



INGÉNIERIE ÉCONOME EN RESSOURCES · LIVRE BLANC

Le coût de fonctionnement de l'IA

Le maintenir bas et prévisible après le projet pilote



Rebecca Enonchong, FREng · Fondatrice et CEO, AppsTech Labs

Ebot Tabi · Directeur de la recherche, AppsTech Labs

Septembre 2026

AppsTech Labs

Breakthrough Technology. Disciplined Engineering.

appstechlabs.com

INTRODUCTION

Le coût de fonctionnement après le projet pilote

Une entreprise qui adopte l'intelligence artificielle (IA) commence en général par un projet pilote, c'est-à-dire un essai sur un petit volume de travail. La construction d'un pilote peut représenter une dépense réelle, en particulier pour une petite ou moyenne entreprise. Son coût de fonctionnement, la facture mensuelle d'utilisation du modèle d'IA, reste en revanche généralement faible, car le pilote traite quelques requêtes par jour. Une fois le système en service dans toute l'entreprise, il peut lire chaque facture fournisseur et répondre à chaque message client. La qualité des réponses peut rester stable tandis que la facture mensuelle augmente à chaque nouvel usage.

Ce livre blanc explique d'où vient le coût de fonctionnement d'un système d'IA et comment le maintenir bas.

AppsTech Labs appelle son approche l'ingénierie économe en ressources. Pour chaque système, nos ingénieurs mesurent ce que coûte le traitement d'une transaction, par exemple une facture.

LE DIAGNOSTIC

Quatre causes de la hausse des coûts de l'IA

Facturation au token. Un modèle d'IA est le programme qui lit une requête et rédige la réponse. Les modèles qui font fonctionner des outils comme ChatGPT, Claude ou Gemini sont des grands modèles de langage, ou LLM. Les entreprises louent en général l'accès à un modèle d'IA auprès d'un fournisseur comme Anthropic, Google ou OpenAI, qui facture au token (jeton). Un token est un fragment de texte, souvent un mot court ou une partie d'un mot plus long. Le fournisseur du modèle d'IA compte chaque token envoyé au modèle et chaque token que le modèle renvoie, à chaque requête. Selon les grilles tarifaires actuelles des trois fournisseurs, un token rédigé par le modèle coûte entre cinq et environ huit fois plus cher qu'un token lu^{[1][2][3]}.

Croissance de l'usage. Un pilote commence en général dans une seule équipe, par exemple la comptabilité fournisseurs. Lorsqu'il fonctionne, d'autres équipes le demandent et de nouvelles tâches s'ajoutent, comme la lecture de contrats ou la réponse aux courriels des clients. Chaque nouvel usage ajoute des requêtes, et le fournisseur du modèle d'IA facture chacune d'elles. Contrairement à une licence logicielle, dont le prix annuel est forfaitaire, la facturation à l'usage suit le volume. Dix fois plus de requêtes donnent une facture dix fois plus élevée, et cette hausse atteint le budget sans passer par une approbation d'achat.

Requêtes surdimensionnées. Une requête type comporte des instructions permanentes, plusieurs exemples, des documents entiers et l'historique de la conversation. L'entreprise paie chacun de ces éléments à chaque requête. Des requêtes plus longues réduisent aussi la fiabilité des réponses. Une étude de juillet 2025 portant sur 18 modèles de premier plan a montré que leurs performances baissaient à mesure que le texte fourni s'allongeait^[4].

Coût non mesuré. Une enquête Deloitte a interrogé 200 directeurs financiers nord-américains à la mi-2026. Pour 46 % d'entre eux, l'incertitude ou le manque de transparence sur les coûts constitue la principale préoccupation interne liée à l'usage de l'IA dans l'entreprise^[5]. Dans la même enquête, 43 % évoquent une visibilité insuffisante sur les outils d'IA ou leur usage^[5]. Sans chiffre sur le coût d'une transaction, la facture mensuelle du fournisseur du modèle d'IA est le premier signe d'une hausse des coûts.

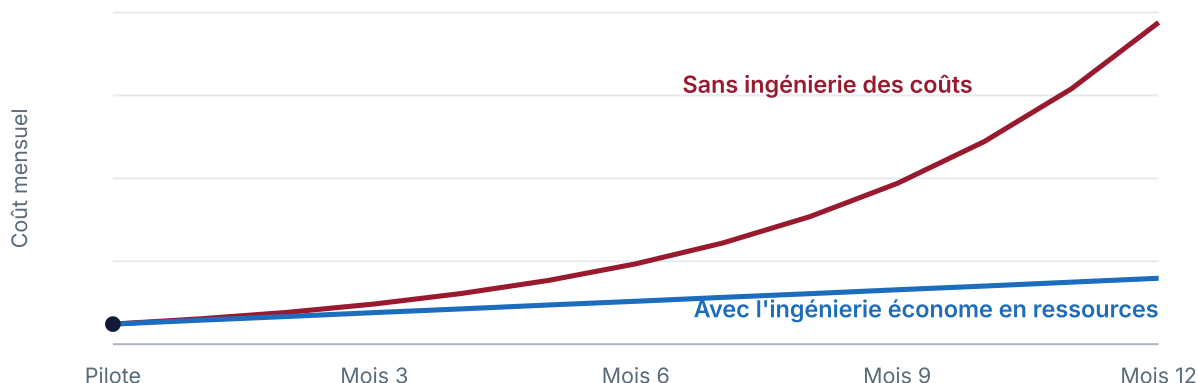


Figure 1. Coût mensuel sur douze mois de deux systèmes au coût identique pendant le pilote. Illustration.

Cloudflare indique qu'un modèle peut accéder à l'ensemble de ses services avec une description d'environ 1000 tokens, contre 1,17 million par la méthode habituelle^[6].

DISCIPLINE · 1

Mise en cache du prompt et taille des requêtes

Les économies les plus importantes portent sur le contenu de chaque requête, quel que soit le modèle choisi.

Les fournisseurs de modèles d'IA peuvent conserver la partie fixe d'une requête, comme les instructions et les exemples, et la réutiliser pendant quelques minutes ou jusqu'à une heure. Ce mécanisme s'appelle la mise en cache du prompt (prompt caching). Le prompt est le texte complet envoyé au modèle avec une requête. Anthropic, Google et OpenAI facturent tous le texte conservé à environ un dixième du prix normal en entrée sur leurs principaux modèles^{[1][2][3]}. La réduction s'applique lorsque la partie fixe vient en premier et le document en dernier. Nos ingénieurs traitent donc l'ordre d'un prompt comme une décision de coût.

Une deuxième source d'économies consiste à tenir les données volumineuses à l'écart du modèle. Un système peut extraire dix mille lignes d'une base de données pour répondre à une seule question. Dans la conception coûteuse, le modèle reçoit les dix mille lignes. Dans la conception économe, un court programme filtre d'abord les lignes et transmet le résultat au modèle. Dans un exemple détaillé publié en novembre 2025, Anthropic indique que ce changement a ramené une tâche de 150 000 tokens à 2 000, soit

une économie de 98,7 %^[7]. Cloudflare a observé un effet comparable en février 2026 pour la description des services qu'un modèle peut utiliser. Donner à un modèle l'accès à l'ensemble des services de Cloudflare de cette manière a demandé environ 1000 tokens de description, contre 1,17 million par la méthode habituelle^[6].

DISCIPLINE · 2

Taille et prix du modèle

Aux prix de septembre 2026, le modèle d'IA le plus performant de la grille tarifaire de Google coûte environ sept fois plus cher par token que le plus petit. L'écart est de dix fois chez Anthropic et de cent fois chez OpenAI^{[1][2][3]}. La bonne taille dépend de la tâche, et nos ingénieurs la déterminent par un test.

Le test est rédigé en premier, sous la forme d'un ensemble de cas réels tirés du travail du client, y compris les plus difficiles, chacun accompagné de sa réponse correcte. Les ingénieurs essaient le plus petit modèle susceptible de réussir, et passent à un modèle plus grand lorsque ce premier modèle échoue.

Pour le tri à grand volume, comme le classement de documents ou l'orientation de messages vers la bonne équipe, un petit modèle entraîné pour cette seule tâche donne souvent le meilleur résultat. Une comparaison indépendante publiée en janvier 2026 a porté sur 32 expériences de ce type. Un petit modèle entraîné pour la tâche a traité 277 éléments par seconde, contre 12 pour un petit modèle généraliste qui rédige ses réponses. Les modèles entraînés ont aussi été plus précis sur trois des quatre tâches testées^[8].

DISCIPLINE · 3

Traitement par lots et auto-hébergement

Une partie du travail d'IA peut attendre quelques heures son résultat. Le rapprochement des comptes pendant la nuit en est un exemple, et un stock de documents numérisés en est un autre. Anthropic, Google et OpenAI traitent ce type de travail à moitié prix grâce à leurs services de traitement par lots (batch), qui renvoient les résultats dans un délai d'un jour^{[1][2][3]}.

Pour les modèles qu'AppsTech Labs exécute sur des serveurs qu'elle gère, la même logique s'applique au matériel. Lorsqu'un serveur traite de nombreuses requêtes ensemble, l'énergie consommée par token peut être divisée par trois à cinq, selon des mesures portant sur 46 modèles publiées par ML.ENERGY en janvier 2026^[9]. Un modèle installé sur les propres serveurs du client n'entraîne pas de frais par question, et les données du client restent à l'intérieur de son réseau.

DISCIPLINE · 4

Compression des modèles

La compression réduit la taille d'un modèle afin qu'il fonctionne plus vite sur moins de matériel. Dans la distillation, un grand modèle entraîne un modèle plus petit à reproduire ses réponses. Dans la quantification, les nombres internes du modèle sont stockés avec une précision moindre.

Ces méthodes restent utiles pour les modèles que nous exécutons nous-mêmes. Leur gain réel dépend toutefois du nombre de requêtes que le serveur traite en même temps, et il faut donc le mesurer dans les conditions d'utilisation réelles. L'étude de janvier 2026 le montre avec la quantification. Elle a comparé les mêmes modèles dans deux versions, l'une stockant ses nombres sur 16 bits, le format habituel, l'autre sur 8 bits. La version sur 8 bits devrait consommer moins d'énergie. Pourtant, avec 8 à 16 requêtes traitées en même temps, elle a consommé jusqu'à 56 % d'énergie de plus que la version sur 16 bits. Elle n'a permis d'économiser de l'énergie qu'au-delà d'environ 64 requêtes simultanées, soit 11 % d'économie dans le cas typique^[9].

DISCIPLINE · 5

Coût par transaction d'IA

L'unité qui compte pour l'entreprise est le coût d'une transaction, par exemple une facture lue ou un message client traité. Nos ingénieurs le mesurent en faisant passer cent éléments réels dans le système terminé, puis en divisant la facture par cent. Ils refont la mesure chaque fois que les instructions changent de manière significative.

EXEMPLE CHIFFRÉ · FACTURES FOURNISSEURS

Pour chaque facture, le système envoie au modèle environ 3 000 tokens, composés des instructions, de deux exemples et du texte de la page. Le modèle renvoie environ 400 tokens, avec les principales informations de la facture présentées dans des champs fixes. Sur Claude Sonnet 5 ou GPT-6 Sol, qui affichaient le même prix catalogue en septembre 2026, une facture coûte 0,01 dollar^{[1][3]}.

Lorsque les instructions et les exemples sont conservés pour être réutilisés, environ 2 200 des 3 000 tokens sont facturés au dixième du prix, et la facture revient à environ 0,006 dollar. Lorsque les factures peuvent attendre un traitement de nuit, le tarif par lots divise ce montant par deux, soit environ 0,003 dollar.

Coût mensuel pour 20 000 factures	
Traitement standard	200 \$
Avec mise en cache du prompt	121 \$
Avec mise en cache et traitement par lots de nuit	60 \$

Les trois lignes utilisent le même modèle et produisent les mêmes réponses.

Calcul effectué à partir des prix publiés par Anthropic et OpenAI en septembre 2026, en dollars américains. Le résultat est identique pour les deux. Volumes donnés à titre d'illustration.

Cinq questions sur le coût de fonctionnement de l'IA

QUESTIONS POUR TOUTE PROPOSITION D'IA, AVANT SIGNATURE PUIS CHAQUE TRIMESTRE

1. Combien coûte une transaction, et qui la mesure ?
2. Comment évolue le coût mensuel lorsque l'usage double ?
3. Le modèle proposé est-il le plus petit qui réussit un test construit sur les propres cas de l'entreprise ?
4. Qu'envoie chaque requête au modèle, et pourquoi ?
5. Combien coûtera le système au douzième mois, par rapport au chiffre du premier mois indiqué dans la proposition ?

Une réponse utile à chaque question est un chiffre ou une décision de conception précise, donnée par écrit afin que la direction financière puisse la vérifier.

LA POSITION

L'ingénierie économe en ressources chez AppsTech Labs

Nos ingénieurs appliquent les cinq disciplines présentées dans ce livre blanc. Chaque requête place les instructions fixes en premier afin qu'elles puissent être mises en cache, et les données volumineuses sont filtrées avant d'atteindre le modèle. Avant de choisir un modèle, nous rédigeons un test à partir des propres cas du client et retenons le plus petit modèle qui le réussit. Le travail qui peut attendre est traité par lots pendant la nuit. Lorsque les données d'un client doivent rester à l'intérieur de son réseau, le système fonctionne sur les propres serveurs du client. Pour les modèles que nous exécutons nous-mêmes, nous utilisons la compression lorsque des mesures sur la charge de travail réelle montrent une économie. Le coût d'une transaction est mesuré pour chaque système, puis mesuré de nouveau chaque fois que ses instructions changent.

appstechlabs.com

Breakthrough Technology. Disciplined Engineering.

À PROPOS DES AUTEURS

Rebecca Enonchong est la fondatrice et CEO d'AppsTech Labs et d'AppsTech. Elle est membre (Fellow) de la Royal Academy of Engineering du Royaume-Uni.

Ebot Tabi est directeur de la recherche d'AppsTech Labs, où il dirige la recherche et le développement des systèmes d'intelligence artificielle de l'entreprise.

SOURCES

1. Anthropic, tarifs de Claude, y compris la mise en cache du prompt et le traitement par lots. platform.claude.com/docs/en/about-claude/pricing. Consulté le 27 septembre 2026.
2. Google, tarifs de l'API Gemini, y compris la mise en cache du contexte et le mode Batch. ai.google.dev/gemini-api/docs/pricing. Consulté le 27 septembre 2026.
3. OpenAI, tarifs de l'API, y compris les entrées en cache et le niveau Batch. developers.openai.com/api/docs/pricing. Consulté le 27 septembre 2026.
4. Chroma, « Context Rot: How Increasing Input Tokens Impacts LLM Performance », 14 juillet 2025. trychroma.com/research/context-rot
5. Deloitte, « Q2 2026 CFO Signals Survey », 200 directeurs financiers nord-américains d'entreprises réalisant au moins 1 milliard de dollars de chiffre d'affaires, enquête menée du 22 mai au 7 juin 2026. deloitte.com/us/en/about/press-room/deloitte-q2-2026-cfo-signals-survey.html
6. M. Carey, Cloudflare, « Code Mode: give agents an entire API in 1,000 tokens », 20 février 2026. blog.cloudflare.com/code-mode-mcp
7. Anthropic, « Code execution with MCP: building more efficient AI agents », 4 novembre 2025. anthropic.com/engineering/code-execution-with-mcp
8. A. Jacobs, « Beating BERT? Small LLMs vs Fine-Tuned Encoders for Classification », 4 janvier 2026. alex-jacobs.com/posts/beatingbert
9. ML.ENERGY, « Diagnosing Inference Energy Consumption with the ML.ENERGY Leaderboard v3.0 », 29 janvier 2026. ml.energy/blog