

Gérer la consommation de jetons d'IA pour favoriser la résilience opérationnelle



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

Communications Security
Establishment Canada
Canadian Centre
for Cyber Security

Canada



Agences partenaires



Communications Security
Establishment Canada

**Canadian Centre
for Cyber Security**

Centre de la sécurité des
télécommunications Canada

**Centre canadien
pour la cybersécurité**

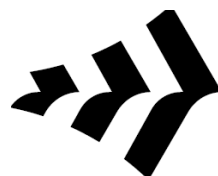


Australian Government

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 des matières

Mieux comprendre les jetons utilisés par l'intelligence artificielle	3
Comment les systèmes d'IA utilisent les jetons	4
Risques liés à l'utilisation des jetons d'IA.....	5
Hausse et imprévisibilité des coûts	5
Réduction de l'accès aux services essentiels.....	5
Visibilité et surveillance limitées	5
Performance globale.....	6
Épuisement des jetons.....	6
Attaques par épuisement des jetons et attaques par explosion des coûts.....	6
Gérer la consommation de jetons d'IA	10
Prioriser l'attribution des jetons.....	10
Concevoir des flux de travaux efficaces	10
Améliorer la formulation des requêtes et les pratiques des utilisatrices et utilisateurs	11
Tirer avantage de la mise en cache et du contexte réutilisable.....	11
Optimiser les données avant leur soumission.....	11
Tirer efficacement parti de l'IA	12
Optimiser la consommation de jetons d'IA	12
Adapter le modèle à la tâche	12
Gouvernance et surveillance.....	13
Résilience et gestion opérationnelle	13
Sensibilisation organisationnelle et utilisation responsable.....	13
Surveillance et amélioration continue	14
Pour en savoir plus	14



Mieux comprendre les jetons utilisés par l'intelligence artificielle

De nombreuses organisations adoptent des services d'intelligence artificielle (IA) de tiers qui utilisent de grands modèles de langage (GML) pour générer du contenu, résumer de l'information, répondre à des questions, analyser des fichiers et automatiser des tâches.

Ces services fonctionnent selon des modalités d'utilisation et de licence différentes de celles des logiciels conventionnels. Au lieu de payer par utilisatrice ou utilisateur, ou par licence, les organisations paient les services d'IA en fonction de leur utilisation, qu'on mesure en jetons. Bien que de nombreux services infonuagiques et logiciels reposent déjà sur des modèles de tarification à l'utilisation (fondés sur des mesures comme le débit de traitement, le nombre de transactions ou l'utilisation des ressources de calcul), les services d'IA se distinguent par le fait qu'ils facturent généralement en fonction de la quantité et du type de données fournies en entrée et générées en sortie, ainsi que du traitement effectué.

Cette tarification peut être plus difficile à prévoir et à gérer pour les organisations, puisque la consommation varie selon la longueur des requêtes et des réponses, le contexte fourni au modèle et les capacités d'IA utilisées.

Les jetons d'IA sont des unités de données (texte, caractères ou groupes de mots) que les modèles d'IA utilisent pour lire, comprendre et traiter le langage.

Il existe deux types de jetons :

- Les jetons d'entrée sont ceux qu'une utilisatrice ou un utilisateur transmet au modèle.
 - Un sous-ensemble des jetons d'entrée est appelé jetons mis en cache.
 - Les jetons mis en cache permettent au modèle d'IA de réutiliser des données provenant de requêtes déjà traitées.
- Les jetons de sortie correspondent à ceux que l'outil d'IA génère dans la réponse fournie à l'utilisatrice ou à l'utilisateur.

Le modèle d'IA découpe le texte en petites unités afin de mieux analyser le langage et de produire une réponse appropriée. Il s'agit du processus de conversion en jetons.

La conversion en jetons transforme des données brutes (texte, images, contenu audio ou vidéo) en séquences numériques qu'un modèle peut utiliser pour son apprentissage. Il s'agit à la fois d'un mécanisme permettant de mesurer l'utilisation d'un service et de la représentation sous-jacente qu'utilise le modèle pour traiter le langage.

Les jetons sont les unités fondamentales à partir desquelles les modèles d'IA traitent et génèrent le langage. Chaque donnée fournie en entrée et chaque réponse produite en sortie sont transformées en une séquence de jetons analysés par le modèle.



Le nombre de jetons utilisés varie selon la complexité de la tâche. Les tâches simples, comme répondre à une question de base, peuvent nécessiter quelques centaines de jetons. Les processus d'IA plus avancés, comme l'analyse de longs documents ou l'exécution de tâches automatisées comportant plusieurs étapes, peuvent en consommer beaucoup plus, parfois des dizaines de milliers. L'utilisation réelle varie selon le service d'IA, le modèle utilisé et la quantité d'information traitée.

La présente publication explique comment les organisations peuvent mieux comprendre, gérer et réduire leur consommation de jetons d'IA afin de favoriser des activités fiables et rentables. Peu importe leur taille ou leur secteur d'activités, les organisations devraient consulter ces conseils pour mieux comprendre la consommation de jetons et les modèles de tarification avant d'entreprendre des démarches auprès de fournisseurs ou de conclure des ententes de niveau de service ou des contrats.

Les responsables de la cybersécurité sont encouragés à prendre connaissance de ces conseils, compte tenu de l'utilisation croissante d'outils intégrant l'IA et de la dépendance opérationnelle accrue à l'égard de cette technologie.

Il convient également pour les propriétaires et exploitants d'infrastructures essentielles de consulter ces conseils afin de les aider à cerner les fonctions essentielles à leur mission, à comprendre leurs dépendances à l'IA et à déterminer les mesures à prendre afin d'assurer la résilience, la planification de la continuité des activités et la priorisation des ressources nécessaires en cas de dégradation ou d'indisponibilité des services d'IA.

Comment les systèmes d'IA utilisent les jetons

La consommation de jetons d'IA dépend de la manière dont les systèmes sont conçus et utilisés. En comprenant bien ces facteurs, les organisations peuvent gérer leur consommation plus efficacement.

La consommation de jetons et les coûts qui y sont associés augmentent dans les situations suivantes :

- des entrées et des sorties plus volumineuses, comme de longs textes ou des images, obligent l'outil d'IA à traiter et à générer davantage de contenu;
- de vastes ensembles de données de référence et les charges utiles liées aux connecteurs ne sont pas mis en cache ou optimisés;
- des renseignements, des données ou un historique de conversation supplémentaires sont inclus dans les requêtes;
- des flux de travaux comportant plusieurs étapes regroupent plusieurs actions ou tâches au sein d'un même processus;
- des requêtes répétées entraînent le traitement des mêmes données à plusieurs reprises;



- des systèmes intégrés récupèrent ou traitent des données provenant de plusieurs plateformes connectées;
- des flux de travaux agentiques entraînent une accumulation des coûts, car chaque étape nécessite de traiter de nouveau du contenu antérieur et des résultats produits par des outils.

À mesure que l'IA évolue et s'intègre davantage aux activités des organisations, la consommation de jetons augmentera rapidement à son tour. Les organisations devraient établir des mécanismes de surveillance et de gestion rigoureux de l'utilisation des jetons d'IA. Il conviendra également d'évaluer si un modèle hébergé par un fournisseur externe est réellement l'outil le mieux adapté à la tâche visée. Des modèles de plus petite taille, des modèles exécutés localement provenant de sources de confiance ou des outils d'IA spécialisés peuvent parfois produire des résultats comparables tout en consommant moins de ressources. Les organisations devront donc déterminer si un modèle exécuté localement ou un modèle offert sous forme de service répond le mieux à leurs besoins opérationnels.

Risques liés à l'utilisation des jetons d'IA

Si elle n'est pas gérée adéquatement, l'utilisation de l'IA peut entraîner des risques opérationnels, financiers et de sécurité. Voici quelques-uns des principaux risques auxquels les organisations peuvent être confrontées dans la gestion de leur consommation de jetons d'IA.

Hausse et imprévisibilité des coûts

La consommation de jetons d'IA peut augmenter rapidement selon la conception des systèmes et la façon dont ceux-ci sont utilisés. Les flux de travaux complexes, les requêtes répétées et les processus automatisés peuvent rendre les coûts difficiles à prévoir et à gérer au fil du temps.

Réduction de l'accès aux services essentiels

Sans mécanisme clair de priorisation, des activités à plus faible valeur ajoutée peuvent consommer des ressources nécessaires aux opérations importantes. En périodes de forte demande, cela peut entraîner des retards, réduire la performance ou interrompre des services essentiels.

À l'inverse, une priorisation qui privilégie uniquement les charges de travail essentielles peut retarder l'exécution d'activités moins prioritaires. Les organisations devraient donc équilibrer l'attribution des ressources de manière à assurer la disponibilité des services essentiels tout en maintenant un niveau d'accès raisonnable pour les autres utilisatrices et utilisateurs et pour les autres fonctions opérationnelles.

Visibilité et surveillance limitées

En l'absence d'outils de surveillance et de production de rapports, il peut être difficile de comprendre comment les jetons sont utilisés à l'échelle de l'organisation. Cette situation réduit la capacité de



repérer les cas d'inefficacité, de détecter les activités inhabituelles et d'appliquer des mesures de contrôle de l'utilisation. L'adoption d'outils de surveillance et de production de rapports adaptés aux besoins de l'organisation permettra à celle-ci de mieux comprendre la consommation de jetons.

Performance globale

Une consommation accrue de jetons peut nuire à la performance globale des systèmes. Par exemple, un outil d'IA qui traite de grands volumes de jetons peut répondre plus lentement, ce qui allonge le délai d'attente avant qu'une utilisatrice ou un utilisateur reçoive une réponse. Un nombre élevé de jetons exerce également une pression supplémentaire sur les ressources informatiques de l'organisation et exige davantage de mémoire. De plus, des requêtes très longues peuvent entraîner des résultats incohérents, des réponses contradictoires ou la production de contenu en double.

Épuisement des jetons

Les jetons risquent d'être épuisés lorsqu'une organisation atteint les limites de jetons qui lui ont été attribuées, ses quotas d'utilisation, ses limites de débit ou les seuils budgétaires établis. Cette situation peut découler d'une utilisation légitime à fort volume, d'applications mal optimisées ou d'une hausse imprévue de la demande.

Lorsque les jetons sont épuisés, les utilisatrices et utilisateurs ainsi que les systèmes peuvent subir des retards, une baisse de performance, des requêtes échouées ou une perte d'accès aux capacités d'IA. Les organisations qui dépendent de services intégrant l'IA pour soutenir des fonctions opérationnelles ou essentielles à leur mission peuvent être particulièrement touchées si aucune mesure d'urgence n'a été mise en place.

Les organisations peuvent configurer des alertes pour signaler un risque imminent d'épuisement des jetons. Ces alertes peuvent être déclenchées progressivement à mesure que la consommation approche des seuils définis par l'organisation. Par exemple, une alerte informative pourrait être générée lorsque 70 % des jetons alloués ont été consommés. À 80 %, l'alerte pourrait prendre la forme d'un avertissement. Enfin, à 90 %, l'alerte pourrait être considérée comme critique.

Attaques par épuisement des jetons et attaques par explosion des coûts

L'attaque par épuisement des jetons (DoT pour *Denial of Tokens*) est fondée sur la consommation excessive de ressources. Les auteurs de menace génèrent délibérément une utilisation excessive des services d'IA afin d'épuiser les allocations de jetons, de consommer les budgets et les ressources disponibles ou de déclencher les limites de débit. Ils peuvent y parvenir au moyen de requêtes automatisées, de l'exploitation abusive d'interfaces ou d'outils d'IA accessibles au public, ou encore de la manipulation de flux de travaux intégrant l'IA afin de traiter des volumes de données anormalement élevés. En épuisant les jetons disponibles ou en augmentant les coûts d'exploitation, une attaque par épuisement des jetons peut réduire la disponibilité des services, perturber les activités légitimes des



utilisatrices et utilisateurs d'une organisation et nuire à sa capacité d'accéder aux services d'IA lorsqu'elle en a besoin.

Les attaques par explosion des coûts (DoW pour *Denial of Wallet*) visent à porter atteinte aux finances d'une organisation. Les auteurs de menace exploitent délibérément les modèles de tarification fondés sur la consommation, l'utilisation ou le paiement à l'usage afin de faire augmenter les coûts. Voici quelques exemples de techniques utilisées dans le cadre d'attaques par explosion des coûts :

- soumettre aux outils d'IA des requêtes complexes nécessitant des analyses de données poussées et produisant de longues réponses;
- créer des boucles d'IA, par exemple en concevant des entrées qui amènent l'outil à réinjecter ses propres réponses dans l'outil.



Considérations pour les responsables de la cybersécurité

Les responsables de la cybersécurité ont l'occasion de tirer parti des outils d'IA pour contrer les cybermenaces. L'IA peut les aider à repérer les vulnérabilités plus rapidement et à plus grande échelle, ce qui permet d'accélérer les mesures correctives.

L'intelligence artificielle recèle d'innombrables possibilités pour les responsables de la cybersécurité. Lorsqu'elle est utilisée de façon sûre, sécurisée et responsable, l'IA peut :

- renforcer la priorisation des cyberrisques;
- améliorer la détection des menaces et des vulnérabilités;
- accélérer l'intervention et la reprise des activités;
- réduire le recours à des tâches manuelles répétitives.

Bien que l'IA offre des avantages potentiels en matière de cybersécurité, son utilisation comporte des coûts. Les tâches et les activités de cybersécurité feront augmenter l'utilisation globale ainsi que les dépenses connexes au sein des organisations. Pour gérer ces coûts, les responsables de la cybersécurité et les organisations devraient envisager l'utilisation de modèles de plus petite taille capables d'exécuter les tâches de cybersécurité moins complexes. Les organisations devraient également accorder la priorité aux tâches de cybersécurité plus complexes et déployer de façon stratégique les GML et les agents d'IA qui consomment plus de jetons pour répondre à ces besoins. Une approche reposant sur plusieurs modèles peut aider les organisations à réduire leur consommation de jetons et les coûts connexes.

Malgré les possibilités qu'offre l'IA, les responsables de la cybersécurité et les organisations doivent garder à l'esprit qu'elle ne remplace pas les principes fondamentaux rigoureux de la cybersécurité. La surveillance humaine, la gouvernance et les pratiques de sécurité intégrée dès la conception demeurent essentielles, conformément à la publication Secure by Design (en anglais seulement).



Considérations liées aux infrastructures essentielles

Les organisations du secteur des infrastructures essentielles (IE) évoluent dans des environnements à risque élevé, où les perturbations peuvent avoir des répercussions étendues. La gestion de la consommation de jetons d'IA devrait donc mettre l'accent sur la résilience, le maintien des opérations et la compréhension des systèmes et fonctions qui dépendent de l'IA.

Les organisations du secteur des IE devraient :

- comprendre comment les services d'IA soutiennent les opérations critiques et les services essentiels;
- cerner les dépendances entre les systèmes d'IA, les fournisseurs de services et les activités opérationnelles;
- comprendre l'incidence des différents modèles de déploiement de l'IA sur la disponibilité des services, la capacité opérationnelle et le rétablissement en cas d'interruptions;
- établir des plans en cas de réduction de la disponibilité des services d'IA attribuable à une panne, à un incident ou à une période de forte demande;
- tenir compte de la manière dont les situations d'urgence ou une hausse du niveau de menace pourraient accroître l'utilisation des services d'IA;
- intégrer l'utilisation de l'IA à la planification globale de la continuité des activités;
- maintenir des capacités non fondées sur l'IA, comme des processus d'automatisation numériques antérieurs ou une formation sur les opérations manuelles, au cas où les systèmes d'IA deviendraient indisponibles ou non fiables.

Pour obtenir d'autres conseils sur les IE, consultez les publications intitulées Résilience des infrastructures essentielles face à l'intensification des menaces et Bulletin conjoint sur l'isolation des systèmes essentiels. Ces publications présentent des recommandations additionnelles sur l'isolation des systèmes d'importance critique lors de cyberincidents majeurs ou dans des environnements à risque élevé prolongé.

Nous recommandons également aux organisations de consulter les conseils présentés dans le document d'orientation Principles for the Secure Integration of Artificial Intelligence in Operational Technology (en anglais seulement). Ce document souligne quatre principes clés que devraient suivre les propriétaires et exploitants pour tirer avantage de l'intégration de l'IA dans les systèmes de technologie opérationnelle (TO), tout en réduisant les risques. Il porte principalement sur l'apprentissage automatique, l'IA fondée sur les grands modèles de langage et les agents d'IA en raison des considérations et des défis de sécurité complexes qu'ils soulèvent. Les conseils s'appliquent également aux systèmes améliorés au moyen de modèles statistiques traditionnels ou de mécanismes d'automatisation fondés sur des règles logiques.



Gérer la consommation de jetons d'IA

Afin de mieux maîtriser les coûts liés à l'IA et de maintenir des activités fiables, les organismes ayant rédigé les documents d'orientation recommandent aux organisations d'adopter un modèle de gouvernance à plusieurs niveaux. Cette approche vise à prioriser l'utilisation des jetons en fonction des répercussions sur les activités de l'organisation, afin de protéger les fonctions essentielles tout en gérant adéquatement les activités moins prioritaires.

Prioriser l'attribution des jetons

Les organisations devraient envisager de répartir les allocations de jetons entre les niveaux ci-dessous. Cette approche peut contribuer à éviter que des activités courantes ou à fort volume consomment les ressources nécessaires aux fonctions essentielles à la mission. Bien que l'attribution de capacités dédiées puisse faire augmenter les coûts, elle peut améliorer la résilience et favoriser la continuité des activités.

Niveau 1 : Fonctions essentielles à la mission

Ce niveau comprend les fonctions hautement prioritaires, comme la cyberdéfense, l'intervention en cas d'incident et le renseignement sur les menaces. Ces services devraient bénéficier de la plus haute priorité en matière de disponibilité des jetons et de continuité des activités.

Niveau 2 : Fonctions opérationnelles

Ce niveau englobe les processus opérationnels essentiels, tels que l'administration des prestations, les activités de conformité et l'analyse des politiques. Il convient de surveiller et de gérer rigoureusement la consommation de jetons associée à ces services afin de soutenir la fiabilité des opérations et la maîtrise des coûts.

Niveau 3 : Productivité

Ce niveau comprend les activités quotidiennes des utilisatrices et utilisateurs, comme la rédaction de courriels, la synthèse de réunions et la recherche générale d'information. Ces activités de moindre priorité peuvent être assujetties à des limites d'utilisation ou à un accès réduit pendant les périodes de forte demande.

Concevoir des flux de travaux efficaces

Les fournisseurs de services d'IA appliquent souvent des tarifs différents selon les jetons d'entrée, les jetons de sortie, les jetons mis en cache et d'autres activités de traitement. Les organisations devraient donc concevoir leurs flux de travaux d'IA de manière à utiliser les ressources efficacement.



En pratique, il s'agit de fournir suffisamment de contexte pour accomplir une tâche tout en évitant les requêtes inutiles, les réponses excessivement longues ou le traitement répété des mêmes renseignements.

Améliorer la formulation des requêtes et les pratiques des utilisatrices et utilisateurs

Il convient de donner de la formation aux utilisatrices et utilisateurs afin de leur permettre d'interagir efficacement avec les systèmes d'IA. Des requêtes claires et précises peuvent contribuer à réduire la consommation inutile de jetons et à améliorer la qualité des résultats. Par exemple, il peut être possible de réduire les coûts en fournissant uniquement les extraits pertinents d'un document plutôt qu'en faisant traiter un fichier entier.

Lorsque c'est possible, les organisations devraient tirer parti des fonctionnalités qui favorisent une utilisation efficace et uniforme des services d'IA. Il peut s'agir notamment d'instructions réutilisables, de modèles et de configurations au niveau du système qui encouragent des réponses concises et réduisent la nécessité de fournir à répétition le même contexte.

Tirer avantage de la mise en cache et du contexte réutilisable

Certains services d'IA offrent des mécanismes de mise en cache qui permettent de réutiliser les résultats de contenus déjà traités, tels que des documents volumineux, des dépôts de code ou des instructions système, dans le cadre de requêtes ultérieures.

Lorsqu'elle est activée, la mise en cache peut réduire les besoins de traitement, améliorer la performance et diminuer les coûts lorsque les mêmes renseignements sont utilisés à plusieurs reprises.

Optimiser les données avant leur soumission

Les organisations peuvent également envisager d'utiliser des outils de prétraitement, d'optimisation ou de filtrage afin de réduire la quantité de données inutiles avant leur soumission à un système d'IA. Ces outils peuvent aider à éliminer le contenu redondant, à réduire la consommation de jetons et à améliorer l'efficacité globale.

Avant leur déploiement, tous les outils devraient être évalués conformément aux cadres de gestion des risques de l'organisation ainsi qu'aux exigences applicables en matière de sécurité, de protection de la vie privée et de protection des données.



Tirer efficacement parti de l'IA

Nous recommandons aux organisations de mettre en place une démarche d'encadrement de l'IA. Cette démarche consiste à concevoir l'infrastructure logicielle qui soutient les agents d'IA de façon à assurer l'exécution fiable et autonome des tâches et des requêtes.

Les organisations devraient intégrer l'IA à leurs structures de gouvernance, à leurs flux de travaux ainsi qu'aux procédures et processus décisionnels pertinents dans les cas où il est judicieux et avantageux de le faire, et lorsqu'il est possible de bien évaluer et sécuriser cette intégration. En recensant les cas d'utilisation où l'IA peut améliorer les activités, la performance des systèmes ou la résolution de problèmes, les organisations peuvent maximiser le rendement de leurs investissements dans les outils d'IA.

Les organismes ayant rédigé les documents d'orientation recommandent également aux organisations de mener des évaluations des risques avant d'autoriser le fonctionnement autonome d'agents d'IA dans leur environnement.

Optimiser la consommation de jetons d'IA

Les organisations devraient adopter une approche proactive et fondée sur les risques pour gérer la consommation de jetons d'IA dans l'ensemble de leurs systèmes et de leurs flux de travaux. Les mesures et pratiques recommandées ci-dessous peuvent les aider à gérer efficacement cette consommation.

Adapter le modèle à la tâche

Il n'est pas nécessaire d'utiliser le même modèle coûteux pour tous les flux de travaux. Les organisations peuvent réduire leurs coûts en choisissant un modèle adapté à la tâche à accomplir. Cette approche permet de réduire le coût par jeton. De plus, le réglage du niveau de raisonnement ou d'effort approprié peut contribuer à limiter la génération de jetons de sortie, qui sont souvent plus coûteux.

Comme l'intégralité du contexte est transmise lors de chaque conversation, les GML peuvent tirer parti de mécanismes de mise en cache peu coûteux pour réutiliser des résultats déjà calculés à partir d'un même texte d'entrée. Les organisations peuvent donc bénéficier du positionnement des renseignements contextuels statiques ou réutilisables au début de la définition d'une compétence ou des fichiers d'architecture connexes. Cette approche maximise l'efficacité de la mise en cache et contribue à l'optimisation globale des coûts. Les organisations devraient toutefois garder à l'esprit que chaque GML possède ses propres méthodes et exigences en matière de mise en cache. Il faudra peut-être faire quelques essais et ajustements.



Gouvernance et surveillance

- Mettre sur pied un groupe de travail sur la gouvernance de l'IA réunissant des parties prenantes des secteurs de la sécurité des technologies de l'information, des finances, des services juridiques et des opérations.
- Réaliser une évaluation des charges de travail liées à l'IA afin de déterminer leur importance opérationnelle et les risques connexes.
- Définir des procédures de reprise après sinistre et des guides opérationnels qui prévoient une supervision humaine aux principaux points de décision durant les pannes ou lors de comportements inhabituels des systèmes.
- Exiger une autorisation humaine avant l'exécution d'actions importantes propulsées par l'IA, conformément aux évaluations des risques, aux politiques, aux procédures et aux mesures de protection de l'organisation.

Résilience et gestion opérationnelle

- Adopter une architecture à fournisseurs multiples afin de diversifier les fournisseurs et de réduire la dépendance envers un seul d'entre eux.
- Choisir le modèle le moins coûteux qui répond aux exigences opérationnelles.
- Rediriger les charges de travail liées à l'IA vers des fournisseurs disponibles en cas de panne ou d'interruption de service afin d'accroître la résilience.
- Déployer des tableaux de bord pour surveiller la consommation de jetons et établir des seuils d'alerte visant les personnes responsables.
- Reporter les cas d'utilisation liés à la productivité qui sont moins prioritaires jusqu'à ce que les charges de travail essentielles à la mission aient été mises à l'essai et jugées capables de répondre à la demande prévue.
- Réutiliser des modèles approuvés pour les tâches courantes.
- Reconnaître les situations où les outils traditionnels, les modèles de plus petite taille, les modèles exécutés localement provenant de sources de confiance ou les processus manuels sont plus efficaces que les GML.

Sensibilisation organisationnelle et utilisation responsable

- Offrir de la formation sur les pratiques de rédaction de requêtes efficaces et responsables, notamment :
 - l'activation de la mise en cache des requêtes;
 - l'élimination de la régénération inutile de réponses et du traitement répété du même contenu;
 - la fourniture d'un contexte pertinent;
 - l'imposition d'une limite à la longueur des réponses;
 - la liaison des boucles agentiques;
 - la limitation des requêtes aux renseignements nécessaires à la tâche, par exemple en fournissant des extraits plutôt qu'un document complet.



- Sensibiliser le personnel aux politiques organisationnelles sur l'utilisation de l'IA, aux processus d'approbation et aux exigences en matière d'utilisation acceptable.
- Donner de la formation au personnel afin de lui permettre d'utiliser des outils d'IA et de reconnaître les situations où il convient d'utiliser l'IA, comparativement à celles où les outils traditionnels ou les processus manuels sont plus efficaces.
- Sensibiliser le personnel aux effets qu'une utilisation excessive ou inutile de l'IA peut avoir sur la capacité opérationnelle, la disponibilité des services et les coûts.
 - Comprendre comment un contexte excessif, des réponses très longues et des flux de travaux comportant plusieurs étapes peuvent accroître la consommation de jetons et les coûts.

Surveillance et amélioration continue

- Établir des niveaux de référence et des pratiques permettant de faire des prévisions afin de soutenir la budgétisation et la planification des capacités.
- Mettre en place des alertes d'utilisation permettant de repérer une consommation inhabituelle de jetons ou l'atteinte imminente des limites établies.
- Surveiller les hausses inhabituelles de la consommation de jetons, car il peut s'agir d'un indicateur potentiel d'exfiltration de données, de la même manière qu'une augmentation anormale de la bande passante peut signaler un transfert de données non autorisé.
- Intégrer la télémétrie liée aux jetons d'IA à une plateforme de gestion des informations et des événements de sécurité (GIES) et la corrélérer avec les journaux d'identité, d'infonuagique, d'API et de systèmes.
- Examiner régulièrement les tendances d'utilisation de l'IA et les profils de consommation de jetons afin de repérer les inefficacités ou les augmentations de coûts inattendues.
- Envisager l'utilisation de passerelles, de serveurs mandataires ou de solutions centralisées de surveillance d'IA afin d'améliorer la visibilité sur l'utilisation de l'IA, la consommation de jetons, les coûts, le respect des politiques et les tendances opérationnelles dans l'ensemble des services d'IA.
 - Les serveurs mandataires d'IA constituent une solution de plus en plus pratique pour surveiller et mesurer l'utilisation et la consommation, appliquer des quotas, imposer des mesures de protection et faciliter l'imputation des coûts.
- Évaluer les fournisseurs tiers de services d'IA en fonction de la fiabilité de leurs services, des limites de quotas et des risques opérationnels liés à la disponibilité des jetons.
- Mettre à jour les orientations organisationnelles, les contrôles et les politiques d'utilisation à mesure que l'adoption de l'IA et les besoins en jetons évoluent.

Pour en savoir plus

- [Les 10 mesures de sécurité en matière d'intelligence artificielle : Introduction \(ITSAP.10.049\)](#)
- [L'intelligence artificielle de pointe \(ITSAP.10.050\)](#)
- [Bulletin conjoint sur l'adoption prudente des services d'intelligence artificielle agentive](#)



- [Conseils sur l'IA de l'Australian Cyber Security Centre \(en anglais seulement\)](#)
- [Conseils sur l'IA de la Cybersecurity and Infrastructure Security Agency \(en anglais seulement\)](#)
- [Conseils sur l'IA du National Cyber Security Centre de la Nouvelle-Zélande \(en anglais seulement\)](#)
- [Conseils sur l'IA du National Cyber Security Centre du Royaume-Uni \(en anglais seulement\)](#)