

MiniMax H3 : le vrai coût

API, local, 2K — ce qu’il faut budgéter avant production

Le tarif API est lisible.
La facture complète ne l’est pas.

BASE DE LECTURE : 10 S DE SORTIE

0,80 \$

API · 768P

1,30 \$

API · 2K

hors entrées, régénérations, GPU, stockage et droits

LA BONNE QUESTION

Quel est le coût par livrable —
une fois le droit d’usage validé ?

À 10 secondes de sortie

Prix publics MiniMax · 23.08.2026 · hors coûts indirects

API · 768P



0,80 \$

API · 2K



1,30 \$

Régénération 768P →
2K



0,50 \$

0 \$

1,30 \$

Ce qui change réellement la décision

- | | | |
|----|------------------|---|
| 01 | Context-IR | 0,90 \$/M tokens en entrée · 3,60 \$/M en sortie |
| 02 | Régénérations | chaque reprise ajoute temps, stockage et coût de sortie |
| 03 | Infrastructure | GPU, stockage, orchestration et maintenance si local |
| 04 | Données & droits | références client, contrat, territoire et licence |

À RETENIR

Le bon comparatif n’est pas « prix / seconde ». C’est : coût par livrable + infrastructure + droit d’usage.