

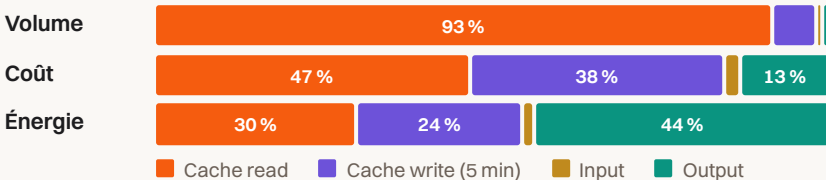
Scale-up française, un mois de Claude Code

70 développeurs, volumes relevés dans la console Anthropic.
Répartition estimée par l'équipe : **Opus 75 %, Sonnet 25 %**. Montants en dollars, la devise de facturation d'Anthropic.

Output	Input	Cache write	Cache read
131 400 000	88 600 000	1 540 000 000	23 870 000 000

Tokens traités 25,6 Md dont 93 % relus du cache	Coût aux prix publics 21,5 k\$ Opus 5 et Sonnet 5, sans remise	Émissions 489 kg CO₂e pour 1,32 MWh d'électricité	Eau 7,8 m³ dont 98 % pour produire l'électricité
--	---	---	--

Les tokens générés : 0,5 % du volume, 13 % du coût, 44 % de l'énergie

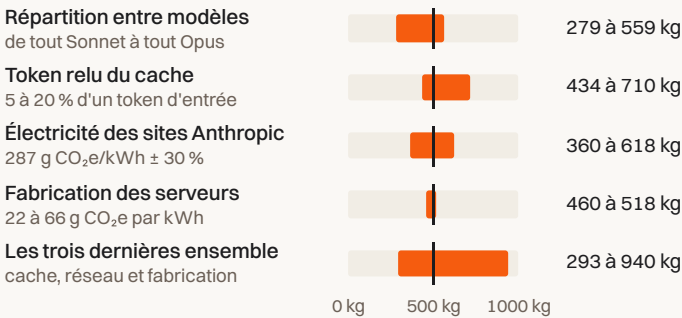


Coût. Le cache pèse 85 % de la facture : 10,1 k\$ en cache read, 8,2 k\$ en cache write. L'output en fait 13 %.

Énergie. L'output passe en tête, à 44 %.

La fourchette autour des 489 kg

Chaque barre fait varier une seule hypothèse, les autres restent fixes.
Le trait noir marque le résultat retenu.



Le cache en chiffres

- Sans cache, le même trafic coûterait 111 k\$ et consommerait 5,9 MWh. Le cache évite 81 % de la facture et 78 % de l'énergie.
- Envoyer un message dans une session de 100 000 tokens sur Opus coûte 0,05 \$ et 1,9 Wh avant le premier token de réponse, autant que 380 tokens générés. À 500 000 tokens : 0,25 \$ 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.
- Les 1,32 MWh sont comptés à 287 g CO₂e par kWh, la moyenne des régions américaines d'AWS. Anthropic ne publie pas la répartition entre ses sites, d'où la marge de 30 %.

Le calcul du coût, ligne à ligne

En \$, prix par M tokens	Volume M tokens	Opus 5 75 % du trafic		Sonnet 5 25 % du trafic	
		Prix	Coût	Prix	Coût
Input	88,6	5	332	2	44
Output	131,4	25	2 464	10	328
Cache write 5 min	1 540,0	6,25	7 219	2,50	962
Cache read	23 870,0	0,50	8 951	0,20	1 194
Total			18 966		2 529

Coût = volume × part du modèle × prix. 18 966 + 2 529 = 21 495 \$. [Le calculateur](#) affiche 18 763 €, au taux BCE du 22 juin 2026 (1 € = 1,1456 \$).

Par développeur

	Par mois	Par an
Tokens	366 M	4,4 Md
Coût	307 \$	3,7 k\$
Émissions	7,0 kg	84 kg
Énergie	18,8 kWh	226 kWh

Base 70 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 10,1 à 25,3 k\$ et émet de 279 à 559 kg. C'est l'hypothèse qui pèse le plus, et la console Anthropic la donne.

Ensuite, les leviers, classés par gain mensuel.

Levier	Coût évité	CO ₂ e évité
10 points de trafic passés d'Opus à Sonnet	1 517 \$	28 kg
10 % de contexte relu en moins	1 014 \$	15 kg
10 % d'écritures de cache en moins	818 \$	12 kg
10 % de tokens générés en moins	279 \$	22 kg
Les quatre appliqués ensemble	3 480 \$	73 kg

Ce total n'est pas la somme des lignes (3 629 \$) : une fois passé à Sonnet, chaque token évité coûte moins cher, donc rapporte moins.

Le carbone ne suit pas le même ordre. Générer moins de tokens rapporte peu en dollars (279 \$) mais se classe deuxième en CO₂e (22 kg) : un token généré consomme 21 fois plus d'énergie qu'un token d'input.

Le cache d'une heure n'est rentable qu'au-delà d'un seuil. Il facture l'écriture 2 fois le prix de l'input au lieu de 1,25, soit 4 909 \$ de plus par mois à volume égal. Il ne se rembourse que s'il évite plus de 37,5 % des écritures, celles qui suivent une pause de 5 minutes à une heure. Les transcrits de Claude Code permettent de compter ces reprises.

Pistes pour le contexte. Fichiers CLAUDE.md plus courts, serveurs MCP inutilisés désactivés, sessions redémarrées plutôt que prolongées. Pour le routage : exploration et tâches répétitives confiées à Sonnet, en mesurant l'effet sur la qualité équipe par équipe.