12 MIN DE LECTURE · ÉDITÉ LE 19/05/2026 À 11H42

Vous brûlez vos tokens Claude sans le savoir

Guide pratique - 16 astuces pour ne plus jamais manquer de tokens en pleine session

Une production N-Gage - Mai 2026

Pourquoi ce guide

Vos crédits Claude s'évaporent avant la fin de votre session. Vous atteignez la limite à 15h alors que vous avez du travail jusqu'à 19h. Et vous regardez l'écran en vous demandant ce qui s'est passé.

Bienvenue dans la nouvelle réalité de Claude.

Depuis fin mars 2026, Anthropic a resserré ses fenêtres d'utilisation. La raison : la montée en puissance des usages agentiques (Claude Code, sessions longues, workflows complexes) qui sollicitent massivement les serveurs pendant les heures de pointe. Résultat : les abonnés Pro et Max atteignent leurs limites beaucoup plus vite qu'avant.

Mais voilà ce que peu de gens réalisent : la majorité des utilisateurs gaspillent leurs tokens sans s'en rendre compte. Pas parce que Claude est gourmand. Parce que la façon dont l'outil est utilisé est inefficace.

Ça se règle. Et ça prend dix minutes.

Les limites d'utilisation de Claude

Avant les astuces, il faut comprendre **comment Claude limite votre utilisation** selon votre abonnement. Beaucoup d'utilisateurs subissent leur limite sans jamais avoir compris la mécanique.

Le principe : sessions de 5 heures + limite hebdomadaire

Anthropic ne compte pas vos messages individuellement, ni vos tokens directement. Le système fonctionne avec **deux compteurs combinés** :

- **Limite glissante de 5 heures** : un compteur qui mesure votre consommation sur une fenêtre de 5h roulante. Quand vous atteignez le plafond, vous attendez que la jauge redescende. Anthropic communique officiellement une estimation d'**environ 45 messages par session de 5h sur Pro** - mais c'est valable pour des conversations courtes et peu intensives. Dans la vraie vie, le chiffre est très variable : un message de 2 lignes ne pèse pas pareil qu'un message de 50 lignes avec un PDF de 30 pages joint. Plus vos messages sont longs, plus vous joignez de fichiers, plus vous activez d'outils (Research, connecteurs), plus vous utilisez Opus, plus vous tapez la limite vite.
- **Limite hebdomadaire** (principalement sur les plans Max) : un plafond global qui se reset chaque semaine.

Concrètement : un fil de conversation long avec beaucoup de fichiers consomme votre budget bien plus vite qu'une dizaine de courtes questions ponctuelles.

Les plans Anthropic en mai 2026

Plan	Tarif (mensuel)	Limite d'usage	Modèles	Projects	Pour qui ?
Pro	20 $ (17 $ annuel)	~45 messages / 5h, reset toutes les 5h	Tous (Sonnet, Opus, Haiku)	Privés uniquement	Indépendant, dirigeant solo
Max 5x	100 $	5x plus que Pro + limite hebdo	Tous + accès prioritaire nouveaux modèles	Privés uniquement	Power user individuel
Max 20x	200 $	20x plus que Pro + limite hebdo	Identique Max 5x	Privés uniquement	Très gros usage individuel
Team Standard	25 $/siège (min 5 sièges)	1,25x Pro par membre	Tous	**Partageables** avec l'équipe	PME, équipes de 5+ personnes
Team Premium	125 $/siège	6,25x Pro par membre	Tous	Partageables	Équipes intensives
Enterprise	Sur devis	Configurable	Tous + API	Partageables + gouvernance (SCIM, audit, HIPAA)	Grandes organisations

Trois choses à retenir :

1. **Pour partager des Projects** entre collègues, il faut au minimum **Team**. Sur Pro et Max, vos Projects sont strictement personnels.
2. Les nouveaux **add-ins Office** (Excel, PowerPoint, Word) sont accessibles dès **Pro**. Word reste en beta avec liste d'attente sur Pro en mai 2026 - Team et Enterprise y ont accès complet.
3. **Claude Design** et les **tâches planifiées** sont disponibles sur tous les plans payés (Pro et au-delà).

Comment surveiller votre consommation

Anthropic propose un tableau de bord d'utilisation accessible dans vos paramètres de compte. Il affiche vos limites en temps réel selon les outils utilisés et vous indique combien de temps avant que la jauge redescende.

Question simple : est-ce que vous le consultez régulièrement ? Ou vous découvrez que vous êtes bloqué quand c'est déjà trop tard ?

Prenez l'habitude de le vérifier en début de journée si vous avez des sessions intensives prévues. Planifiez le travail lourd en dehors des heures de pointe si possible. Anticipez. Ne subissez pas.

Ce qui se passe quand vous atteignez la limite

- **Limite des 5h atteinte** : Claude vous affiche un message d'attente avec l'heure à laquelle votre quota se libère. Vous pouvez relancer une nouvelle session à ce moment-là.
- **Limite hebdomadaire atteinte (Max)** : c'est verrouillé jusqu'au reset hebdomadaire.

L'option extra usage : continuer après la limite

Depuis 2026, Anthropic propose une option **pay-as-you-go** pour ne pas être bloqué quand vous atteignez vos limites. C'est disponible sur les plans Pro, Max 5x et Max 20x.

Le principe : au lieu d'attendre le reset, vous pouvez **continuer à utiliser Claude aux tarifs API standard**, avec un compteur séparé qui s'incrémente à chaque message.

Comment l'activer :

1. Ajoutez un moyen de paiement dans votre compte
2. Fixez un **plafond mensuel** maximum pour ne pas perdre le contrôle (recommandé : commencez petit, par exemple 20 $)
3. Optionnel : activez le rechargement automatique avec un seuil (par exemple "recharger 10 $ quand le solde tombe sous 2 $")

Le plafond journalier maximum est de 2 000 $.

Bon réflexe : utilisez les extras pour les sessions critiques que vous ne pouvez pas reporter. Mais fixez un plafond serré (20-50 $/mois suffit pour la plupart des usages) pour ne pas dériver. Le pay-as-you-go peut vite coûter cher si vous ne fixez pas de limite.

Comprendre comment Claude compte

Un token, qu'est-ce que c'est ?

Un token est l'unité que Claude utilise pour découper et analyser le texte avant de le traiter. Concrètement, un token peut être :

- Un mot court entier ("chat", "bien", "non")
- Un fragment d'un mot plus long ("auto" + "matisation" = 2 tokens)
- Un signe de ponctuation, un chiffre, un emoji
- Un espace ou un saut de ligne

La règle de pouce pour vos calculs : en français, comptez en moyenne **1,3 à 1,5 token par mot**. Un message de 100 mots en français représente donc entre 130 et 150 tokens. C'est plus dense qu'en anglais, parce que les mots français sont souvent plus longs et le découpage est moins optimisé pour notre langue.

Vous n'avez pas à compter manuellement. Mais avoir cet ordre de grandeur en tête vous aide à comprendre pourquoi un long fil de conversation devient cher.

À tester : le Tokenizer Playground

Pour visualiser concrètement comment un texte est découpé en tokens, allez sur le **Tokenizer Playground** de Xenova :

> https://huggingface.co/spaces/Xenova/the-tokenizer-playground

Tapez n'importe quel texte (français, anglais, code, emoji...) et l'outil vous montre en temps réel :

- Comment le texte est découpé en tokens (chaque token est colorisé)
- Le nombre total de tokens
- Le nombre total de caractères
- Vous pouvez choisir différents tokenizers pour comparer (GPT-4, Llama, BERT...)

Petite précision : ce playground utilise les tokenizers de modèles open source, pas celui d'Anthropic Claude (qui n'est pas public). Le découpage exact peut donc varier de quelques pourcents avec celui que Claude utilise réellement. Mais le **principe** et les **ordres de grandeur** sont les mêmes : c'est parfait pour visualiser ce qui se passe.

Essayez avec une phrase en français comme "Le développement de l'intelligence artificielle générative" et regardez combien de tokens ça fait. Vous comprendrez immédiatement pourquoi un long fil consomme aussi vite.

La taille de la fenêtre de contexte selon les modèles

La **fenêtre de contexte**, c'est la quantité maximale de texte que Claude peut tenir en mémoire dans une conversation : votre historique de messages, vos fichiers joints, vos instructions personnalisées, le contenu d'un Project. Au-delà de cette fenêtre, Claude commence à oublier le début.

En mai 2026, voici les fenêtres officielles selon le modèle utilisé :

Modèle	Fenêtre de contexte	Équivalent en mots	Équivalent en pages A4
Opus 4.7 / 4.6	1 000 000 tokens (1M)	~750 000 mots	~1 500 pages
Sonnet 4.6	1 000 000 tokens (1M)	~750 000 mots	~1 500 pages
Sonnet 4.5 et antérieurs	200 000 tokens	~150 000 mots	~300 pages
Haiku 4.5	200 000 tokens	~150 000 mots	~300 pages

Concrètement avec Opus 4.7 ou Sonnet 4.6, vous pouvez littéralement charger l'équivalent d'un livre entier dans une conversation et poser des questions dessus. C'est ce qui permet les usages avancés : analyser un dossier juridique complet, croiser plusieurs textes réglementaires, traiter une base documentaire entière.

Mais attention au piège : plus la fenêtre est utilisée, plus chaque message coûte cher. Même une question courte force Claude à scanner potentiellement la fenêtre entière. C'est exactement le mécanisme expliqué ci-dessous, et c'est pour ça que les astuces "un sujet = un fil" et "videz le contexte régulièrement" restent valables même avec un modèle 1M tokens. Une grosse fenêtre n'est pas une excuse pour empiler.

Le mécanisme caché qui vous coûte cher

Claude relit l'intégralité de la conversation depuis le début à chaque nouveau message. Chaque. Nouveau. Message.

Votre premier échange dans la session coûte une poignée de tokens. Le trentième ? Il force Claude à avaler les vingt-neuf précédents avant de traiter votre question.

Faites le calcul. Une conversation de 30 échanges, avec des messages de taille moyenne, peut coûter dix à vingt fois plus que si vous aviez organisé vos échanges intelligemment. Ce n'est pas Claude qui devient cher, c'est votre session qui devient un boulet.

Et c'est là que ça relie aux limites vues plus haut : ce qui consomme votre budget de session, ce ne sont pas les messages comptés un par un. C'est le **travail réel** que Claude doit produire à chaque échange. Plus le contexte est gros, plus Claude travaille, plus la jauge descend vite.

C'est bien pour ça que la même conversation peut consommer trois fois plus chez votre voisin que chez vous, alors que vous avez le même abonnement.

Où passent vraiment vos tokens : la décomposition qui change tout

Voilà le chiffre qui surprend tout le monde : sur une conversation typique, **votre message lui-même ne représente que 3 % des tokens consommés**. Le reste, ce sont des coûts cachés que vous ne voyez pas passer.

Voici la répartition observée sur une session normale d'un utilisateur Pro :

Source de tokens	Part de la consommation
Historique de conversation (relu à chaque message)	**~51 %**
Système prompt + fichier de contexte (style, instructions)	~13-20 %
MCP serveurs activés (même non utilisés dans la réponse)	~13 %
Apps & outils connectés (Drive, Calendar, Slack…)	~6 %
Mémoire / souvenirs persistants	variable
Votre message + la réponse de Claude	**~3 %**

Trois enseignements à retenir :

1. **L'historique est le poste numéro 1**, et de très loin. Toutes les astuces qui le réduisent (un sujet = un fil, vider le contexte régulièrement, fiche projet de transition) sont les plus rentables.
2. **Les MCP et connecteurs activés coûtent même quand ils ne servent pas** dans la réponse. Claude doit charger leur description en contexte. C'est pour ça que désactiver ce qui ne sert pas est aussi efficace.
3. **Votre message ne représente quasiment rien**. Pas la peine de couper en deux ou trois prompts pour "économiser" : groupez au contraire dans un seul message bien structuré.

Les 16 astuces pour ne plus jamais manquer de tokens

1. Un sujet = un fil de conversation. Point.

C'est l'erreur numéro un. Vous commencez par une analyse de contrat, vous glissez vers un email de prospection, et vous terminez sur un débat stratégique. Tout dans la même fenêtre.

Vous venez de payer pour que Claude relise l'analyse de contrat à chaque fois que vous posez une question sur l'email de prospection.

Règle simple : dès que vous changez de sujet, vous ouvrez un nouveau fil. Pas de négociation.

2. Groupez vos questions en un seul message

"Tu peux m'aider avec X ?" puis "En fait voilà le contexte..." puis "Et j'aurais besoin que tu le fasses dans ce format..."

Trois messages. Trois relectures complètes de l'historique. Pour un résultat que vous auriez obtenu avec un seul message bien structuré.

Formatez vos questions en liste à puces si nécessaire. Vous obtenez exactement le même résultat pour un tiers de la consommation. C'est énorme.

3. Choisissez le bon modèle pour la bonne tâche

Opus est le modèle le plus puissant de Claude. C'est aussi le plus gourmand en ressources. L'utiliser pour trier des emails ou corriger un texte, c'est sortir une Porsche pour aller chercher le pain.

La logique à intérioriser :

- **Haiku** pour les tâches répétitives et courtes : extraction, résumé, reformatage, correction orthographique
- **Sonnet** pour l'immense majorité du travail professionnel : rédaction, analyse, code standard, recherche
- **Opus** pour ce qui justifie vraiment sa puissance : raisonnement complexe, code avancé, décisions stratégiques, création de Skills

Réservez Opus. Ne le gaspillez pas.

4. Désactivez les outils que vous n'utilisez pas, et chargez-les seulement quand c'est nécessaire

La recherche web, le mode Research, les connecteurs Slack ou Google Drive, les MCP tiers (Apollo, Pappers, datagouv...) : chacun consomme des tokens supplémentaires à chaque réponse, même quand vous n'en avez pas besoin. Anthropic le confirme dans sa documentation : ces outils sont particulièrement gourmands.

Le bon réflexe niveau 1 : tout désactiver par défaut. Tout activer ponctuellement quand la tâche le

Le bon réflexe niveau 2 (méconnu) : pour les outils que vous gardez activés, ouvrez le menu **Connecteurs** > **Accès aux outils**, et choisissez l'option **"Charger les outils si nécessaire"** plutôt que "Outils déjà chargés".

La différence : - **"Charger les outils si nécessaire"** (recommandé) : les outils ne sont chargés en contexte qu'au moment où Claude en a besoin. La conversation se compresse moins, vous économisez des tokens à chaque échange. - **"Outils déjà chargés"** : les outils sont préchargés en permanence. Claude y accède plus vite, mais le contexte est plus lourd à chaque message, donc votre quota descend plus vite.

Pour 95% des usages, "Charger si nécessaire" est le bon choix. Sauf si vous travaillez intensivement sur un workflow où vous savez d'avance que vous allez sans cesse appeler les mêmes outils.

5. Distinguez "Recherche" et "Recherche Web" - ce ne sont pas la même chose

Dans le menu "+" en bas de votre conversation, vous avez **deux boutons distincts** que beaucoup d'utilisateurs confondent :

- **Recherche Web** (icône globe) : mode **recherche ponctuelle**. Claude fait 1 ou 2 requêtes web pour répondre à une question factuelle, en quelques secondes. Coût modéré. Vous pouvez la laisser activée en permanence sans gros surcoût. *Exemples* : "quel est le tarif actuel de tel logiciel ?", "quelle est la météo à Toulouse ?", "trouve-moi le contact de cette personne sur LinkedIn".
- **Recherche** (icône loupe) : mode **Research / Recherche approfondie**. Active une vraie investigation : Claude planifie ses recherches, enchaîne 5+ requêtes web, analyse les sources, et produit un **rapport structuré** en 1 à 3 minutes. Active automatiquement aussi le **raisonnement étendu** (extended thinking). Coût élevé en tokens. *Exemples* : "fais-moi une analyse comparative des SaaS RH du marché français", "rapport sur la concurrence dans tel secteur", "étude de marché pour tel produit".

En clair : Recherche Web = chercher une info ponctuellement / Recherche = lancer une vraie étude. Pas le même usage, pas le même coût.

Le bon réflexe : Recherche Web peut rester activée par défaut. N'activez **Recherche** que ponctuellement, pour les vraies études de fond qui le justifient. Sinon, vous brûlez votre quota inutilement.

Note : il existe aussi un mode **raisonnement étendu seul** (extended thinking sans web), accessible via certains réglages avancés. Idéal pour les problèmes qui demandent de la réflexion sans avoir besoin d'infos web (logique, math, analyse stratégique sur des données déjà fournies).

6. Exploitez les Projects pour vos ressources récurrentes

Uploader le même document dans dix conversations différentes, c'est le faire lire dix fois.

Les Projects Claude règlent ce problème : un fichier uploadé une fois est mis en cache et reste disponible pour toutes les conversations du Project, sans token supplémentaire à chaque session.

Si vous avez une base documentaire que vous sollicitez régulièrement (grilles tarifaires, textes réglementaires, templates internes, contexte client), les Projects sont faits pour vous. Pas pour les usages ponctuels. Pour le récurrent.

7. Surveillez votre consommation avant de manquer d'air

Anthropic propose un tableau de bord d'utilisation accessible dans vos paramètres de compte. Il affiche vos limites en temps réel selon les outils utilisés.

Question : est-ce que vous le consultez régulièrement ? Ou vous découvrez que vous êtes bloqué quand c'est déjà trop tard ?

Prenez l'habitude de le vérifier en début de journée si vous avez des sessions intensives prévues. Planifiez le travail lourd en dehors des heures de pointe si possible. Anticipez. Ne subissez pas.

8. Convertissez vos fichiers texte avant de les uploader

Pour un PDF essentiellement textuel, extrayez le texte et collez-le dans un fichier Markdown ou texte brut avant d'uploader. Claude est facturé deux fois sur un PDF natif : extraction du texte ET analyse visuelle page par page.

Pour un PDF avec graphiques, tableaux complexes ou mise en page importante, gardez le PDF natif : l'analyse visuelle est utile.

Même logique pour les captures d'écran : quand l'information est textuelle, copiez-collez plutôt que de capturer.

Le gain réel : un PDF de 15 pages textuelles converti en Markdown passe typiquement de 6 000-7 000 tokens à 2 500-3 000 tokens en entrée. Soit une **division par 2 à 3 du coût d'analyse**, à chaque fois que vous interrogez le document. Si vous travaillez régulièrement avec ce document (Project, conversations multiples), l'économie est massive.

Trois méthodes simples pour convertir vos PDF en Markdown (gratuites, sans inscription) :

- **pdf2md.morethan.io** : outil dédié, glissez-déposez le PDF, conversion instantanée dans le navigateur, vous récupérez un fichier `.md` propre. Aucun envoi de fichier sur leurs serveurs - tout se fait en local. Parfait pour les documents confidentiels.
- **copymarkdown.com/pdf-to-markdown** : alternative équivalente, le fichier est supprimé immédiatement après conversion. Interface très épurée.
- **Via ChatGPT (gratuit) ou Gemini** : uploadez le PDF, demandez "Convertis ce PDF en Markdown propre, structuré avec des titres et sous-titres". Plus lent mais utile si le PDF a des tableaux ou une structure complexe que l'IA peut mieux interpréter qu'un convertisseur automatique.

À retenir : pour un document de référence que vous allez interroger plusieurs fois, **investissez 30 secondes à le convertir une fois pour toutes**. L'économie de tokens se rentabilise dès la deuxième utilisation.

9. Modifiez votre requête, ne la corrigez pas dans le fil

Ce message s'empile dans l'historique et sera relu intégralement à chaque échange suivant.

Le bon réflexe : cliquez sur le crayon à côté de votre message initial. Modifiez-le directement. Envoyez à nouveau. L'échange est remplacé, pas empilé. Zéro surcoût, même résultat.

C'est une petite habitude qui change beaucoup sur des sessions longues.

10. Clôturez intelligemment vos sessions longues - la règle des 15-20 messages

C'est l'astuce la plus importante du guide. Si vous n'en retenez qu'une, retenez celle-là.

Règle simple : au-delà de 15 à 20 messages dans une même conversation, le coût explose. C'est le seuil à partir duquel l'effet boule de neige devient vraiment douloureux. Au 30e message, vous payez environ 30 fois ce que coûtait votre 1er message.

Le bon réflexe : avant de dépasser ce seuil, demandez à Claude :

> Résume tout ce que nous avons fait d'important dans cette conversation sous la forme d'une fiche projet de 300 à 400 tokens : objectif, décisions prises, prochaines étapes, points en suspens. Format Markdown.

Vous récupérez une fiche compacte de quelques lignes. Vous ouvrez une nouvelle conversation. Vous collez la fiche en premier message. Vous repartez sur une ardoise quasi-vierge avec tout le contexte essentiel - mais sans traîner 25 messages d'historique.

C'est radical, c'est efficace, et ça change littéralement la façon dont vous travaillez sur des projets longs.

Le bonus instructions personnalisées : ajoutez dans vos instructions globales (Paramètres > Profil) une consigne du type :

> Quand tu détectes qu'une conversation devient longue (15-20 messages échangés), suggère-moi spontanément de créer une fiche projet de transition pour démarrer une nouvelle conversation. Tu peux me la générer à la demande.

Claude prendra l'initiative à votre place quand le moment sera venu.

Pour les artefacts : générez vos documents Word, présentations, tableaux en **fin de session** plutôt qu'en cours de route. La génération de fichiers est coûteuse en tokens : mieux vaut affiner le contenu en mode conversation d'abord, puis déclencher la création en une seule fois à la fin.

11. Activez la mémoire de Claude, mais maîtrisez ce qu'elle alimente

Claude peut accéder à vos conversations passées pour récupérer du contexte, ce qui vous évite de répéter les mêmes informations à chaque nouvelle session. Deux fonctionnalités dans Paramètres > Fonctionnalités :

- **La recherche dans les conversations passées** : vous pouvez demander explicitement à Claude de retrouver ce qui a été discuté lors d'échanges précédents.
- **La mémoire contextuelle** : Claude retient automatiquement les informations clés d'une session à l'autre.

Mais attention au piège : il existe une option "**générer de la mémoire à partir de l'historique des conversations**". Activée, Claude alimente automatiquement sa mémoire avec ce qu'il extrait de toutes vos sessions. Résultat : votre mémoire enfle au fil du temps, et chaque nouveau message paie le coût de la lire intégralement.

Le bon réflexe : désactivez cette génération automatique. Préférez écrire vous-même, une fois pour toutes, dans **Paramètres > Profil > Instructions personnalisées**, le contexte qui vous concerne :

- Votre métier, votre secteur, vos clients types
- Votre style d'écriture préféré (concis, paragraphes courts, ton décontracté...)
- Vos contraintes (langue, format, ce que Claude ne doit jamais faire)

Une mémoire **écrite proprement par vous** est beaucoup plus dense et plus utile qu'une mémoire générée automatiquement à partir de l'historique. Et elle ne grossit pas.

Bon à savoir : les conversations issues des Projects ne sont pas intégrées à la mémoire globale. Elles ont leur propre espace. La mémoire globale ne couvre que les conversations hors Projects.

12. Bonus - Réduisez la verbosité de Claude dans vos instructions personnalisées

Par défaut, Claude répète souvent le contexte avant de répondre, donne des explications longues, propose plusieurs options. C'est pédagogiquement utile, mais ça consomme des tokens en sortie.

Ajoutez dans vos instructions personnalisées (Paramètres > Profil > Instructions personnalisées) : "Réponds de manière concise. Va droit au but. Ne répète pas le contexte de ma question dans ta réponse. Pose une question si tu as besoin de plus d'info plutôt que de proposer 3 versions."

Gain typique : 30 à 50% de tokens en sortie en moins. Sur des sessions longues, c'est massif.

13. Bonus - Videz le contexte régulièrement, même sur le même thème

L'astuce 1 dit "un sujet = un fil". Mais on peut aller plus loin : même sur un même sujet, dès que vous changez de tâche concrète, ouvrez un nouveau fil.

Exemple : vous travaillez sur la rédaction d'une proposition commerciale. Vous avez fini la partie "contexte client". Vous attaquez la partie "chiffrage". Ouvrez un nouveau fil et collez juste les éléments dont vous avez besoin. Vous évitez que Claude relise toute la partie contexte chaque fois que vous itérez sur le chiffrage.

C'est un peu contre-intuitif. Mais le gain est très significatif sur des projets longs.

14. Cowork : planifiez dans le chat, exécutez dans Cowork

Claude Cowork (les espaces multi-agents et tâches longues) est la solution Claude qui consomme le plus de tokens. Logique : elle fait tourner plusieurs agents en parallèle, gère des tâches complexes, et son travail de planification est lourd.

L'erreur classique : ouvrir Cowork et lui demander directement "Crée-moi un rapport complet sur X avec un Excel et un PowerPoint à la fin". Cowork va planifier, structurer, déléguer, itérer... et brûler vos tokens à une vitesse folle, surtout si vous changez d'avis en cours de route.

Le bon réflexe : faites toute la **phase de planification dans le chat classique** (avec Sonnet, peu coûteux). Demandez :

> Quel serait pour toi le meilleur plan pour produire un rapport sur le modèle financier de X, avec un fichier Excel et un PowerPoint en livrables ? Donne-moi les étapes détaillées et la structure attendue.

Vous itérez sur le plan, vous l'ajustez, vous le validez. Tout ça dans le chat, donc à coût modéré.

Une fois le plan stabilisé, **vous le copiez** et **vous le collez dans Cowork** comme premier message en disant "Voici la planification validée, exécute-la". Cowork ne paie alors plus que la production - pas la planification ni les itérations.

Le gain est massif. Sur un projet complexe, on parle facilement de **moitié de tokens économisés** dans Cowork.

15. Laissez Claude vous poser ses questions plutôt que rédiger un prompt parfait

Vous voulez écrire un prompt impeccable, alors vous mettez 8 paragraphes de contexte au cas où. Résultat : vous gonflez votre input de tokens dont la moitié est inutile, et Claude se perd parfois dans le contexte.

Le bon réflexe : commencez par dire ce que vous voulez en quelques mots, puis demandez à Claude :

> De quoi as-tu besoin comme informations pour me produire la meilleure réponse possible ?

Claude va vous **poser ses questions ciblées** (3 à 5 questions précises selon ce qu'il a vraiment besoin de savoir). Vous répondez juste à ces questions. Vous évitez le contexte inutile, Claude obtient exactement ce qu'il faut, le résultat est meilleur, et vous économisez des tokens d'input et d'aller-retours.

Idéal pour : rédactions complexes (emails sensibles, propositions commerciales, posts LinkedIn), analyses spécifiques, briefs créatifs.

16. Bonus avancé - Le skill "Caveman" pour diviser par 4 la verbosité de sortie

Pour les utilisateurs avancés (qui connaissent les Skills) : un skill open source baptisé **Caveman** ("homme des cavernes") force Claude à répondre dans un style ultra-condensé, façon télégramme

gain d'environ **75 % sur la sortie**.

Le style reste lisible (pas du langage codé). C'est juste Claude qui supprime tous les mots de remplissage : "OK voici la liste", "Bien sûr, je peux faire ça", "Cela représente plusieurs avantages"... Tout ce qui n'apporte rien à l'information.

À utiliser quand : vous voulez juste l'info, pas l'enrobage. Tâches répétitives (extraction de données, listes, transformations). Sessions longues où vous voulez économiser au maximum.

À éviter quand : vous avez besoin d'explications pédagogiques, de nuances, ou de réponses qui se lisent bien comme un texte humain.

Pour l'installer : Cloud > Créer un plugin > Ajouter depuis la marketplace > coller l'URL du repo GitHub Caveman > synchroniser. Le skill devient activable comme n'importe quel autre.

Ce qu'il faut retenir

Vous n'avez pas un problème de limite. Vous avez un problème d'organisation.

Les limites d'Anthropic ont changé, oui. Mais la plupart des blocages que je vois chez mes clients viennent de pratiques inefficaces, pas d'un manque de crédits. Un fil de conversation bien géré consomme trois à cinq fois moins qu'un fil mal géré sur le même sujet.

Ces 16 astuces ne demandent aucun abonnement supplémentaire. Aucun outil externe payant. Juste des réflexes à intégrer une bonne fois pour toutes.

Appliquez-en cinq dès aujourd'hui (la règle des 15-20 messages, la conversion PDF en Markdown, la désactivation des outils non utilisés, le bon choix de modèle, les instructions personnalisées concises). Vous sentirez la différence avant la fin de la semaine.

Aller plus loin avec N-Gage

N-Gage forme et accompagne les dirigeants et les équipes à utiliser Claude et l'IA générative au quotidien, sans jargon, avec des cas d'usage concrets et durables.

- **Site** : https://n-gage.fr
- **Contact** : contact@n-gage.fr
- **LinkedIn** : Nicolas Graillot

Si ce guide vous a aidé, partagez-le. Si vous voulez aller plus loin sur Claude (Projects, Skills, MCP, Connecteurs, Claude Design…), parlons-en.