Où sont vraiment passés les 32 341 $ de la Partie 2

Compagnon de détail financier de la Partie 2 de la trilogie SHORA. Pour les lecteurs curieux de la décomposition ligne par ligne des dépenses AWS qui ont produit le taux de navigation de 9 % sur 10,5 millions de pages produit.

Ce post compagnon décompose, ligne par ligne, où sont passés les 32 341 $ de dépenses AWS de l'expérience de la Partie 2. **Source :** requêtes AWS Cost Explorer sur le compte de production SHORA, sur la fenêtre 2025-08-01 → 2026-05-14.

Les chiffres clés

Métrique	Valeur
Dépense AWS	**32 341,37 $**
Fenêtre calendaire	août 2025 → mai 2026 (10 mois ; 7 de combustion intense)
Mois de pic	février 2026 — 20 518,40 $
Jour de pic	2026-02-12 — 5 407,37 $ (dont 5 163,47 $ sur Claude Opus 4.6 seul)
Modèle d'inférence principal	Claude Opus 4.6 — 15 589,38 $ (48,2 % du total)
Région principale	eu-west-3 (Paris) — 22 538,41 $ (69,7 %)

> **Le LLM a consommé 65,8 % du budget.** L'ensemble des modèles Claude (Opus 4.6 + Opus 4.5 + Sonnet 4.5 + Haiku 4.5 + autres) totalisent 21 289,92 $ sur les 32 341,37 $. L'argument architectural de la Partie 2 — que le coût d'une couche de mesure par agent LLM est dominé par le modèle de langage lui-même — se reflète directement dans la facture.

La version intégrale — courbe de combustion mensuelle, ventilation par service AWS, répartition par région, courbe quotidienne de février 2026, économie des tokens Bedrock (entrée / sortie / cache), économie unitaire et ce que cette preuve soutient ou ne soutient pas — est en anglais.

Lire le compagnon financier complet (anglais) →