

AGENT LIBRE

WEBINAIRE GRATUIT · 45 MIN · EN LIGNE

SERVICES GÉRÉS D'INFÉRENCE IA

Comprendre votre facture d'IA avant qu'elle décide pour vous

Structure des coûts, effet sur vos marges, projection sur cinq ans et
risque de dépendance à un seul fournisseur.

 Hébergé et opéré au Canada | Montréal, Québec

AU PROGRAMME

- La structure et la dynamique réelles de votre facture IA
- Son impact sur vos marges et votre compétitivité
- Projection sur cinq ans : coûts et usage à la hausse, et vous ?
- Tout miser sur un seul fournisseur : quels risques concrets
- L'IA fait maintenant partie de votre marge opérationnelle — et entre la géopolitique, le dollar américain et les tarifs, qui la contrôle ?

LE POINT DE DÉPART

Les démos d'agents sont impressionnantes. Elles ne sont pas gratuites.

On voit passer les vidéos : un agent qui gère le support, un autre qui écrit le code, un autre qui fait la comptabilité. Personne ne montre le compteur qui tourne pendant ce temps-là.

UN HUMAIN

~5 000

tokens générés par jour. Il dort, il mange, il prend des vacances.

UNE FLOTTE D'AGENTS

~100 000 000

tokens par jour : 20 agents à 50 tokens/seconde, 24 h sur 24. Elle ne dort pas, et chaque tour relit tout l'historique.

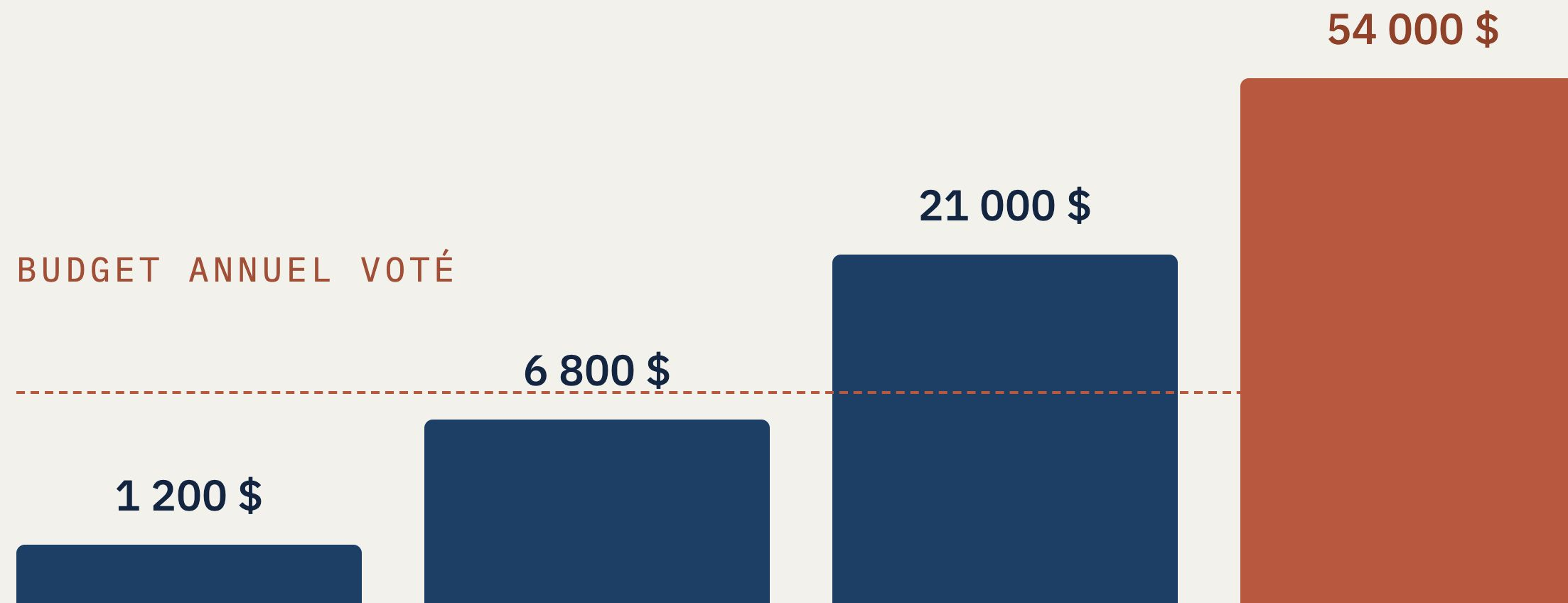
L'ÉCART

20 000×

Votre facture ne suit plus le nombre d'employés.

Calcul : un agent de codage autonome tient environ 50 tokens/seconde, soit 4,3 M de tokens par jour – 20 agents en continu donnent ~86 M. Côté humain, 2 000 mots écrits par jour valent environ 3 000 tokens.

Plus on l'adopte, plus ça coûte



Progression typique observée sur 12 à 18 mois, à l'usage.

Une adjointe qui rédige ses courriels, puis le service à la clientèle qui répond avec un agent IA, puis la facturation et les soumissions, puis la prospection en continu. Chaque étape est une bonne nouvelle pour l'entreprise.

Et chaque étape est une mauvaise nouvelle pour la facture. Alors on freine — quotas par personne, outils limités, projets remis à plus tard.

On finit par optimiser la facture au lieu des processus.

OÙ VOUS SITUEZ-VOUS

Trois profils, un même horizon

ENTREPRISES DE
30 À 60 EMPLOYÉS

01 **2 000 \$ / mois**

« C'est suffisant. »

CONSTAT

Adoption IA encore faible. L'outil est là, les processus ne le sont pas.

La transformation arrive dans les 16 prochains mois.

02 **15 000 \$ / mois**

« C'est suffisant. »

CONSTAT

Adoption IA alignée sur les tendances actuelles — sauf que vous payez peut-être jusqu'à 5× le coût réel.

La compétition va tirer vos volumes vers le haut, et votre facture avec.

03 **15 000 \$ / mois**

« Et je suis limité. »

CONSTAT

Adoption IA supérieure aux tendances actuelles. Vous freinez déjà vos équipes.

Le plafond n'est plus technique, il est budgétaire.

Pour les profils 2 et 3, souveraineté et élimination des intermédiaires arrivent d'elles-mêmes.

Chacun son rythme. Mais c'est une éventualité certaine.

CE QUE VOUS PAYEZ VRAIMENT

Entre le GPU et vous, il y a du monde

01

Le centre de données

Bâtiment, électricité, refroidissement, GPU. C'est ici que le travail se fait.

+

02

L'hyperscaleur

Loue le matériel à l'heure, en dollars américains, avec sa marge.

+

03

Le laboratoire de modèles

Amortit son entraînement, sa R&D et son marketing mondial sur chaque token que vous générez.

+

04

L'API à l'usage, puis l'agrégateur

Facture au token, sortie 3 à 5× l'entrée, refacture chaque relecture d'historique — et le revendeur ajoute sa couche.

=

Vous

Votre facture

Vous payez le même GPU quatre fois — et vous ne savez pas combien il coûte.

Coût marginal réel d'un million de tokens d'entrée sur un serveur dédié :
~0,09 \$ CAD

Facturé de **20 à 45×** ce prix.

Soyez à l'affût de votre dépense IA indirecte

Vous achetez un super logiciel boosté à l'IA — il traite votre support à la clientèle, il fait votre prospection. Ce que le contrat ne dit pas fort : il vous relaie sa facture à l'usage, avec sa marge par-dessus.

C'est une facture de tokens que vous payez sans la voir, sans la mesurer et sans pouvoir la négocier.

LE RÉFLEXE À PRENDRE

Demandez à faire rouler ce logiciel sur votre infrastructure IA. Les fournisseurs sérieux vous laissent brancher vos propres clés.

TROIS RAISONS DE L'EXIGER

01 Vos coûts

Ce volume passe dans votre débit déjà réservé, au lieu d'être refacturé au token par un tiers.

02 Votre gouvernance de données

Vos dossiers clients et vos données sensibles restent sur votre couche d'inférence, avec vos journaux d'audit.

03 Votre culture IA

Vos propres garde-fous internes s'appliquent partout : mêmes règles, même filtrage, peu importe l'outil qui appelle.

La question à poser avant de signer : « **Est-ce que je peux utiliser mes propres clés API d'inférence IA ?** »

Les deux réflexes qui ne survivent pas au long terme

« On va le faire rouler sur nos laptops »

- Un laptop est loin des 50 tokens/s qu'un seul agent demande.
- GLM 5.2 n'entre pas dans un laptop : le modèle réduit qui y entre n'est pas celui testé.
- La mémoire explose bien avant 256K tokens.
- La machine est bloquée, et rien n'est mutualisé.

UNE RÉPONSE DE 1 500 TOKENS

Laptop, modèle réduit : ~8 tokens/s

3 min 07 s

Débit réservé, ~99 tokens/s par agent

15 s

Bon pour une démo, jamais pour la production.

« On achète un forfait individuel personnel par employé »

L'équivalent de les laisser travailler avec leur Gmail personnel.

- Aucune surveillance de l'utilisation : on s'en sert pour planifier ses voyages et le reste.
- Détournement : votre capacité part dans des projets personnels.
- Les données clients restent dans un historique personnel, hors de votre contrôle.
- Un siège, c'est pour un humain : vos agents ne s'arrêtent pas à 17 h.

Dépense d'outillage, pas stratégie d'inférence.

« La minute » avait été inventée elle aussi

Les télécoms facturaient à la minute. Leur coût réel, lui, était de l'infrastructure : des tours, du cuivre, de la fibre — payés une fois, puis amortis. La minute n'apparaissait nulle part dans leur bilan.

01

Une unité de facturation, pas de coût

La minute a été créée pour transformer une infrastructure fixe en revenu variable. Plus vous parliez, plus ça rapportait — sans que ça coûte davantage.

02

Et tout le monde s'est limité

On comptait ses appels, on rappelait après 18 h, on gardait l'interurbain pour Noël. Exactement ce que vos équipes font aujourd'hui avec leurs tokens.

03

Ça s'est réglé par le forfait

Les illimités ne sont pas arrivés par générosité, mais par concurrence : quelqu'un a fini par facturer ce qui coûtait vraiment — l'accès au réseau.

Le token, c'est la minute de 2026. Le coût réel, c'est le GPU — et un GPU, ça se réserve.

LE RENVERSEMENT

Vous n'achetez plus des tokens. Vous réservez du débit.

Un fournisseur de services gérés opère votre courriel, votre site web ou votre système de sécurité.

Un fournisseur de services gérés d'inférence fait pareil pour votre IA : capacité dimensionnée, opérations surveillées, facture fixe. Ni laboratoire de modèles, ni revendeur.

LE MODÈLE À L'USAGE

Vous achetez des tokens à l'unité

Vous absorbez chaque variation de charge

Vous surveillez les quotas et les limites

Le fournisseur décide du modèle et du contexte

Vous subissez les hausses et le taux de change

LES SERVICES GÉRÉS D'INFÉRENCE

Vous réservez un débit, pas des tokens

La variation est mesurée; le forfait ne bouge pas

La capacité et la disponibilité sont gérées

La charge est syndiquée entre les membres

Vous éliminez les intermédiaires du GPU jusqu'à vous

À l'usage, vous payez surtout le droit de consommer n'importe quand

À L'USAGE

Le droit d'y aller quand ça vous chante

Zéro engagement : vous pouvez ne rien consommer ce mois-ci, ou dix fois plus demain matin

Le fournisseur doit garder de la capacité en réserve pour absorber cette liberté

Cette réserve est payée par tout le monde, à chaque token

C'est une prime de flexibilité.

AU DÉBIT RÉSERVÉ

Des tokens à la seconde, garantis

Vous vous engagez sur une capacité, elle vous est réservée en tout temps

Aucune réserve à financer : la capacité est déjà à vous

Pas de prime à payer, donc un prix qui reflète le calcul réel

Vous payez un serveur, pas une option.

Pourquoi payer le premium du « n'importe quand »? **Vos agents roulent tous les jours, à un rythme que vous connaissez déjà.**

LE MÉCANISME

Le syndicat

Votre débit réservé sur une infrastructure partagée. Ce que le forfait couvre :

01

L'exploitation technique

Débit garanti, réservé à l'avance

Optimisation continue du modèle sur le GPU

À matériel constant, le débit augmente

02

L'administration

Appariement, arrivées et départs des membres

La capacité libérée accueille d'autres membres

Baux et partenaires de centre de données

03

La négociation des coûts

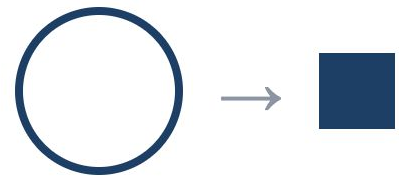
Arbitrage à mesure que le volume cumulé grandit

Veille continue du marché des GPU

Renégociation des ententes aux 6 à 12 mois

Vous n'avez qu'à relayer votre volume vers la passerelle partagée.

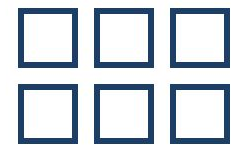
Le compteur, ou vos propres panneaux



LE COMPTEUR

Hydro-Québec

Vous branchez, vous consommez, vous payez au kilowattheure. Aucune immobilisation — et aucun mot à dire sur le tarif. C'est le grand fournisseur d'IA.



LES PANNEAUX, TOUT SEUL

Sur votre propre toit

Vous achetez, vous installez, vous entretenez, vous gérez les surcharges et les pannes. Autonomie réelle — mais c'est une expertise que vous n'avez pas.



LES PANNEAUX, OPÉRÉS

Le fournisseur géré

Un fournisseur de services gérés installe un système qui vous appartient — ou qui vous est loué — et prend en charge l'expertise. Vous éliminez quand même tous les intermédiaires entre vous et l'électricité.

L'installation est à vous, l'expertise vient avec — et entre vous et l'électricité, il n'y a plus personne.

LES CHIFFRES

Ce que la même capacité coûte, selon l'unité de facturation

FORFAIT FIXE — SERVICES GÉRÉS

18 000 \$

Débit réservé : **1 980 tok/s**

≈ **40** agents à 50 tok/s — ou 20 agents à **99 tok/s** chacun

Le montant ne change pas si vos agents roulent 24 h sur 24.

LA MÊME CAPACITÉ, FACTURÉE À L'USAGE

50 % d'utilisation

52 700 \$

par mois

70 % d'utilisation

73 800 \$

par mois

100 % d'utilisation

105 500 \$

par mois

Calcul : tokens générés par le débit réservé sur un mois complet, au tarif de sortie de Claude Sonnet (15 \$ US/M), converti en CAD.

Et si votre compétiteur paie 600 % moins cher pour la même tâche ?

Ça n'a rien à voir avec la technologie : c'est un avantage de coût sur chaque soumission, chaque courriel, chaque dossier traité.

OPTION A

Il baisse ses prix

Son coût de traitement par dossier est six fois plus bas. Il peut couper ses tarifs et rester rentable — vous, non. Vous perdez des appels d'offres sans jamais voir pourquoi.

OPTION B

Il garde sa marge et en fait six fois plus

Même prix, même marge — mais six fois plus de prospection, de suivis clients, de contenu, de nouveaux marchés testés. À budget égal, il avance six fois plus vite.

Le coût de l'inférence n'est plus une ligne de TI. **C'est intégré à votre structure de prix.**

« Oui, mais le modèle, il est bon ? »

GLM 5.2, le modèle à poids ouverts de Z.ai, selon des évaluations publiques et indépendantes.

ÉVALUATION	GLM 5.2	Claude Sonnet 5	Claude Opus 4.8	GPT-5.5
SWE-bench Pro	62,1	63,2	69,2	58,6
Terminal-Bench 2.1	81,0	80,4	85,0	84,0
FrontierSWE	74,4	—	75,1	72,6
Détection IDOR (Semgrep)	39 % F1	—	32 % F1	—

Pour le gros du travail agentique quotidien, l'écart de qualité est mince. L'écart de coût, lui, ne l'est pas — c'est ce qui rend le forfait possible.

Et la tendance joue pour vous : le retard des poids ouverts sur les laboratoires privés est passé de 6-9 mois à 3-5 mois.

LA NUANCE QUI COMPTE

Ce que vous expérimentez au quotidien vient principalement du harness, et non du modèle

Le harness d'agent, c'est tout ce qu'on construit autour du modèle pour le faire travailler : les outils auxquels il a accès, ce qu'il se rappelle, ce qu'il fait quand une étape échoue, les règles qu'il doit respecter et la façon dont une tâche est découpée.

LE HARNESS

Ce qui fait la différence ressentie

Les bons outils branchés au bon moment

Le contexte qu'on lui donne — et celui qu'on lui épargne

Les relances et la vérification quand une étape échoue

Vos règles internes, appliquées à chaque appel

ET CÔTÉ COÛT

Le harness est un coût fixe

Il s'écrit une fois, il vous appartient, il s'améliore

Il ne se refacture pas au token

Il fonctionne avec le modèle de votre choix

Changer de modèle ne le jette pas aux poubelles

Vous investissez dans le harness une fois. **Le modèle, lui, devient interchangeable — et c'est exactement ce qu'on veut.**

OÙ EST LE PIÈGE

Un modèle ouvert, ça existe comment ?

L'équivalent le plus proche : l'article scientifique publié en accès libre. La méthode est publique, n'importe qui peut la reprendre, la vérifier, bâtir dessus.

01

Publié, pas donné

Ce qui est ouvert, ce sont les paramètres du modèle : le résultat de l'entraînement, publié comme un papier de recherche. Vous pouvez le télécharger, l'auditer, le faire rouler où vous voulez.

02

Pourquoi ils publient

Une immense communauté de scientifiques et de développeurs publie par contribution et réciprocité. En échange, celui qui publie garde la main sur sa communauté — et récolte la réputation qui vient avec.

03

Le piège est ailleurs

Le modèle est ouvert; le faire rouler, non. Ça demande des GPU, de l'exploitation, de l'optimisation et de la disponibilité — le métier d'un fournisseur de services gérés d'inférence.

Les modèles sont ouverts. **Le matériel qui les roule, lui, se paie — et c'est la seule facture qui reste.**

PROJECTION SUR CINQ ANS

L'inférence est devenue une ligne de votre marge. Qui la contrôle ?

LE TAUX DE CHANGE

Vos tarifs sont en dollars américains. Votre revenu, non. Chaque mouvement du taux passe direct dans votre marge.

LES TARIFS ET LA GÉOPOLITIQUE

Matériel, énergie, accès aux modèles : trois dossiers politiques sur lesquels vous n'avez aucun levier.

LE FOURNISSEUR UNIQUE

Il change son prix, son quota, son modèle ou ses conditions : vous l'apprenez par courriel, et vous encaissez.

L'USAGE QUI MONTE

Vos volumes agentiques vont croître d'un ordre de grandeur. À l'usage, votre facture suit la même courbe.

Un débit réservé, contractuel, hébergé et opéré au Canada : c'est une décision de marge, pas une décision d'outillage.

C'est aussi la seule version de ce calcul que votre équipe finance peut engager sur trois ans.

Vos processus critiques. Vos IA clés s'ajoutent à vos employés clés.

Un employé, vous savez gérer ça

- Les échelles salariales vous donnent une prévisibilité de coût sur l'année.
- S'il est débordé, il vient vous voir : avant d'embaucher, vous discutez d'optimisation.
- Le roulement de personnel, vous l'avez déjà vécu : ça se remplace, ça se forme.

C'est prévisible.

Un fournisseur d'IA unique, moins

- Il change son prix, son quota, son modèle ou ses conditions : vous l'apprenez par courriel.
- Vos processus critiques sont bâtis sur son comportement : ses réponses, sa latence, ses limites.
- Changer de fournisseur, ce n'est pas un remplacement : c'est retester chaque processus, chaque prompt, chaque garde-fou.

Ça reste très abstrait.

Plus vous bâtissez dessus, moins vous pouvez en sortir. **Un débit réservé et des modèles à poids ouverts, c'est ce qui garde votre souveraineté.**

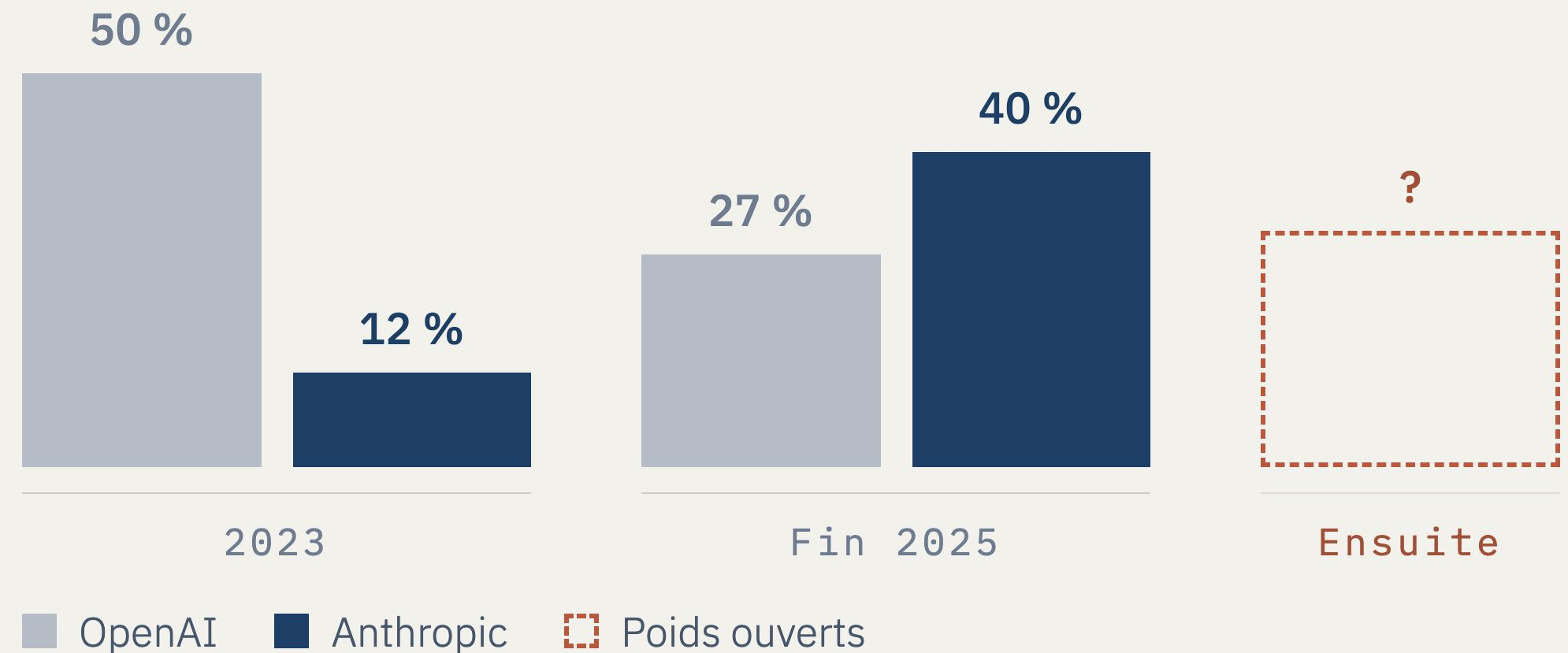
UN PEU D'HUMILITÉ

En 2023, tout le monde ne jurait que par ChatGPT

« On met ça sur ChatGPT » était la réponse à tout. Trois ans plus tard, en entreprise, c'est Claude qui mène — et personne ne s'en excuse.

La question qui se pose : pourquoi Claude serait-il la fin de l'histoire ?

Les modèles à poids ouverts sont peut-être en train de vivre leur propre moment « ChatGPT » — celui que Claude a fait vivre à ChatGPT.



Part du marché des API LLM en entreprise, Menlo Ventures (2023 → fin 2025). Sur le code seulement, Claude est passé à 54 % contre 21 % pour OpenAI.

AGENT LIBRE

WEBINAIRE · 45 MIN

Merci

Des questions ? C'est le moment.

Libérez les agents. Fixez la facture.

 Hébergé et opéré au Canada · Montréal, Québec