows involve several actions or tasks within a single process
- repeated requests process the same data multiple times
- integrated systems retrieve or process data across connected platforms
- agentic workflows compound costs, with each step reprocessing prior content and tool outputs

As AI continues to advance and integrate into business operations, token usage will also increase rapidly. Organizations should implement robust oversight and management of AI token usage and should consider whether an externally hosted model is the most appropriate tool for a task. Smaller models, locally run models from trusted sources, or specialized AI tools may achieve similar outcomes while using fewer resources. Organizations will have to determine whether using a model running locally or a model offered as a service best suits the business need.

Risks related to AI token usage

AI can introduce operational, financial and security risks if it is not properly managed. Below are common risks that organizations may face when managing AI token usage.



Rising and unpredictable costs

AI token usage can increase rapidly depending on how systems are designed and used. Complex workflows, repeated requests and automated processes can make costs difficult to predict and manage over time.

Reduced access to critical services

Without clear prioritization, lower-value activities can consume capacity needed for important operations. During periods of high demand, this can lead to delays, degraded performance or disruption of critical services.

Conversely, prioritization that only favours critical workloads may result in the delay of low-priority activities. Organizations should balance resource allocation to ensure critical services remain available while maintaining reasonable access for other users and business functions.

Limited visibility and oversight

If monitoring and reporting tools are not implemented, it can be difficult to understand how tokens are being used across the organization. This limits the ability to identify inefficiencies, detect unusual activity and enforce usage control. Ensuring the use of the most appropriate monitoring and reporting tools will help organizations better understand token usage.

Overall performance

An increased use in tokens can lead to poorer performance overall. For example, the response time of an AI tool consuming larger volumes of tokens can lag, creating longer wait times for the user to receive a response. Larger token counts also put pressure on organizational computing resources and require additional memory. Larger prompts can also lead to inconsistent results, conflicting responses, or duplicate content.

Token exhaustion

Token exhaustion can occur when an organization reaches its allocated token limits, usage quotas, rate limits, or budget thresholds. This may result from legitimate high-volume use, poorly optimized applications, or an unexpected increase in demand.

When token exhaustion occurs, users and systems may experience delays, degraded performance, failed requests, or loss of access to AI capabilities. Organizations that depend on AI-enabled services for operational or mission-critical functions may be particularly affected if contingency measures are not in place.



Organizations can set alerts for impending token exhaustion. These alerts can be set to occur progressively as users near organizationally set benchmarks for token exhaustion. For example, an information-only alert could be triggered when an organization reaches 70% of the token usage quota. Once the organization reaches 80% of token exhaustion, the alert would then be a warning. Finally, once the organization reaches 90% of token exhaustion, the alert would be considered critical.

Denial-of-tokens and denial-of-wallet attacks

A denial-of-tokens attack is a form of resource-consumption attack. Threat actors intentionally generate excessive AI usage to exhaust token allocations, consume budgets and resources, or trigger rate limits. They achieve this through automated prompts, abuse of publicly accessible AI interfaces or tools, or manipulation of AI-enabled workflows to process unusually large volumes of data. By depleting available tokens or increasing operational costs, a denial-of-token attack can reduce service availability, disrupt the organization's legitimate users, and impair an organization's ability to access AI services when needed.

Denial-of-wallet attacks aim to cause organizations financial harm. Threat actors intentionally exploit consumption-based, usage-based, and pay-per-use pricing models to increase costs. Examples of how threat actors conduct denial-of-wallet attacks include:

- prompting AI tools with complex inputs that require robust data analysis and result in long outputs
- creating AI loops (for example, inputs that require the AI tool to feed its response output back into the tool)



Considerations for cyber defenders

Cyber defenders have the opportunity to leverage AI tools to their advantage to mitigate cyber threats. AI can assist cyber defenders in discovering vulnerabilities faster and at a larger scale, allowing for expedited remediation efforts to be taken.

AI presents a significant opportunity for cyber defenders. When used safely, securely and responsibly AI can:

- strengthen prioritisation of cyber risks
- improve detection of threats and vulnerabilities
- support faster response and recovery
- reduce reliance on repetitive manual tasks

While AI provides potential benefits in cyber defence, it does not come without a cost. Token usage within cyber defence tasks and actions will increase overall usage and associated costs for organizations. To contend with this, cyber defenders and organizations should look for ways to leverage smaller models that can complete less complex cyber security tasks. Organizations should also prioritize complex cyber security tasks and strategically deploy LLMs and agents that require larger token expenditures to meet these needs. A multi-model approach can help organizations reduce token usage and associated costs.

Despite the opportunities presented, cyber defenders and organizations must take note that AI is not a replacement for strong cyber security fundamentals. Human oversight, governance and Secure by Design practices remain essential.

For more information, read our joint guidance [Opportunities for AI in cyber defence](#).



Considerations for critical infrastructure

Critical infrastructure (CI) organizations operate in higher risk environments where disruptions can have widespread impacts. AI token usage should be managed with a focus on resilience, sustained operations and awareness of which systems and functions depend on AI.

CI organizations should:

- understand how AI services support critical operations and essential services
- identify dependencies between AI systems, service providers and business operations
- understand how AI deployment models may affect service availability, capacity and recovery during disruptions
- plan for reduced AI availability during outages, incidents or periods of high demand
- consider how emergencies or heightened threat activity could increase AI usage demands
- integrate AI usage into broader organizational continuity planning
- maintain non-AI capabilities, such as previous digital automation controller processes or manual operations training, in case the AI becomes unavailable or untrustworthy

For more guidance on CI, read [Critical infrastructure resilience and escalated threat navigation](#) and our [joint guidance Isolating vital systems](#). These publications offer additional guidance on isolating systems of critical importance during significant cyber incidents or sustained high risk environments.

We also recommend organizations read our joint guidance [Principles for the Secure Integration of Artificial Intelligence in Operational Technology](#). This guidance outlines four key principles owners and operators can follow to realize the benefits of integrating AI into OT systems while reducing risk. It focuses on machine learning, large language model-based AI, and AI agents because of the complex security considerations and challenges they pose. The guidance also applies to systems augmented with traditional statistical modeling and logic-based automation.



Managing AI token usage

To effectively control AI costs and maintain reliable operations, the authoring agencies recommend organizations adopt a tiered governance model. This approach prioritizes token usage based on business impact, ensuring critical functions are protected while lower priority activities are appropriately managed.

Prioritize token allocation

Organizations should consider separating token allocations across the following tiers. This can help ensure that routine or high-volume activities do not consume resources needed for mission-critical functions. While dedicated capacity may increase costs, it can improve resilience and support operational continuity.

Tier 1: Mission critical

This tier includes high-priority functions such as cyber defence, incident response and threat intelligence. These services should receive the highest priority for token availability and operational continuity.

Tier 2: Operational

This tier covers essential business processes such as benefits processing, compliance activities and policy analysis. Token usage for these services should be monitored and managed to support reliable operations and cost control.

Tier 3: Productivity

This tier includes everyday user activities such as drafting emails, summarizing meetings and doing general research. These lower-priority activities may be subject to usage limits or reduced access during periods of high demand.

Design efficient workflows

AI service providers often charge different rates for input tokens, output tokens, cached tokens, and other processing activities. As a result, organizations should design AI workflows to use resources efficiently.

In practice, this means providing sufficient context to complete a task while avoiding unnecessary prompts, excessive outputs, or repeated processing of the same information.



Improve user practices and prompting

Users should be trained to interact effectively with AI systems. Clear and specific prompts can help reduce unnecessary token consumption and improve the quality of results. For example, users may be able to reduce costs by providing only the relevant excerpts of a document rather than processing an entire file.

Where supported, organizations should take advantage of features that promote efficient and consistent use of AI services. Examples include reusable instructions, templates, and system-level configurations that encourage concise outputs and reduce the need to repeatedly provide the same context.

Leverage caching and reusable context

Some AI services offer caching mechanisms that allow the results of previously processed content, such as large documents, code repositories, or system instructions, to be reused by subsequent requests.

When enabled, caching can reduce processing requirements, improve performance, and lower costs when the same information is processed repeatedly.

Optimize data before submission

Organizations may also consider preprocessing, optimization, or filtering tools to reduce unnecessary data before it is submitted to an AI system. These tools can help eliminate redundant content, reduce token consumption, and improve overall efficiency.

Any tools should be assessed in accordance with the organization's risk management frameworks, as well as security, privacy, and data protection requirements before deployment.

Harness AI effectively

We recommend organizations implement AI harnessing. This requires organizations to engineer the software infrastructure supporting the AI agent to help ensure reliable and autonomous execution of tasks and prompts.

Organizations should embed AI into existing governance structures, workflows and applicable decision-making processes and procedures when it makes sense, when it is beneficial, and when it can be properly assessed and secured. By identifying use cases where AI can improve operations and performance or can provide enhanced problem solving, organizations can get a better return on investment into AI tools.



The authoring agencies recommend that organizations conduct risk assessments before allowing AI agents to operate autonomously in their environment.

Optimizing AI token usage

Organizations should take a proactive and risk-informed approach to managing AI token usage across systems and workflows. Tokens can be managed effectively by following the recommended actions and practices below.

Right-sizing the model

Using the same high-cost model across all workflows is not necessary. Organizations can save costs by right-sizing the model to the specific task at hand. This will reduce the per-token cost. Additionally, configuring the appropriate reasoning or effort level helps limit the generation of output tokens, which are typically more expensive.

Since full context is sent with every conversation, LLMs leverage low-cost caching mechanisms for previously calculated answers based on input text. As a result, organizations can benefit from positioning static or reusable context information toward the beginning of the skill definition and associated architectural files. This approach maximizes cache efficiency and further contributes to overall cost optimization. Organizations should note that each LLM has specific approaches and requirements for setting up the cache feature, so some trial and error may be necessary.

Governance and oversight

- Establish an AI governance working group that includes IT security, finance, legal, and operational stakeholders
- Conduct an AI workload assessment to determine operational importance and risk
- Define disaster recovery procedures and operational playbooks that include human oversight at key decision points during outages or unusual system behaviour
- Require human authorization for consequential AI-driven actions as per organizational risk assessments, policies, procedures, and guardrails

Resiliency and operational management

- Adopt a multi-provider architecture that supports vendor diversity and reduces dependency on a single provider
- Select the lowest cost model that meets operational requirements
- Route AI workloads to available providers during outages or service disruptions to improve resiliency
- Deploy dashboards to monitor token consumption and set thresholds for alerting the appropriate contacts



- Defer lower-priority productivity use cases until mission-critical workloads have been tested to ensure they can handle the expected demand
- Reuse approved templates for common tasks
- Recognize when traditional tools, smaller or locally run models from trusted sources, or manual processes are more efficient than LLMs

Organizational awareness and responsible use

- Provide training on responsible and efficient prompting practices, including concepts and examples such as:
 - enabling prompt caching
 - avoiding unnecessary regeneration of response and repeated processing of the same content
 - providing and sending relevant context
 - applying a cap on output length
 - binding agentic loops
 - limiting requests to the information required for the task (for example, providing excerpts instead of the full document)
- Educate staff on organizational AI usage policies, approval processes and acceptable use requirements
- Train staff to use AI tools and be able to recognize when AI use is appropriate versus when traditional tools or manual processes may be more effective
- Promote awareness of how excessive or unnecessary AI usage can affect operational capacity, service availability and cost
 - Understand how excessive context, lengthy outputs, and multi-step workflows can increase token consumption and costs

Monitoring and continuous improvement

- Establish token usage baselines and forecasting practices to support budgeting and capacity planning
- Implement usage alerts to identify unusual token consumption or approaching usage limits
- Monitor unusual increases in token consumption as a potential indicator of data exfiltration, similar to how abnormal bandwidth spikes can signal unauthorized data transfers
- Feed AI token telemetry into a security information and event management platform and correlate it with identity, cloud, API and system logs
- Regularly review AI usage trends and token consumption patterns to identify inefficiencies or unexpected cost increases
- Consider using AI gateways, proxies or centralized monitoring solutions to provide visibility into AI usage, token consumption, costs, policy compliance, and operational trends across AI services



- AI proxies are becoming an increasingly practical way for organizations to monitor and measure consumption and use, enforce quotas, apply guardrails and support chargeback
- Assess third-party AI providers for service reliability, quota limitations and operational risks related to token availability
- Update organizational guidance, controls and usage policies as AI adoption and token demands evolve

Learn more

- [Top 10 artificial intelligence security actions: A primer \(ITSAP.10.049\)](#)
- [Frontier artificial intelligence \(ITSAP.10.050\)](#)
- [Joint guidance on the careful adoption of agentic artificial intelligence services](#)
- [Australian Cyber Security Centre AI guidance](#)
- [Cybersecurity and infrastructure security agency AI guidance](#)
- [New Zealand National Cyber Security Centre AI guidance](#)
- [United Kingdom National Cyber Security Centre AI guidance](#)