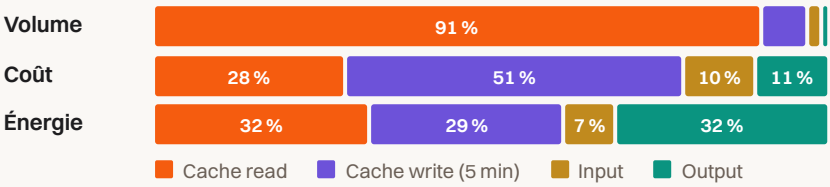
50 développeurs, un mois de Claude Code

Organisation type faite de chiffres publiés, aucun ne vient d'un client : 13 $ par développeur et par jour actif sur 15 jours, 54 % des sessions sur Opus (Anthropic), tokens répartis comme dans 2 676 sessions Claude Code (TraceLab).

Output	Input	Cache write	Cache read
75 061 726	361 870 333	1 447 481 330	20 244 387 526

Tokens traités	Coût aux prix publics	Émissions	Eau
22,1 Md	**9,8 k$**	**308 kg CO_2e**	**4,9 m³**
dont 91 % relus du cache	Claude Opus et Claude Sonnet	pour 0,82 MWh d'électricité	dont 98 % pour produire l'électricité

Les tokens générés : 0,3 % du volume, 11 % du coût, 32 % de l'énergie

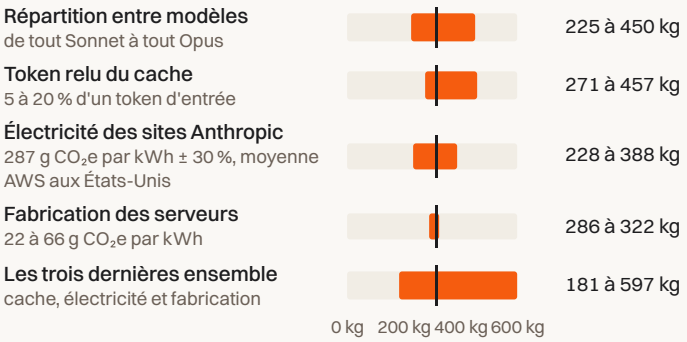


Coût. Le cache pèse 79 % de la facture : 2,8 k$ en cache read, 5,0 k$ en cache write.

Énergie. Un token d'output coûte 5 fois le prix d'un token d'input et consomme 21 fois plus d'énergie.

La fourchette autour des 308 kg

Chaque barre fait varier une seule hypothèse. Le trait noir marque les 308 kg retenus.



Le cache en chiffres

- Sans cache, le même trafic coûterait 61 k$ et consommerait 3,9 MWh. Le cache évite 84 % de la facture et 79 % de l'énergie.
- Envoyer un message dans une session de 100 000 tokens sur Opus coûte 0,02 $ et 1,9 Wh avant le premier token de réponse, autant que 380 tokens générés. À 500 000 tokens : 0,10 $ et 9,5 Wh.
- Le cache expire après 5 minutes sans activité, ou une heure si l'outil le demande. Passé ce délai, le message suivant réécrit tout le contexte.

Le calcul du coût, ligne à ligne

		Claude Opus 37 % des tokens		Claude Sonnet 63 % des tokens	
En $, prix par M tokens	Volume M tokens	Prix	Coût	Prix	Coût
Input	361,9	4	535	2	456
Output	75,1	20	555	10	473
Cache write 5 min	1 447,5	5	2 677	2,50	2 280
Cache read	20 244,4	0,20	1 498	0,10	1 276
Total			**5 265**		**4 485**

Coût = volume × part du modèle × prix. 5 265 + 4 485 = 9 750 $. Le calculateur affiche 8 470 €, à la moyenne mensuelle des taux BCE de septembre 2026 (1 € = 1,1513 $).

Par développeur

	Par mois	Par an
Tokens	443 M	5,3 Md
Coût	195 $	2,3 k$
Émissions	6,2 kg	74 kg
Énergie	16,4 kWh	196 kWh

Base 50 développeurs, le mois étudié répété douze fois.

Nos recommandations

D'abord, mesurer la répartition entre modèles. Selon la part réelle d'Opus, le mois coûte de 7,1 à 14,2 k$ et émet de 225 à 450 kg. La console Anthropic donne cette répartition.

Ensuite, les leviers, classés par coût évité.

Levier	Coût évité	CO_2e évité
10 points de trafic passés d'Opus à Sonnet	712 $	22 kg
10 % d'écritures de cache en moins	496 $	9 kg
10 % de contexte relu en moins	277 $	10 kg
10 % de tokens générés en moins	103 $	10 kg
Les quatre appliqués ensemble	**1 524 $**	**49 kg**

Ce total n'est pas la somme des lignes (1 588 $) : après le passage à Sonnet, chaque token évité coûte moins cher.

En CO_2e, l'ordre change. Réduire les tokens générés, dernier levier en dollars (103 $), devient le troisième en émissions (10 kg).

Le cache d'une heure. Il facture l'écriture 2 fois le prix de l'input au lieu de 1,25 : 2 974 $ de plus par mois à volume égal. Il n'est rentable que s'il évite plus de 37,5 % des écritures, celles qui suivent une pause de 5 minutes à une heure. Les transcripts de Claude Code permettent de les compter.

Pistes. Fichiers CLAUDE.md plus courts, serveurs MCP inutilisés désactivés, sessions redémarrées plutôt que prolongées. Exploration et tâches répétitives confiées à Sonnet, en mesurant la qualité équipe par équipe.