L'IA est-elle rentable ?

Anatomie d'un business model qui brûle des milliards

Sophie — The Monocle Bear

themonoclebear.com
Février 2026

OpenAI a généré 20 milliards de dollars de revenus en 2025. Et perdu au minimum 9 milliards. Anthropic lève 30 milliards en février 2026. DeepSeek propose la même performance à 1/30ème du prix. Le seul qui s'enrichit, c'est Nvidia — le vendeur de pelles pendant la ruée vers l'or. Enquête sur un business model que personne ne comprend, et que personne ne maîtrise.

Le paradoxe fondamental : plus tu vends, plus tu perds

L'IA générative a un problème structurel que ni le cloud computing ni le SaaS classique n'ont jamais eu : le coût marginal ne tend pas vers zéro.

Chaque requête consomme de la puissance de calcul. Chaque token généré mobilise des GPU qui coûtent des dizaines de milliers de dollars l'unité. Contrairement à un fichier hébergé sur un serveur cloud — qui coûte quelques fractions de centime à servir une fois stocké — une inférence LLM est un calcul intensif. À chaque fois. Sans exception.

Résultat : sur l'année 2025, OpenAI a dépensé environ **1,69 $ pour chaque dollar de revenu généré** (22 milliards de dépenses projetées contre 13 milliards de revenus). Au premier semestre, le ratio était encore pire — près de 2 $ dépensés pour chaque dollar gagné (8,5 milliards pour 4,3 milliards). Deutsche Bank estime que les pertes cumulées avant rentabilité pourraient atteindre **143 milliards de dollars** entre 2024 et 2029 — et ce chiffre n'inclut même pas l'engagement récent d'OpenAI de 1 400 milliards de dollars en infrastructure de data centers.

Pour mettre ce nombre en perspective : Uber a perdu 18 milliards sur six ans avant d'être rentable. Tesla, 9 milliards sur neuf ans. OpenAI vise un ordre de grandeur au-dessus.

La guerre des prix qui ruine tout le monde

Le prix des tokens API baisse à une vitesse vertigineuse. Le coût d'un million de tokens pour une performance équivalente à GPT-4 est passé de 20 $ fin 2022 à environ 0,40 $ début 2026. Les prix d'inférence chutent en médiane de **50x par an**.

C'est une excellente nouvelle pour l'adoption. C'est une catastrophe pour la rentabilité.

Les développeurs — devenus les plus gros consommateurs de tokens — en mangent à longueur de journée via les API. Claude Code d'Anthropic génère à lui seul 2,5 milliards de dollars de revenus annualisés en février 2026, un chiffre qui a doublé depuis le début de l'année.

État des lieux tarifaire — début 2026

Modèle	Input ($/M tokens)	Output ($/M tokens)	Note
GPT-4o	2,50	10,00	
Claude Sonnet 4	3,00	15,00	
GPT-5.2 Pro	21,00	168,00	Haut de gamme

Qwen3-Max	0,46	1,84	
Qwen3.5-Plus	0,12	—	1/18e de Gemini 3 Pro
GLM-4.7	0,60	2,20	
GLM-4.7-Flash	Gratuit	Gratuit	Sans quota
DeepSeek V2	0,14	0,28	

DeepSeek a posé une bombe en proposant des prix 95 à 98 % inférieurs aux concurrents occidentaux. Et DeepSeek n'a pas besoin d'être rentable — c'est un projet de recherche fondamentale financé par un hedge fund chinois, High-Flyer, qui a réalisé 56,6 % de retour sur investissement en 2025. Le fondateur, Liang Wenfeng, a explicitement déclaré que DeepSeek est motivé par la curiosité, pas par le profit.

Et DeepSeek n'est pas seul. Alibaba distribue gratuitement Qwen sous licence Apache 2.0 — le modèle open source le plus téléchargé au monde — et le monétise indirectement via son cloud. Zhipu AI (GLM) vient de faire la plus grosse IPO "modèle fondationnel" de l'histoire à Hong Kong, avec des revenus de 45 millions de dollars pour 350 millions de pertes. L'écosystème chinois attaque le marché par tous les flancs : le prix, l'open source, et le volume.

> *Quand vos concurrents les plus agressifs ne cherchent même pas à gagner de l'argent sur le modèle lui-même, vous avez un problème structurel.*

Le piège de l'abonnement à 20 €

ChatGPT Plus à 20 $/mois, Claude Pro au même tarif. Ces abonnements sont le cœur du revenu grand public de l'IA.

Le modèle tient sur une asymétrie classique : la majorité des abonnés utilisent peu, ce qui compense les utilisateurs intensifs qui consomment bien plus que ce que leur abonnement couvre. C'est le même principe que les salles de sport — le business model repose sur ceux qui paient sans venir.

Mais contrairement à une salle de sport, l'usage de l'IA **augmente** avec le temps, pas l'inverse. À mesure que les modèles deviennent plus capables, que les agents se multiplient, que les intégrations se généralisent, chaque utilisateur consomme davantage. OpenAI compte déjà 800 millions d'utilisateurs actifs hebdomadaires. Le jour où une fraction significative d'entre eux utilise l'IA comme un outil de travail quotidien — ce qui est précisément l'objectif — l'équation économique explose.

La preuve : depuis le 9 février 2026, OpenAI affiche de la publicité dans ChatGPT pour les utilisateurs gratuits et Go. Seulement 5 % des 800 millions d'utilisateurs paient un abonnement. Omnicom Media a confirmé que plus de 30 de ses clients participent au programme pilote. Sam Altman avait pourtant qualifié la publicité dans l'IA de "uniquely unsettling" en mai 2024, la décrivant comme un "last resort". Le last resort est arrivé. Quand une entreprise à 20 milliards de chiffre d'affaires bascule dans la pub,

c'est que les abonnements ne suffisent pas.

L'état financier des principaux acteurs

OpenAI

• **Revenus 2025** : plus de 20 milliards $ (triplement vs 2024). H1 2025 : 4,3 milliards $ de CA, 8,5 milliards $ de dépenses

• **Pertes H1 2025** : 4,7 milliards $ (perte opérationnelle de 7,8 milliards $). Cash burn : 2,5 milliards $, laissant 17,5 milliards $ en réserve

• **Levées de fonds** : négocie début 2026 un tour pouvant atteindre 100 milliards $ pour une valorisation de ~730 milliards $

• **Rentabilité espérée** : 2029-2030. Revenus cibles : 200 milliards $/an d'ici 2030

Anthropic

• **Revenus fin 2025** : 9 milliards $ annualisés (dont 2,5 milliards rien que pour Claude Code)

• **Levée Series G** (février 2026) : 30 milliards $ à 380 milliards $ de valorisation

• **Total levé** : 67,3 milliards $. Rentabilité espérée : 2028

• **Clients** : 300 000+ entreprises, Microsoft dépense ~500 millions $/an

Google (Gemini / DeepMind)

• **Capex IA prévu 2026** : 175-185 milliards $ (plus du double de 2025)

• **Stratégie** : subsidie massivement l'API Gemini pour capter des parts de marché

• **Avantage** : les revenus pub et cloud d'Alphabet absorbent les pertes IA

DeepSeek

• **Modèle** : financé par High-Flyer (hedge fund quantitatif, ~10 milliards $ sous gestion)

• **Coût V3** : 6 M$ en heures GPU déclarées (coût réel estimé ~1,3 milliard $ incluant infra et R&D)

• **Objectif** : recherche fondamentale, pas de commercialisation immédiate

Qwen (Alibaba)

• **Investissement** : 53 milliards $ en infrastructure IA sur trois ans

• **Stratégie** : open source agressif (Apache 2.0) — 600 M+ downloads. Monétisation indirecte via Alibaba Cloud

• **Dernier modèle** : Qwen 3.5 (février 2026), 397B paramètres / 17B actifs (MoE), 60 % moins cher à opérer

• **Logique** : les revenus e-commerce et cloud (130 milliards $/an) absorbent les coûts du modèle

Zhipu AI / Z.ai (GLM)

• **IPO** : janvier 2026 à Hong Kong — 558 M$ levés, titre +173 % en un mois

• **Revenus 2024** : ~45 M$. Pertes : ~350 M$. R&D > 700 % du revenu

• **Business model** : 85 % du CA en on-premise pour gouvernements et SOEs chinois (marge brute 66 %)

• **Contrainte** : blacklisté Entity List US. Réentraîne sur puces Huawei Ascend

Le seul qui gagne de l'argent : Nvidia

Pendant que les fournisseurs d'IA brûlent du cash, un acteur empoche le jackpot : le vendeur de pelles.

Nvidia affiche un revenu de **57 milliards de dollars pour le seul Q3 de son exercice fiscal 2026** (octobre 2025), dont 51,2 milliards uniquement pour la division Data Center — en hausse de 66 % sur un an. Sur l'année fiscale 2026, les projections tournent autour de **213 milliards de dollars de revenus** avec une croissance de 63 %.

Nvidia détient **86 % du marché des puces IA** (contre 25 % en 2021). La demande est si forte que Jensen Huang parle d'un carnet de commandes de **500 milliards de dollars** pour les plateformes Blackwell et Rubin.

> *L'analogie historique est limpide : pendant la ruée vers l'or, ce sont les vendeurs de pioches et de jeans qui se sont enrichis, pas les chercheurs d'or.*

L'analogie AWS : promesse ou mirage ?

Le parallèle le plus souvent invoqué pour justifier les pertes actuelles est celui d'AWS. Amazon Web Services a fonctionné à perte pendant environ trois ans après son lancement en 2006, avant de devenir rentable vers 2009. Aujourd'hui, AWS génère des dizaines de milliards de profit annuel avec des marges opérationnelles de 30 %.

L'argument : l'IA suit le même schéma. Investir massivement maintenant, dominer le marché, puis récolter les profits une fois l'infrastructure amortie et les clients verrouillés.

Sauf que la comparaison a des limites structurelles :

Ce qui se ressemble :

- Investissements initiaux massifs en infrastructure
- Course aux parts de marché via des prix bas
- Effets de réseau et coûts de migration qui créent du verrouillage

Ce qui diffère fondamentalement :

- Le cloud a des coûts marginaux quasi nuls. L'inférence IA est un calcul intensif à chaque requête.
- AWS est passé de pertes à profits en 3 ans. OpenAI prévoit au minimum 5 ans de pertes supplémentaires — à une échelle 100x supérieure.
- Le cloud vendait du calcul standardisé. L'IA vend de l'intelligence — un produit qui exige des investissements continus en R&D.
- AWS n'avait pas un écosystème chinois entier — DeepSeek, Qwen, GLM — qui attaque simultanément par le prix, l'open source et le volume.

Le parallèle AWS est séduisant. Mais il pourrait aussi être dangereux s'il sert de justification pour ignorer des différences structurelles fondamentales.

Les signaux d'alarme

95 % des organisations ne voient aucun retour sur leurs investissements en IA générative, selon un rapport du MIT Media Lab d'août 2025. Si les clients ne tirent pas de valeur mesurable de l'IA, ils finiront par réduire leurs dépenses.

58 % du capital-risque mondial allait vers des startups IA début 2025 — 73 milliards de dollars rien qu'au premier trimestre. Une concentration qui rappelle furieusement les niveaux de la bulle dot-com, mais à une échelle 100 fois supérieure.

Les PDG eux-mêmes alertent. Sam Altman a reconnu qu'une bulle IA est en cours. Demis Hassabis (DeepMind) parle de startups "wildly overpriced". Ray Dalio (Bridgewater) compare directement la situation à la bulle dot-com.

La consolidation est inévitable. VentureBeat prédit des faillites massives parmi les "AI wrappers" d'ici fin 2026. On s'attend à ce que 2 à 3 acteurs dominent le marché des modèles fondationnels d'ici 2028.

Alors, est-ce rentable — un jour ?

La réponse courte : probablement oui, pour 2 ou 3 acteurs. Pas pour l'industrie dans son ensemble.

Ce qui pourrait sauver le modèle :

- La baisse des coûts d'inférence via la quantification (réduction de 60-70 %), le décodage spéculatif (latence ÷2-3x), et les puces spécialisées
- Le routage intelligent : 90 % des requêtes vers un petit modèle, 10 % vers un frontier = -86 % de coûts
- L'émergence d'un vrai verrouillage client (données propriétaires, intégrations profondes, agents)
- L'IA comme couche d'infrastructure invisible intégrée dans chaque logiciel

Ce qui pourrait faire tout s'effondrer :

- L'absence de retour mesurable pour les entreprises clientes
- Un effondrement du financement VC (répétition du scénario dot-com)
- L'open source et les modèles locaux qui érodent la valeur des API cloud — quand un 32B Q8 tourne sur un Mac Studio, pourquoi payer une API ?
- La réglementation (GDPR, AI Act) qui contraint l'usage des données

Ce que ça signifie pour les décideurs

Sur la pérennité du fournisseur : votre fournisseur peut-il survivre à un assèchement du capital-risque ? OpenAI et Anthropic brûlent des milliards. Si le financement se tarit — et l'histoire montre que ça arrive toujours — que devient votre intégration ?

Sur la dépendance : si vous construisez votre infrastructure sur une API cloud, quel est votre plan B ? Le coût de migration est gérable aujourd'hui. Demain, avec des agents et des données contextuelles profondément intégrés, il pourrait ne plus l'être.

Sur le coût réel : les prix API actuels ne reflètent probablement pas le coût réel de l'inférence. Ils sont subventionnés par le capital-risque. Budgétisez une hausse de 30 à 50 % à moyen terme.

Sur l'alternative locale : pour certains usages, un modèle local bien quantifié peut offrir 80 % de la performance à une fraction du coût récurrent. C'est une stratégie de résilience, pas un rejet du cloud.

L'industrie de l'IA est dans une phase que les économistes appellent la "destruction créatrice" — des milliards investis, des fortunes détruites, et au bout du tunnel, probablement une transformation réelle

de l'économie. Comme l'internet avant elle.

Mais entre-temps, quelqu'un paie la facture. Aujourd'hui, ce sont les investisseurs en capital-risque. Demain, si les modèles économiques ne se stabilisent pas, ce seront les clients — par des hausses de prix, des dégradations de service, ou la disparition pure et simple de leur fournisseur.

La question n'est pas de savoir si l'IA va transformer le monde.
La question est : **qui sera encore debout quand la facture arrivera ?**

Sources et données

— OpenAI ARR $20B — PYMNTS (pymnts.com)
— OpenAI prévisions de pertes $143B — Deutsche Bank via eMarketer
— OpenAI pertes et projections — Fortune (novembre 2025)
— OpenAI H1 2025 results — The Information
— Anthropic Series G $30B — Bloomberg (février 2026)
— Anthropic revenus $9B run rate — Bloomberg (janvier 2026)
— Nvidia Q3 FY2026 — Nvidia Newsroom
— Prix API comparaison 2026 — IntuitionLabs
— DeepSeek business model — TechCrunch
— AI bubble vs dot-com — IntuitionLabs
— MIT Media Lab 95% zero return — MIT (août 2025)
— Alphabet Q4 2025 capex — CNBC (février 2026)
— High-Flyer 56% returns — Bitcoin Ethereum News
— Alibaba Qwen challenging proprietary AI economics — AI News
— Alibaba Qwen 3.5 launch — CNBC (février 2026)
— Zhipu AI / Z.ai deep dive — Financial Content (février 2026)
— Qwen API pricing — Alibaba Cloud Model Studio

The Monocle Bear — themonoclebear.com

AI Workflows & Agentic UX — Brussels, Belgium