

changer d'abonnement

Claude relit toute ta conversation à chaque message : voilà pourquoi ta limite tombe avant les 45 annoncés. Les 4 postes de dépense, les 5 réflexes.

BLYX-AGENCY.COM · 24 JUILLET 2026 · PAR MELVIN BRETON

Ton quota Claude ne baisse pas au rythme de tes messages, il baisse au poids de chaque échange. Claude relit la conversation entière à chaque fois que tu lui écris : ta trentième question part donc avec les vingt-neuf précédentes, même si tu tapes trois mots.

Le geste le plus rentable est aussi le plus simple : ouvrir une conversation neuve dès que tu changes de sujet. Les quatre autres tiennent dans une journée de travail, sans compétence technique et sans changer d'abonnement.

1. Pourquoi ton quota fond si vite

À 20 euros par mois, tu as droit à environ 45 messages toutes les 5 heures. À 200 euros, environ 900. Entre les deux, un facteur 20.

Sauf que la plupart des gens à 20 euros tapent leur limite bien avant le quarante-cinquième message, et qu'ils n'y comprennent rien. La raison tient en une phrase, et presque personne ne la connaît.



Ton message ne part jamais seul. Il part avec l'intégralité de l'échange qui a précédé, si bien que plus la conversation avance, plus chaque nouveau message pèse lourd.

Ce que tu crois payer	Ce que tu paies vraiment
Ta question du moment	Ta question, plus tout l'historique relu
Un message court, donc léger	Un message court dans une longue conversation, donc lourd
Un quota qui baisse au nombre de messages	Un quota qui baisse à la taille de chaque échange

Au trentième message, tu ne paies presque plus ta question : tu paies la relecture des vingt-neuf précédentes. C'est exactement pour ça que ton quota fond sans prévenir, alors que tu as l'impression de ne rien avoir demandé d'énorme.

La bonne nouvelle, c'est qu'une fois le mécanisme compris, tu récupères une marge considérable sans rien changer à ton abonnement.

2. Les quatre postes de dépense

Avant de chercher des solutions, il faut savoir où fuit l'eau. Il y a quatre postes, et trois d'entre eux ne se voient nulle part dans l'application





L'historique
cumulé

Tout ce qui a été dit avant, relu à
chaque nouveau message

Le poste numéro
un, de loin

La mémoire

Ce que Claude retient de toi et
réinjecte dans chaque conversation

Modéré, mais
permanent

Les fichiers
collés

PDF, captures et pavés de texte
ajoutés en cours de route

Fort, et presque
toujours oublié

Le modèle

Un modèle très puissant mobilisé
pour une tâche simple

Fort, et totalement
invisible

Retiens un seul principe et tu peux oublier le tableau : tout ce que Claude doit lire pour te répondre te coûte du quota. Ton travail consiste à réduire ce qu'il doit lire, sans réduire la qualité de ce qu'il te rend.

3. Couper la conversation au bon moment

C'est le réflexe le plus rentable des cinq. Dès que tu changes de sujet, ouvre une nouvelle conversation : tu repars d'un historique vide, donc d'un coût plancher.

Le piège classique est la conversation fourre-tout, ouverte le matin et gardée toute la journée. Au bout de quelques heures, chaque message qu'on y écrit traîne derrière lui un dossier entier.

Reste le cas des vrais sujets longs, ceux qu'on ne peut pas couper sans perdre le fil. Il existe une manœuvre simple pour eux.



lignes que je pourrai réutiliser », puis colle ce résumé dans une conversation neuve. Tu gardes le contexte utile et tu jettes le poids mort.

4. Mettre la mémoire au régime

Si tu as activé la mémoire, Claude retient des éléments sur toi et les réinjecte au début de chaque nouvelle conversation. C'est pratique, mais cela consomme du quota en permanence, surtout quand la mémoire s'est remplie de vieux détails devenus inutiles.

Va dans les réglages, à l'entrée Mémoire. Trois actions y sont possibles :

- ✓ supprimer les souvenirs devenus inutiles, un par un ;
- ✓ mettre la mémoire en pause, elle existe toujours mais n'est plus réinjectée ;
- ✓ couper la génération de souvenirs à partir de l'historique si tu n'en as pas l'usage.

Fais ce ménage une fois par mois. Deux minutes de travail, un gain sur chacune de tes conversations suivantes. Les libellés de l'application changent d'une version à l'autre : si tu ne trouves pas l'entrée, cherche le mot mémoire dans la recherche des réglages.

5. Choisir le modèle qui suffit



Tous les modèles ne pèsent pas la même chose. Prendre le plus

Blyx Pour tout, c'est sortir le camion pour aller chercher le pain. 

Modèle	Ce qu'il fait très bien	Poids
Haiku	Reformuler, résumer, corriger, trier	Léger
Sonnet	Analyse, rédaction sérieuse, la majorité des tâches	Moyen
Opus	Raisonnement complexe, gros dossiers, arbitrages	Lourd

La règle tient en une ligne : Sonnet par défaut, Opus seulement quand la tâche résiste, Haiku pour tout ce qui est mécanique. Rien que ce tri change l'ordre de grandeur de ta consommation.

Le même raisonnement vaut au niveau de l'entreprise, quand il s'agit de décider quel assistant confier à quel métier. Le comparatif est traité à part dans [Claude, ChatGPT ou Copilot selon votre métier](#).

6. Alléger ce que Claude doit lire

Un PDF de quarante pages collé dans une conversation n'est pas lu une fois : il est relu à chaque message qui suit. Une capture d'écran lourde suit la même logique.

- ✓ Ne colle que la partie utile d'un document, jamais le document entier.
- ✓ Extrais la page, le tableau ou le paragraphe qui porte la réponse.
- ✓ Quand tu as fini d'exploiter un fichier, repars sur une conversation neuve.



Les fichiers volumineux sont le piège le plus sournois de la liste : on les a testés, parfois, on les oublie aussitôt, et ils alourdissent tout le reste de l'échange.



Il existe un cas particulier utile à connaître : le contexte que tu réexpliques tous les jours, ton métier, tes clients, ta façon d'écrire. Celui-là n'a rien à faire dans une conversation, il se range dans un fichier que Claude n'ouvre qu'au moment où il en a besoin. Le mécanisme est détaillé dans [le guide des Skills Claude](#).

7. Suivre ta consommation en direct

On ne pilote bien que ce qu'on mesure, et tu es loin d'être aveugle sur ce point.

Ouvre le menu de ton profil, en bas à gauche, puis les réglages, puis l'entrée Usage. Tu y vois deux barres : ta fenêtre glissante de 5 heures et ta limite hebdomadaire. Regarde-les avant de te lancer dans une grosse session, tu sauras si tu as la marge ou s'il vaut mieux attendre la fenêtre suivante.

Des extensions de navigateur affichent la même information dans ta barre d'outils. Nous ne les recommandons pas ici : une extension qui lit ta session Claude voit aussi tout ce que tu y écris, et sur un compte professionnel cela se décide en connaissance de cause, pas au détour d'un guide. La mesure native suffit largement.

8. Le cas particulier de Claude Code





Là, le gaspillage vient surtout des sorties de commandes : un test, un journal d'erreurs, un différentiel de quatre cents lignes que Claude relit intégralement. Trois leviers existent.

Levier	Ce que ça fait
RTK	Comprime la sortie des commandes avant qu'elle n'arrive à Claude, de 60 à 90 % de texte en moins. Gratuit et ouvert, 72 000 étoiles sur GitHub
Mode opusplan	Opus planifie et Sonnet exécute : tu gardes la finesse là où elle décide de quelque chose
Commande /compact	Comprime l'historique de la session en un résumé, quand elle devient trop lourde

Une seule commande branche RTK sur ton installation :

```
rtk init -g
```

Procédure vérifiée le 30/08/2026.

Tape aussi `/usage` directement dans Claude Code pour voir ta consommation de session, sans rien installer. Et garde en tête qu'une tâche programmée consomme le quota de ton compte exactement comme une session normale : c'est le point que rappelle [le guide des routines Claude](#).



9. Quand dix personnes utilisent Claude

Tout ce qui précède est écrit pour une personne seule devant son

 une entreprise, le problème change de nature, et c'est là  que ça coûte vraiment de l'argent.

Le quota est attaché à chaque compte, pas à l'entreprise. Tu n'as donc pas un réservoir commun mais dix réservoirs séparés, qui se vident à des vitesses très différentes selon les habitudes de chacun. La personne qui bloque un mardi après-midi n'est pas celle qui travaille le plus, c'est celle qui garde une conversation ouverte depuis lundi.

Le motif qui revient partout est toujours le même : une tâche identique refaite chaque semaine, réexpliquée depuis zéro à chaque fois. Le rapport du vendredi, la réponse type à une demande de devis, le tri du courrier entrant. Sur une année, c'est le même contexte payé cinquante fois.

Trois décisions règlent la majeure partie du problème, et aucune n'est technique :

- ✓ un contexte d'entreprise écrit une fois, partagé, que personne ne retape ;
- ✓ une règle simple sur le modèle à utiliser selon le type de tâche ;
- ✓ une conversation neuve par sujet, posée comme habitude collective et non comme conseil.

Ces trois décisions relèvent des réflexes, pas de l'outil, et elles survivent au prochain changement de version. C'est précisément la thèse que nous défendons dans [les réflexes qui restent quand les outils changent](#).

Reste le seuil au-delà duquel les réflexes ne suffisent plus. Quand tâche revient chaque semaine, qu'elle se décrit toujours de la même



façon et qu'elle mobilise plusieurs personnes, elle n'a plus sa place dans
une **Blyx** : elle mérite un outil construit une fois pour toutes,  **Blyx** qui ne repaie pas le contexte à chaque exécution. C'est le terrain de
nos [solutions sur mesure](#).

10. Ta checklist de la semaine

Garde-la sous les yeux pendant sept jours, les réflexes viennent vite.

Réflexe	Fréquence
Nouvelle conversation à chaque changement de sujet	À chaque fois
Résumé-relais pour les sujets qui durent	Quand ça traîne
Ménage dans la mémoire	Une fois par mois
Sonnet par défaut, Opus quand la tâche résiste	À chaque tâche
Ne coller que la partie utile d'un document	À chaque fichier
Un œil sur les barres d'usage	Avant les grosses sessions

Le prix de ton abonnement ne change rien à l'affaire, la façon de t'en servir décide de tout. Applique ne serait-ce que les deux premiers réflexes et tu verras la différence dès cette semaine.

Questions fréquentes



Combien de messages un abonnement Claude à 20 euros

coûte-t-il vraiment



Environ 45 messages toutes les 5 heures, contre environ 900 sur l'abonnement à 200 euros. Ce nombre est un ordre de grandeur, pas un compteur : la limite réelle dépend de la taille de chaque échange, et une conversation longue la fait tomber bien avant le quarante-cinquième message.

Pourquoi ma limite tombe avant le nombre de messages annoncé

Parce que Claude relit toute la conversation à chaque message que tu envoies. Ton message ne part pas seul, il part avec l'historique complet de l'échange. Au trentième tour, tu paies surtout la relecture des vingt-neuf tours précédents, pas ta question du moment.

Faut-il couper la mémoire de Claude pour ménager son quota

Pas nécessairement. La mémoire coûte peu tant qu'elle reste courte et à jour, puisqu'elle est réinjectée au début de chaque conversation. Le vrai geste est le ménage mensuel : supprimer les souvenirs périmés, ou mettre la mémoire en pause pendant les périodes où elle ne sert à rien.

Quel modèle Claude choisir pour ne pas gaspiller son quota

Sonnet par défaut, parce qu'il couvre la majorité des tâches de bureau. Haiku pour tout ce qui est mécanique, comme reformuler, résumer ou trier. Opus uniquement quand la tâche résiste vraiment, c'est-à-dire sur du raisonnement complexe ou des arbitrages à enjeu.



Passer à 200 euros par mois est-il la bonne réponse



Rarement en premier. Un abonnement à 20 euros bien piloté fait le

travail d'un abonnement mal utilisé beaucoup plus cher, et les cinq réflexes ci-dessus se mettent en place en une semaine. L'abonnement supérieur devient pertinent quand tu tapes encore la limite alors que les réflexes sont déjà installés.

Blyx — cartographie des process & automatisation · blyx-agency.com

