Oliimpe LEGAL

LIVRE BLANC · OCTOBRE 2026

La tokenisation de la société

ou pourquoi vos coûts d'IA vont compter

Le prix d'un token s'effondre depuis trois ans. Les factures, elles, montent. Ce paradoxe n'est pas une anomalie comptable : c'est le signe qu'une nouvelle unité de compte s'installe dans l'économie du travail intellectuel. **Ce document explique le mécanisme, en chiffre l'ampleur, et expose le pari qu'Oliimpe fait dessus.**

Pauline Lahary, co-fondatrice et CEO d'Oliimpe · 5 pages · lecture 12 minutes · sources en fin de document

I. Le token est devenu une unité de compte

Un *token* n'est pas un mot. C'est un fragment de texte produit par un algorithme de segmentation — le plus souvent un *byte-pair encoding* — qui découpe une phrase en unités statistiquement fréquentes. En anglais, un token vaut environ quatre caractères. En français, davantage de tokens pour le même sens, parce que les vocabulaires ont été construits sur des corpus majoritairement anglophones. Sur un texte juridique français, truffé d'abréviations, de numéros d'articles et de formules consacrées, l'écart se creuse encore. **La même question coûte structurellement plus cher en français qu'en anglais.** Peu de directions financières le savent.

Ce détail technique a une conséquence économique directe : les modèles de langage ne se facturent ni à la requête, ni à l'utilisateur, ni au temps passé. Ils se facturent au token, avec deux compteurs distincts — les tokens *entrants*, ceux que vous donnez à lire au modèle, et les tokens *sortants*, ceux qu'il rédige. Les seconds coûtent généralement trois à cinq fois le prix des premiers. Un même travail peut donc coûter du simple au triple selon qu'on demande une réponse courte ou un rapport.

S'ajoute une propriété que l'on oublie souvent : le coût de calcul du mécanisme d'attention croît de manière quadratique avec la longueur du contexte. Doubler la taille du dossier que l'on colle dans la conversation ne double pas la charge, il la quadruple. Les fournisseurs amortissent cela par du *prompt caching* — la mise en cache des portions de contexte réutilisées — mais le principe demeure : **ce que vous donnez à lire coûte, et ce que vous donnez à lire inutilement coûte pour rien.**

Un ordre de grandeur, pour fixer les idées. Un acte de trente pages représente environ trente mille tokens en français. Le donner à lire à un modèle dix fois dans la journée — une fois par collaborateur, une fois par question — c'est trois cent mille tokens d'entrée avant la moindre réponse. À ce stade, ce n'est pas encore cher. Ce qui coûte, c'est la répétition de ce geste par toute une profession, tous les jours, et ce que le modèle écrit pour lui-même avant de vous répondre — nous y revenons au chapitre III.

> **Trois unités à ne pas confondre**
>
> **Le prix du token** : ce que facture le fournisseur par million d'unités. Il s'effondre. **La consommation** : le nombre de tokens brûlés pour accomplir une tâche. Elle explose. **Le coût par requête** : le produit des deux. C'est le seul chiffre qui atterrit sur votre relevé bancaire, et c'est le seul que presque personne ne mesure.

II. Le prix s'effondre. La facture monte.

Commençons par la bonne nouvelle, qui est spectaculaire. Le *AI Index* de Stanford HAI mesure depuis trois ans le coût d'obtention d'un niveau de performance donné. Pour atteindre la qualité d'un GPT-3.5, il fallait débourser environ **20 dollars par million de tokens en novembre 2022**. En octobre 2024, **7 centimes**[1]. Un facteur 280 en deux ans. Epoch AI, qui suit la même grandeur sur un éventail de tâches, situe la baisse annuelle dans une fourchette large — de 9 à 900 fois par an selon le seuil de performance retenu[2]. Sam Altman, lui, retient un ordre de grandeur simple : un facteur dix par an.

> « Il semble bien que nous soyons sur le point de tenir la promesse d'une intelligence trop bon marché pour être comptée. Nous avons divisé par plus de dix le coût de chaque unité d'intelligence, chaque année, depuis cinq ans. »
>
> **Sam Altman**, conférence de la Réserve fédérale américaine, juillet 2025[3]

Maintenant, la facture. Google traitait environ **480 000 milliards de tokens par mois** en mai 2025. En octobre de la même année, Sundar Pichai annonçait **1,3 million de milliards** — 1,3 quadrillion[4]. Du côté des entreprises, Menlo Ventures chiffre la dépense mondiale en API de modèles de langage à 3,5 milliards de dollars fin 2024, puis **8,4 milliards six mois plus tard**[5]. Le prix unitaire divisé par dix, la dépense multipliée par deux et demi, sur la même période.

÷ 280
le prix d'une performance GPT-3.5, entre nov. 2022 et oct. 2024 *(Stanford HAI)*

× 2,7
les tokens traités par Google entre mai et octobre 2025

93 %
des organisations ont dépassé leur budget IA *(McKinsey, 2026)*

Ce n'est pas un accident : c'est un mécanisme connu depuis un siècle et demi. En 1865, l'économiste britannique William Stanley Jevons observe dans *The Coal Question* que l'amélioration du rendement des machines à vapeur n'a pas réduit la consommation de charbon du Royaume-Uni — elle l'a augmentée, parce que le charbon devenu moins cher a ouvert des usages qui n'existaient pas. Le 26 janvier 2025, au plus fort de la panique boursière provoquée par DeepSeek, le patron de Microsoft a résumé l'affaire en trois mots.

> « Le paradoxe de Jevons frappe encore. »
>
> **Satya Nadella**, PDG de Microsoft, 26 janvier 2025[6]

Les chiffres de gestion confirment que le phénomène a quitté les laboratoires pour atterrir dans les services financiers. La *FinOps Foundation* relevait en 2026 que la part des équipes FinOps chargées de piloter une dépense d'IA était passée de 31 % à **98 % en deux ans**, et que **73 % des organisations** avaient dépensé plus que prévu[7]. Une enquête McKinsey de mai 2026 monte à 93 % de dépassements, en pointant trois causes : une consommation de tokens imprévisible, une tarification à l'usage, et une gouvernance immature[8].

L'imprévisibilité n'est pas un défaut d'organisation. Des travaux de Microsoft Research montrent que **les modèles de pointe sont incapables de prédire leur propre consommation** avant d'exécuter une tâche, et que deux exécutions d'une même tâche peuvent varier d'un facteur trente[9]. Aucun directeur financier n'accepterait cela d'un fournisseur d'énergie. Nous l'acceptons de l'intelligence artificielle parce que nous n'avons pas encore l'habitude de la compter.

III. Le vrai moteur : le raisonnement

Si la facture monte alors que le prix baisse, c'est que la consommation monte plus vite encore. Trois forces s'additionnent, et une seule est visible du client.

1. Les modèles de raisonnement écrivent pour eux-mêmes

Depuis 2024, les modèles les plus performants ne répondent plus directement : ils produisent d'abord une longue chaîne de réflexion interne, qui vous est facturée au prix fort des tokens sortants, et que vous ne lisez jamais. Une étude récente mesurant la part du coût de génération consacrée à cette réflexion interne trouve une **médiane de 94,7 %**, avec plusieurs modèles de pointe au-delà de 99 %[10]. Autrement dit : sur cent euros de génération, quatre-vingt-quinze paient une réflexion invisible. Et la longueur de ces chaînes croîtrait d'environ **cinq fois par an**[11].

2. Les agents rappellent le modèle en boucle

Une requête simple, c'est un aller-retour. Un enchaînement agentique — l'outil qui cherche, lit, recoupe, se corrige — c'est des dizaines d'appels, chacun rechargeant tout ou partie du contexte. Une estimation reprise par EY situe le coût d'une interaction orchestrée autour de **1,20 dollar en 2026, contre 0,04 dollar en 2023**[12] — un facteur trente, pendant que le prix du token était divisé par cent.

3. Le contexte enfle sans qu'on s'en aperçoive

C'est la dérive la plus silencieuse. On colle un dossier entier plutôt qu'un extrait. On active une fenêtre de contexte plus large « au cas où ». On laisse l'historique de conversation s'accumuler. Aucun changement d'infrastructure, aucune alerte, et la facture double. Les praticiens du FinOps ont identifié l'indicateur avancé : **le nombre moyen de tokens par requête** monte bien avant que le budget ne craque[13].

Ce qui baisse	Ce qui monte
Le prix du million de tokens	Le nombre de tokens par tâche
Le coût d'une performance donnée	La performance que l'on exige
Le coût d'un aller-retour	Le nombre d'allers-retours
Le coût du modèle d'hier	Le prix du modèle du jour

Ce tableau est le cœur du sujet. **Chaque ligne de gauche sert d'argument commercial ; chaque ligne de droite atterrit sur la facture.** Et la dernière ligne mérite une attention particulière, parce qu'elle est structurelle : un éditeur qui revend de l'intelligence artificielle voit sa marge directement indexée sur le prix des modèles. Le dernier modèle sorti coûtant toujours plus cher que le précédent, il a un intérêt économique à vous laisser sur l'ancien. Ce n'est pas une accusation, c'est une mécanique d'incitation.

> **Pourquoi le français juridique coûte plus cher**
>
> Les vocabulaires de tokenisation sont construits sur des corpus majoritairement anglophones. Un texte français consomme déjà davantage de tokens qu'un texte anglais de sens équivalent ; un texte juridique français, avec ses références d'articles, ses abréviations et ses formules, davantage encore. Pour une profession francophone, **la facture d'IA part avec un handicap structurel** que personne ne compense.

IV. Pourquoi cela devient un sujet de société

Tant qu'il s'agit de trois requêtes par jour, le token reste une curiosité d'ingénieur. Quand une profession entière s'y met, il devient autre chose : **une unité de mesure du travail intellectuel**. C'est cela que nous appelons la tokenisation de la société — le moment où réfléchir, chercher, rédiger cessent d'être des gestes dont le coût se dilue dans un salaire, pour devenir des opérations comptées à l'unité, facturées à l'acte, et arbitrables.

Quatre conséquences se dessinent déjà.

L'énergie devient la contrainte de fond

L'Agence internationale de l'énergie, dans son rapport *Energy and AI* d'avril 2025, chiffre la consommation électrique mondiale des centres de données à **415 TWh en 2024**, soit environ 1,5 % de l'électricité mondiale, et projette **945 TWh en 2030** — légèrement au-dessus de la consommation totale du Japon aujourd'hui[14]. La demande des centres optimisés pour l'IA ferait plus que quadrupler sur la période. C'est la raison pour laquelle Sam Altman, interrogé par des parlementaires américains, fait converger les deux sujets : « à terme, le coût de l'intelligence convergera vers le coût de l'énergie »[15]. Le token n'est pas une abstraction logicielle : c'est de l'électricité transformée en phrases.

Une dépendance nouvelle, et silencieuse

Un cabinet qui bâtit sa production sur un outil dont le coût est indexé sur le prix des modèles hérite d'une exposition qu'il n'a pas choisie. Il ne négocie pas avec son éditeur : il subit, avec un trimestre de décalage, les arbitrages d'un fournisseur d'infrastructure américain. **La question n'est plus « combien ça coûte » mais « de qui dépend ce que ça coûtera ».**

Une inégalité d'accès au dernier modèle

Si le coût d'une performance donnée s'effondre mais que le prix du modèle de pointe reste élevé, alors l'écart ne se joue plus sur l'accès à l'IA — tout le monde y a accès — mais sur l'accès à *la meilleure* IA, au moment où elle sort. Entre deux cabinets, la différence ne sera pas l'outil : ce sera la génération de modèle que chacun peut se payer, et la qualité des sources qu'il lui donne à lire.

Une question déontologique

Pour un notaire, un avocat, un expert-comptable, la maîtrise du coût n'est pas qu'un sujet de gestion. Le jour où la recherche juridique se facture à l'unité, **l'arbitrage entre chercher davantage et coûter moins cher devient un arbitrage professionnel**, pas seulement budgétaire. Une profession dont l'obligation de conseil est légale ne peut pas laisser ce curseur entre les mains d'un fournisseur.

> **Ce qu'il faudrait mesurer, et que presque personne ne mesure**
>
> - **Tokens par requête**, suivi dans le temps : l'indicateur avancé de toute dérive.
> - **Part des tokens de raisonnement** dans le coût de génération.
> - **Coût complet par dossier traité** — tokens, recherche, outils, reprises.
> - **Taux de réutilisation du cache** : ce que vous payez deux fois.
> - **Exposition au prix du modèle** : quelle part de votre facture suit le tarif d'un tiers.

V. Le pari d'Oliimpe

Nous avons construit Oliimpe Legal sur une conviction qui, en 2024, passait pour un renoncement commercial : **ne pas revendre d'intelligence artificielle**. Pas d'interface, pas de site où se connecter, pas de modèle « maison » facturé en plus. Oliimpe vit à l'intérieur de l'IA que le professionnel utilise déjà, et s'ajoute à son forfait existant.

> « Tout le monde construisait des interfaces par-dessus un modèle. Nous avons parié que dans trois ans, la question ne serait plus "quelle IA utilisez-vous" mais "combien vous coûte celle que vous utilisez déjà". Ce jour-là, le deuxième compteur qu'on avait posé sur le bureau du client sera le premier qu'il coupera. »
>
> **Pauline Lahary**, co-fondatrice et CEO d'Oliimpe

Ce choix a trois conséquences économiques, et elles découlent directement de tout ce qui précède.

Notre prix n'est pas indexé sur le prix des modèles

Nous ne prenons pas de marge sur l'inférence : nous vendons la recherche, pas l'intelligence. Quand le prix d'un modèle de pointe augmente, notre tarif ne bouge pas, et nous n'avons aucun intérêt à retenir nos clients sur une génération ancienne. Le client branche le modèle qu'il veut, le jour de sa sortie.

Moins de tokens brûlés à chercher

Une IA sans sources structurées explore : elle formule, relit, revient en arrière, et chacun de ces détours est facturé. Notre travail consiste à indexer les corpus de lois, à les relier entre eux dans un graphe, et à ramener directement le bon texte dans le bon contexte. Sur nos tests internes, la bonne réponse revient **environ cent fois plus vite** qu'une recherche menée à l'aveugle — et ce qui n'est pas cherché n'est pas payé. *La mesure est la nôtre, sur nos propres jeux de test ; nous publierons le protocole.*

> « Un modèle a une excellente tête et une mémoire courte. Il ne peut pas retenir des centaines de millions de documents, ni se souvenir du chemin pour y naviguer. Nous tenons cette mémoire à sa place. C'est moins spectaculaire que de construire une IA de plus, mais c'est là que se trouve l'économie. »
>
> **Pauline Lahary**

Une dépense lisible

Le compte d'accès au modèle est ouvert au nom du cabinet, facturé au cabinet, résiliable par le cabinet. Il voit ce qu'il consomme, requête par requête. Nous ne nous interposons pas entre lui et son fournisseur d'IA, et nous ne prenons pas de marge au passage. **C'est moins confortable pour nous. C'est la seule position tenable quand le coût de l'IA devient un poste de dépense surveillé.**

Le pari, en une phrase : la valeur se déplace du modèle vers ce qu'on lui donne à lire.

Les modèles convergent et se banalisent. Ce qui différenciera deux intelligences artificielles, ce ne sera plus leur capacité à raisonner, mais la donnée qu'elles savent atteindre, les liens qu'elles savent faire — et le nombre de tokens qu'elles ne gaspillent pas pour y arriver.

— Sources et notes

Les références ci-dessous sont classées dans l’ordre des appels de note du document.

1. Stanford HAI, *AI Index Report 2025* — coût d’inférence à performance GPT-3.5 constante, 20 $ (nov. 2022) → 0,07 $ par million de tokens (oct. 2024). hai.stanford.edu/ai-index
2. Epoch AI, *The plunging price of thought* — baisse de 9× à 900× par an selon la tâche. epoch.ai
3. Sam Altman, conférence de la Federal Reserve, juillet 2025 — « intelligence too cheap to meter », facteur 10 par an sur cinq ans.
4. Sundar Pichai, Google Cloud, octobre 2025 — 1,3 quadrillion de tokens par mois ; 480 000 milliards en mai 2025. the-decoder.com
5. Menlo Ventures, *2025 Mid-Year LLM Market Update* — dépense API entreprise 3,5 Md$ (fin 2024) → 8,4 Md$ (mi-2025). menlovc.com
6. Satya Nadella, X / LinkedIn, 26 janvier 2025 — « Jevons paradox strikes again ». geekwire.com
7. FinOps Foundation, *State of FinOps 2026* — 98 % des équipes pilotent une dépense d’IA ; 73 % de dépassements. data.finops.org
8. McKinsey, enquête mai 2026 — 93 % des organisations dépassent leur budget IA.
9. Microsoft Research — imprévisibilité de la consommation ; variance jusqu’à 30× sur une même tâche agentique.
10. *The Reasoning Tax: Token Economics of LLM Reasoning*, arXiv — part médiane du coût de génération consacrée au raisonnement interne : 94,7 %. arxiv.org
11. OckBench (arXiv), citant Epoch AI — croissance d’environ 5× par an de la longueur des sorties de raisonnement. arxiv.org
12. EY, cité par la presse spécialisée — coût d’une interaction agentique : 0,04 $ (2023) → 1,20 $ (2026).
13. FinOps Foundation · analyses de pratique — « tokens par requête » comme indicateur avancé de dérive.
14. Agence internationale de l’énergie, *Energy and AI*, avril 2025 — 415 TWh (2024) → 945 TWh (2030). iea.org/reports/energy-and-ai
15. Sam Altman, billet de blog, juin 2025, et auditions parlementaires — convergence du coût de l’intelligence vers le coût de l’énergie.

Les chiffres issus de rapports privés (Menlo Ventures, McKinsey, EY) et de prépublications n’ont pas la même robustesse que ceux des sources officielles (AIE, Stanford HAI). Ils sont signalés comme tels. Les mesures de performance d’Oliimpe sont internes et seront publiées avec leur protocole.