ELINKCAT

Estalvia tokens amb Claude

Tot el que necessites saber sobre tokens, context i bons hàbits per treure el màxim a la teva subscripció.

Per Òscar Junyent

Juny 2026

CONTINGUT

En aquesta guia

CAPÍTOL 01

El missatge que ningú vol veure

Ens ha passat a tots. Estàs enmig d'alguna cosa important, analitzant un informe, redactant una proposta, completament concentrat, i de cop apareix el missatge: «Has arribat al límit d'ús. Torna d'aquí a cinc hores.»

Frustració total i, sobretot, desconcert, perquè portes mesos usant ChatGPT i això mai t'havia passat. Hi ha dues notícies, una de bona i una de dolenta.

La dolenta: els límits són reals i hi seran sempre. No són un caprici comercial, sinó un reflex del cost físic de processar informació en un model de llenguatge. Processar tokens consumeix GPUs, memòria i electricitat.

La bona: la gran majoria de vegades que xoquem contra aquests límits no és per culpa del pla contractat, sinó per com estem usant l'eina. Entenent-ho bé, es poden evitar quasi sempre.

IDEA CLAU

El problema gairebé mai és el pla. És l'arquitectura de com uses Claude. Aquesta guia t'ensenya a canviar-la.

PUNTS CLAU

- La major part del consum innecessari ve de tres fonts: no separar converses per tema, no haver configurat bé les capes estables i arrossegar converses que haurien d'haver-se compactat fa temps.
- Entendre com funciona Claude per dins és el primer pas per usar-lo millor.
- Un usuari optimitzat pot obtenir entre el 50% i el 70% més de feina útil amb la mateixa subscripció.

CAPÍTOL 02

Per què Claude talla l'accés (i ChatGPT no)

Una pregunta freqüent entre usuaris nous: «Amb ChatGPT mai em passa això.» Cert, però no perquè ChatGPT sigui més generós.

Quan t'acostes als límits reals de ChatGPT, la conversa simplement continua, però sota condicions molt diferents: el sistema comença a resumir parts anteriors, elimina fragments de l'historial o redueix el nivell de detall. La qualitat no es trenca de cop, però es va degradant de manera silenciosa.

Claude fa exactament el contrari: no amaga el moment en el qual el sistema ja no pot sostenir tot el context amb la mateixa qualitat. Et talla l'accés en lloc de degradar la qualitat. És més honest, però desorientador si no ho saps.

PER RECORDAR

Claude no és menys generós — és més transparent. La degradació silenciosa de ChatGPT i el tall explícit de Claude responen al mateix problema: massa context per processar.

PUNTS CLAU

- Claude prefereix tallar l'accés abans que lliurar respostes de qualitat inferior.
- ChatGPT degrada en silenci: segueix funcionant però perd context, precisió i coherència.
- La idea central d'aquesta guia: tot gira al voltant de gestionar bé el context.

CAPÍTOL 03

Tokens: la unitat real de mesura

Claude no llegeix paraules. Llegeix tokens: fragments de text que poden ser una paraula sencera, una part de paraula, una combinació freqüent de dues paraules curtes o fins i tot un signe de puntuació.

El criteri no és gramatical, és estadístic. Durant l'entrenament, el model va aprendre quines seqüències de caràcters apareixen juntes amb molta freqüència i les tracta com una sola peça.

Un detall important: el recompte en tokens no es correspon amb la teva intuïció de paraules. Un text en català o castellà pesa més en tokens que el mateix text en anglès, perquè els tokenitzadors es van optimitzar originalment per a l'anglès.

Un paràgraf	150 – 200 tokens
Una pàgina	600 – 700 tokens
PDF de 10 pàgines	5.000 – 8.000 tokens
PDF de 50 pàgines	25.000 – 40.000 tokens
Llibre 200 pàgines	100.000 – 150.000 tokens

PUNTS CLAU

- Una pàgina en català ≈ 600–700 tokens; en anglès, uns 400–500.
- Markdown net és significativament més lleuger que un PDF escanejat.
- Els emojis, les taules mal formatades i les imatges pesen molt més del que sembla.

CAPÍTOL 04

El context: tot el que Claude té a la taula

El context és tot el que Claude té davant seu en el moment de respondre. No és només el missatge que acabes d'escriure, inclou l'historial complet, els documents pujats, les instruccions del sistema, les teves preferències i qualsevol eina connectada.

Quan envies un missatge, Claude no processa només aquell missatge. Processa tot un paquet. Sumant-ho tot, un primer missatge arrenca fàcilment prop dels 25.000 tokens abans que Claude hagi processat ni tan sols la teva pregunta.

Sis de les set capes les controles tu. Aquí és exactament on es troben totes les palanques d'estalvi d'aquesta guia.

> LA CLAU
>
> **Sis de les set capes del context les controles tu. Aquí es troba tot el potencial d'optimització.**

Instruccions Anthropic	~4.000 tokens · No controlable
Preferències d'usuari	900 – 3.200 · Sí
Memòria	600 – 1.200 · Sí
Connectors MCP actius	5.000 – 18.000 · Sí
Historial de conversa	creix cada torn · Sí
El teu missatge + fitxers	variable · Sí

CAPÍTOL 05

Per què les converses llargues costen més

Deu torns de conversa no costen deu vegades més que un de sol. Costen molt més, perquè en cada torn s'inclou tot l'anterior. El consum no creix de forma lineal, sinó que s'acumula sobre si mateix.

Hi ha un segon motiu: la qualitat es degrada amb el context enorme i redundant. Quan l'atenció ha de repartir-se entre 100.000 tokens en lloc de 5.000, cada token rep menys focus. Els investigadors anomenen aquest fenomen «lost in the middle».

SENYAL D'ALERTA

A partir dels 60–80 torns: Claude oblida detalls, les respostes es tornen genèriques, apareixen contradiccions, les respostes tarden més. Compacta i recomença.

PUNTS CLAU

- Una conversa de 50.000 tokens no costa cinc vegades més que una de 10.000 — costa de l'ordre de vint-i-cinc vegades més.
- Quan la conversa sembli confusa, no insisteixis reformulant. Compacta amb un resum i obre un xat nou.
- Cada token de sortida avui és context que arrossegues en els torns de demà.

CAPÍTOL 06

Dos límits, dos problemes

Aquesta és probablement la confusió més cara que existeix entre usuaris de Claude. Hi ha dos límits independents, amb causes i solucions distintes.

Finestra de context (200.000 tokens): mida màxima de la conversa activa. Quan s'omple, els detalls antics es resumeixen i es perden matisos. Es reinicia en obrir un xat nou.

Límit d'ús (missatges per 5 hores): quants missatges pots enviar en una finestra temporal. No depèn de la mida de les converses, depèn del temps.

REGLA RÀPIDA

Si Claude respon malament → finestra de context. Si directament no pots escriure → límit d'ús. Dues causes, dues solucions.

Finestra de context	Es buida quan la conversa és molt llarga. Solució: compactar i obrir xat nou.
Límit d'ús	Es buida quan envies massa missatges en 5h. Solució: esperar o optimitzar.

CAPÍTOL 07

Plans: el que tens i el que pots esperar

Tots els plans, Free, Pro i Max, comparteixen exactament la mateixa finestra de context de 200.000 tokens. La diferència entre plans és la quantitat de missatges per hora, no la mida de les converses.

Max no et dóna més context, et dóna més missatges. Si un PDF no t'hi cap en Pro, tampoc hi cabrà en Max. La solució és altra: dividir-lo, convertir-lo a Markdown o ficar-lo en un Project.

Les hores també importen: en hores punta (13h–19h dies laborables) els límits es noten més estrictes. Claude Code i Claude.ai comparteixen la mateixa quota.

ABANS DE PUJAR DE PLA

La diferència entre un usuari optimitzat i un de caòtic pot ser de 3–4x més feina útil. Moltes vegades, «necessito el Max» és en realitat «necessito ordenar com uso el Pro».

Free	200.000 tokens · 15–40 missatges/5h
Pro · 20$/mes	200.000 tokens · ~45 missatges/5h
Max 5x · 100$/mes	200.000 tokens · ~225 missatges/5h
Max 20x · 200$/mes	200.000 tokens · ~900 missatges/5h

CAPÍTOL 08

Les tres regles d'or

Regla 1 · Treballa en ràfegues aprofitant el caché. Claude desa temporalment les capes estables del context. Si en els propers 5–10 minuts envies un altre missatge sense que res hagi canviat, Claude no ho torna a calcular. Un primer missatge: ~25.000 tokens. Missatges en ràfega: 500–1.500 tokens cadascun.

Regla 2 · Cada informació viu en un sol lloc. Instruccions estables → Preferències. Patrons → Memòria. Processos repetitius → Skills. Contingut del projecte → Project. Informació puntual → missatge directe.

Regla 3 · Gestiona el context, no les paraules. L'optimització real no és escriure missatges curts, és prendre decisions d'arquitectura: quan obrir un xat nou, quan compactar, quan desconnectar un MCP.

REGLA DE BUTXACA

Els usuaris avançats no escriuen missatges més curts — gestionen millor l'arquitectura. Xats especialitzats per tema, resums intermedis, Projects ben estructurats, context net.

PUNTS CLAU

- Tens diverses preguntes sobre el mateix tema? Fes-les totes seguides, sense pauses llargues, mentre el caché és calent.
- Obre un xat nou quan canvia el tema. Sempre, sense excepcions.
- Compacta i recomença quan la conversa es faci llarga o confusa.

CAPÍTOL 09

Preferències i Memòria: les capes que ningú audita

Les Preferències d'usuari es carreguen completes a l'inici de cada conversa, parlis del que parlis. Unes preferències sense auditar poden pesar 3.200 tokens de més per missatge. En vint missatges diaris durant un mes, són més d'un milió de tokens desperdiciats.

Què hi posem: qui ets professionalment, idioma, to, format. Què no hi posem mai: detalls de projectes concrets, instruccions puntuals, llistes llargues d'excepcions.

La Memòria és útil quan recull patrons reals i estables. Quan s'omple de projectes tancats i dades obsoletes, es converteix en pes mort. Quinze minuts al mes per revisar-la i netejar-la marquen la diferència.

IMPACTE REAL

Diferència entre preferències optimitzades (~900 tokens) i sense auditar (3.200+): 2.300 tokens de més per missatge. En 20 missatges diaris durant un mes: >1 milió de tokens desperdiciats.

PUNTS CLAU

- Revisa les teves Preferències ara: elimina tot allò que no sigui informació estable sobre tu.
- Quinze minuts al mes per revisar la Memòria i netejar projectes tancats.
- No duplis: si quelcom ja és a les Preferències, no el posis també a la Memòria.

CAPÍTOL 10

Connectors MCP: el cost invisible

Cada connector MCP actiu, Google Drive, Notion, Slack, GitHub..., afegeix entre 5.000 i 18.000 tokens a cada missatge, l'usis o no en aquella conversa. El sistema carrega l'esquema complet de l'eina al principi de cada missatge.

Sis connectors actius «per si de cas» poden suposar fins a 108.000 tokens addicionals per missatge. En vint missatges al dia, parlem de més de 2 milions de tokens setmanals dedicats a connectors que potser no has obert.

ACCIÓ IMMEDIATA

Obre la configuració de connectors ara i desconnecta tots els que no uses activament aquesta setmana. Reconnectes quan els necessitis: triga menys de trenta segons.

EXEMPLE

connectors.txt

```
# Cost acumulat amb 6 connectors actius
Drive:   ~8.000 tokens / missatge
Notion:  ~6.500 tokens / missatge
Slack:   ~7.000 tokens / missatge
GitHub:  ~9.000 tokens / missatge
# TOTAL: ~44.000 tokens addicionals / missatge
# En 20 missatges/dia → ~880.000 tokens extra/dia
```

CAPÍTOL 11

Projects: la palanca més potent

Un Project no és una carpeta de converses. És una unitat arquitectònica que combina: una base de coneixement cacheada (fitxers processats una sola vegada), instruccions que s'apliquen automàticament a totes les converses, i un espai aïllat on el context no es barreja.

Quan puges un document a un Project, Claude el processa complet la primera vegada. Cada accés posterior des del caché costa aproximadament el 10% del cost original. El punt de rendibilitat s'assoleix a la segona consulta.

> REGLA DE BUTXACA
>
> **Si un fitxer l'has consultat (o el consultaràs) més d'una vegada, el seu lloc és un Project. Cada cop que el puges a un xat nou sense Project, el tornes a pagar sencer.**

Sense Project	Document 30.000 tokens × 5 consultes = 150.000 tokens
Amb Project	~35.000 tokens en total (estalvi del 77%)

PUNTS CLAU

- Un Project ben estructurat pot reduir el cost de treball regular en un 70–90%.
- Posa noms descriptius: informe-vendes-Q2-2026.pdf, no document1.pdf.
- Un Project, un propòsit. Millor molts Projects petits que pocs de grans i caòtics.

CAPÍTOL 12

RAG: quan el Project és més gran que la finestra

Quan el teu Project conté més informació de la que cap en la finestra de 200.000 tokens, entra en joc el RAG (Retrieval Augmented Generation). En lloc de carregar-ho tot, el sistema cerca de forma intel·ligent quins fragments són rellevants per a cada pregunta.

La cerca no és per paraules exactes, sinó per significat. Una pregunta sobre rotació de personal pot recuperar un fragment que parlava d'abandó d'empleats, perquè tots dos estan a prop semànticament.

RAG s'activa automàticament quan la base de coneixement supera el límit de context. No cal configurar-lo, però la qualitat dels resultats depèn directament de com tens organitzats els documents.

REGLA DE BUTXACA

Si entren escombraries, surten escombraries. Noms de fitxer descriptius, contingut ben estructurat amb encapçalaments clars, fitxers agrupats temàticament i preguntes específiques.

PUNTS CLAU

- Converteix PDFs escanejats a Markdown abans de pujar-los: pot reduir el pes a la meitat.
- Si RAG retorna resultats confusos, el problema és quasi sempre la manca d'estructura als documents.
- Millora el cache hit rate mantenint els documents ordenats i les preguntes concretes.

CAPÍTOL 13

Skills: processos reutilitzables

Un Skill és una instrucció que defineixes una vegada i Claude aplica cada cop que la crides. Si tens processos que repeteixes amb regularitat, un format d'informe, un protocol de resposta, una estructura d'anàlisi, converteix-los en Skills.

Descriure el format cada vegada costa ~200 tokens per torn. Cridar un Skill costa ~5 tokens. En vint usos del procés, la diferència és de 3.900 tokens. Amb deu processos repetitius ben convertits, l'estalvi mensual és considerable.

QUAN CREAR UN SKILL

Si fas quelcom cinc o més vegades al mes i sempre vols el mateix format o enfocament, és un Skill. El temps d'escriure'l una primera vegada s'amortitza a la tercera vegada que l'uses.

PUNTS CLAU

- Identifica els teus tres processos més repetitius i converteix-los en Skills aquesta setmana.
- Un Skill no és un prompt llarg — és una definició guardada que es crida amb una sola frase.
- Audita els teus Skills cada mes: elimina els que ja no uses i actualitza els que han canviat.

CAPÍTOL 14

Bones pràctiques de conversa

Separa converses per tema. Cada cop que canviïs d'assumpte, obre un xat nou. Tot el context del tema anterior continuarà viatjant amb tu, tot i que no tingui res a veure. Un xat, un tema. Sense excepcions.

Agrupa preguntes relacionades. Quatre dubtes sobre el mateix tema no són quatre missatges, sinó un amb quatre preguntes numerades. Pagues el context una sola vegada.

Edita en lloc de corregir. «No, fes-ho així» afegeix dos torns a l'historial. L'alternativa: editar el missatge original (icona del llapis) i deixar que Claude regeneri. La conversa es manté neta.

Especifica sempre la longitud de resposta. «Resposta màxima de 200 paraules» o «llista breu, sense introducció». Cada token de sortida avui és context que arrossegues demà.

REGLA DE BUTXACA

Trenta segons pensant abans d'escriure t'estalvien minuts d'anada i tornada i, sobretot, diversos torns que es queden a l'historial pesant.

PUNTS CLAU

- Evita la cortesia superflua: «gràcies», «entès», «perfecte» són torns sencers que es queden a l'historial.
- No tractis Claude com un company de Slack. Directe, clar, sense floritures.

CAPÍTOL 15

Els errors que tothom comet

Error 1: Enganxar de nou contingut que ja és a la conversa. Claude veu tots els torns anteriors. Repetir un text el duplica a l'historial, i aquell duplicat viatjarà dues vegades en cada torn futur.

Error 2: Pujar el mateix fitxer a diverses converses. Sense Projects, Claude processa el fitxer sencer cada vegada. Cinc converses = cinc processos complets.

Error 3: Connectors MCP actius per si de cas. Cada MCP connectat pesa 5.000–18.000 tokens per missatge, l'usis o no.

Error 4: Estirar converses fins a l'infinit. Una conversa de 80 torns no sols costa molt, sinó que funciona pitjor.

EL COST REAL

El consum innecessari és acumulatiu i silenciós. Errors 1–4 junts poden representar el 40–60% del teu consum total.

PUNTS CLAU

- Error 5: no definir longitud de sortida → resposta de 1.500 tokens quan en bastaven 300.
- Error 6: pujar PDFs escanejats sense convertir → el doble del pes necessari.
- Error 7: acumular Memòria sense auditar → pes mort en cada missatge.

CAPÍTOL 16

Casos pràctics: caòtic vs. optimitzat

Cas 1, Anàlisi recurrent de reviews. Usuari caòtic: 8.000–12.000 tokens/sessió, 60.000 tokens/setmana, 40% instruccions repetides. Usuari optimitzat (Project + instruccions base): 2.500–3.500 tokens/sessió. Estalvi del 70% per exactament la mateixa feina.

Cas 2, Contingut per a client recurrent. Usuari caòtic: missatge inicial de 2.000 tokens de context explicant el client cada cop. Usuari optimitzat (Project + Skill): missatge inicial de 150 tokens.

Cas 3, Recerca tècnica complexa. Usuari caòtic: 70 torns, Claude comença a contradir-se. Usuari optimitzat: als 40 torns compacta amb un resum executiu, obre un xat nou dins del mateix Project i continua amb context fresc conservant tot l'avenç.

LA DIFERÈNCIA CLAU

La diferència entre el 70% d'estalvi i el 0% gairebé mai és tecnologia — és disciplina d'arquitectura. Xats per tema, Projects per fitxers recurrents, Skills per processos repetitius.

CAPÍTOL 17

Monitoratge: com saber si ets eficient

No pots optimitzar el que no mesures. Claude ofereix un panell d'ús accessible des del teu perfil on pots consultar els missatges consumits en la finestra de cinc hores i els consumits a la setmana.

BANDERES VERMELLES

Mateix Project en 10 converses idèntiques → consolida. Toppes el límit cada dia → optimitza l'arquitectura. Conversa oberta diverses setmanes → compacta i obre un xat nou.

Missatges/setmana	Bo: 50–150 · Alerta: >300 en Pro sense Projects
Durada de conversa	Bo: 10–30 torns · Alerta: >60 sense compactar
Límit assolit	Bo: ocasionalment · Alerta: cada dia a mig matí

PUNTS CLAU

- Revisa el panell d'ús ara i anota els números. Els necessitaràs per comparar en dues setmanes.
- Les hores a les quals arribes al límit et diuen molt sobre l'origen del problema.

CAPÍTOL 18

Pla d'acció: per on començar

No cal canviar-ho tot de cop. Un ordre raonable, repartit en fases, dóna resultats molt més sòlids.

Aquesta setmana (impacte immediat): desconnecta connectors MCP inactius · audita les Preferències · crea un Project per als tres documents més consultats · revisa la Memòria.

El mes vinent: converteix els tres processos més repetitius en Skills · separa converses per tema sistemàticament · afegeix instruccions de longitud als prompts habituals.

De manera continuada: 15 minuts/mes per auditar Memòria, Preferències i Skills · revisar connectors actius cada setmana · consultar el panell d'ús setmanalment.

RESULTAT ESPERAT

Un usuari que aplica aquestes pràctiques pot obtenir entre un 50% i un 70% més de feina útil amb la mateixa subscripció. En molts casos, «necessito el Max» és en realitat «necessito ordenar com uso el Pro».

PUNTS CLAU

- Desconnecta connectors MCP inactius → fins a 18.000 tokens estalviats per missatge.
- Treballa en ràfegues (caché actiu) → cost 10–20x menor per missatge.
- Crea Projects per a fitxers recurrents → reducció de cost del 70–90% per document.

CAPÍTOL 19

Tot en un cop d'ull

La síntesi de totes les estratègies de la guia, per tenir-les a mà en un sol lloc.

1 · Connectors MCP	Desconnecta els inactius → fins a 18.000 tokens/missatge
2 · Preferències	Neteja i audita → fins a 2.300 tokens/missatge
3 · Projects	Per fitxers recurrents → estalvi del 70–90%
4 · Caché en ràfegues	Preguntes seguides → cost 10–20x menor
5 · Memòria	Audita cada mes → elimina pes mort
6 · No duplicar	No repeteixis el que ja és al context
7 · Edita, no corregeixis	Manté l'historial net i lleuger
8 · Longitud de resposta	Especifica sempre → menys context futur
9 · Compacta	Als 60–80 torns, resumeix i recomença
10 · Skills	Processos repetitius → 200 → 5 tokens/ús

PUNTS CLAU

- Comença pels punts 1, 2 i 3 — l'impacte és immediat i es nota en 24 hores.
- Els punts 4, 7 i 8 són canvis d'hàbit que costen poc però sumen molt al mes.
- Els punts 3, 9 i 10 requereixen setup inicial però estalvien molt a llarg termini.

ELINK CAT

Formació i consultoria IA propera i de confiança.

Ajudem empreses i professionals a entendre, adoptar i treure profit real de la intel·ligència artificial.

Sense fum. Sense tecnicismes innecessaris.
Amb criteri, experiència i proximitat.

elink.cat
info@elink.cat