LE POINT SUR...

Le coût de l'intelligence artificielle

Modes de facturation, leviers d'économie et pilotage

Les tarifs unitaires de certains modèles d'IA ont diminué, mais les dépenses des entreprises peuvent augmenter avec les volumes, les fonctionnalités et l'automatisation. Le coût dépend donc moins du prix affiché que de la manière dont l'outil est utilisé. Deux entreprises ayant le même besoin peuvent présenter des factures très différentes.

Comment l'IA est-elle facturée ?

Plusieurs modèles coexistent : abonnement par utilisateur, facturation à l'usage et formules hybrides. L'abonnement apporte une meilleure visibilité budgétaire, mais peut comporter des limites d'usage ou des coûts complémentaires. La facturation à l'usage est fréquente pour les traitements automatisés raccordés aux logiciels de l'entreprise.

Le token est une unité de traitement du texte. Un document de cinquante pages peut représenter plusieurs dizaines de milliers de tokens, mais le volume réel dépend de la langue, de la densité du document, de sa mise en forme et du modèle. Dans de nombreuses API, les tokens d'entrée et de sortie sont facturés séparément, les seconds étant souvent plus chers.

Ce qui peut entrer dans la facture

- Les instructions permanentes peuvent être incluses ou réinjectées à chaque appel, selon le service et son architecture.
- Les documents joints peuvent être transmis intégralement, découpés ou récupérés à la demande ; le traitement dépend du service et d'une éventuelle mise en cache.
- L'historique de la conversation peut être repris, résumé, tronqué ou mis en cache ; il ne faut pas supposer qu'il est toujours renvoyé intégralement.
- La question posée est souvent la plus petite partie du volume, mais cela dépend du contexte fourni.
- La réponse produite est généralement facturée séparément et peut coûter plus cher lorsqu'elle est longue.
- La réflexion interne peut être comptée comme des tokens de sortie par certains modèles ; sa facturation dépend du fournisseur.

Le coût ne vient donc pas seulement de la question : il peut inclure le contexte, la réponse et les outils mobilisés.

Les leviers qui pèsent sur la facture

Levier	Effet sur la facture
Choisir le modèle selon la profondeur réelle de la tâche	Selon les modèles, l'écart de prix entre gammes peut être important : comparer sur une tâche type
Réserver les modes « raisonnement » aux véritables analyses	Le raisonnement peut augmenter le volume de tokens et le coût ; le réserver aux tâches qui le justifient
Éviter de faire relire le même document à chaque demande (la « mise en cache »)	Le cache peut réduire le coût des contextes répétitifs, mais les tarifs, durées et conditions varient selon le fournisseur
Différer ce qui n'est pas urgent	Certains Batch API pour tâches non urgentes proposent une réduction importante, parfois de 50 %, si la requête est éligible
Ouvrir une conversation par tâche	Une conversation longue peut augmenter le contexte traité, mais le service peut aussi appliquer résumé, troncature ou mise en cache

© ATH. Tous droits réservés.

Des instructions précises peuvent réduire les allers-retours et donc le coût total. Il n'est pas nécessaire d'utiliser des techniques savantes : préciser le contexte, le format et le niveau de détail suffisent souvent.

Un ordre de grandeur

Exemple théorique : si un rapport de 35 000 tokens (environ 50 pages) est transmis dix fois, le volume d'entrée atteint 350 000 tokens hors mécanisme de réutilisation.

Avec une mise en cache effectivement disponible, le coût peut être fortement réduit. Le gain réel dépend du fournisseur, du modèle, de la durée du cache et des éventuels coûts d'écriture ou de stockage.

Concrètement pour l'entreprise

Faire l'état des lieux

- Inventorier les outils déjà payés : comptes actifs, formule, crédits et moyen de paiement.
- Supprimer les comptes dormants. C'est l'économie la plus rapide et elle ne suppose aucun arbitrage.

* * * * * *

Choisir le bon niveau de modèle

Exemples	Niveau de modèle adapté
Tâches simples	
Reformater, traduire, classer, extraire des informations d'un document	Modèle léger, si la qualité attendue est suffisante
Rédaction courante	
Courriers, synthèses, comptes rendus, réponses clients	Modèle intermédiaire, à valider sur des cas représentatifs
Analyse complexe	
Comparer des documents longs, raisonner en plusieurs étapes, arbitrer	Modèle avancé, mode raisonnement si la tâche le justifie

Rédiger des instructions économes

- Donner le contexte indispensable plutôt que de le laisser deviner.
- Préciser le format et la longueur attendus : une sortie courte peut réduire le coût de génération, sans garantir une baisse si le contexte domine.
- Joindre un exemple du résultat souhaité lorsque c'est possible : cela peut limiter les allers-retours.
- Figer les consignes récurrentes peut favoriser la réutilisation du contexte, si le service propose une mise en cache compatible.

Piloter dans la durée

- **Poser des plafonds et des alertes** sur chaque accès facturé à l'usage, avant l'ouverture du premier traitement automatisé.
- **Encadrer les traitements qui tournent seuls : nombre d'étapes limité, budget maximal, arrêt automatique et gestion des erreurs.**
- **Interroger le prestataire sur la mise en cache : fonctionnement automatique ou configurable, éléments concernés, durée, coûts d'écriture et visibilité dans le suivi.**
- **Suivre le coût complet par tâche réellement terminée, temps humain, stockage et contrôle compris.**
- **Désigner un référent. Le FinOps désigne une pratique de pilotage des coûts et de la valeur des services numériques, applicable aussi à l'IA. Un poste dédié n'est pas nécessaire au départ : une personne rapprochant chaque mois la facture des usages suffit.**
- **Réexaminer régulièrement les contrats, les usages et les tarifs, qui évoluent rapidement.**

© ATH. Tous droits réservés.

Les points de vigilance

- Les tarifs publics sont souvent exprimés au million de tokens et en dollars : intégrer le change, les taxes, les engagements et le coût d'une tâche type.
- Les traitements automatisés qui bouclent ou réessaient sont un risque majeur ; les usages individuels peuvent aussi augmenter la facture.
- Les abonnements souscrits individuellement peuvent créer un risque de confidentialité : avant d'y transmettre des données clients ou personnelles, vérifier conservation, réutilisation, accès, sous-traitants et transferts hors UE.
- Une facture qui augmente pendant qu'un traitement devient beaucoup plus rapide peut rester favorable : le coût de l'IA doit être comparé au temps et aux coûts évités.
- Les formules présentées comme illimitées peuvent comporter des limites de modèle, de débit ou de volume, qui ralentissent ou bloquent le service une fois atteintes.

À savoir

Tant que la facture n'est pas rapprochée des usages, aucun arbitrage n'est possible : il faut connaître les volumes, les modèles et les tâches réellement achevées. Le budget doit être construit sur des cas d'usage et des volumes, puis révisé lorsque les tarifs ou les conditions changent.

Les tarifs et les conditions d'utilisation évoluent rapidement. Un budget d'IA se construit donc sur des volumes, des cas d'usage et des hypothèses révisables, plutôt que sur un prix unitaire figé.

À retenir

1. Plusieurs modèles coexistent : abonnement, usage et formules hybrides. Le coût dépend du contexte, de la réponse, des outils et du niveau d'automatisation.
2. Trois leviers sont essentiels : le modèle adapté à la tâche, la réutilisation du contexte lorsque le service le permet, et le traitement par lots des tâches non urgentes.
3. Des instructions précises peuvent réduire les allers-retours et le coût total.
4. Les traitements automatisés doivent être encadrés dès le départ : plafond de budget, limite d'étapes et arrêt automatique.
5. Le bon indicateur est le coût complet par tâche terminée, temps humain, contrôle et stockage compris.

Pour aller plus loin

Pages tarifaires et conditions contractuelles officielles des fournisseurs, vérifiées à la date de diffusion.

Documentation officielle sur la mise en cache, le traitement par lots, les limites d'usage et la protection des données.

© ATH. Tous droits réservés.