

# L'IA locale : votre équipe, votre capacité

Quatre modèles Qwen/Swift · deux GPU R9700 · aucun quota de tokens fournisseur

**174,7 tokens/s pour une réponse · 855,5 tokens/s cumulés à huit demandes · ~128K tokens d'entrée.** Une capacité concrète pour les assistants, les dossiers longs et les agents en parallèle. Le serveur reste disponible selon votre organisation, sans crédits API à racheter ni remise à zéro d'un quota fournisseur.

**855,5**

tokens/s cumulés · 8 demandes soumises

**174,7**

tokens/s · une réponse (C1)

**0 USD**

facturation API locale

Quatre modèles, huit demandes soumises ensemble : débit mural (tokens/s)

A · Qwen3.8-27B



471,1 [471,1-471,2]

B · Swift1.5 Qwen3.8-27B



412,2 [412,2-412,2]

C · Qwen3.8 Flash Next original



693,3 [536,0-850,6]

D · Swift1.5 Flash Next



855,5 [855,2-855,9]

Médiane [min-max] de deux répétitions ; cercle R1, carré R2. 128K réels · low · sortie ≤4096 · fenêtre 262144.

A | Qwen3.8-27B / Paition | 2 R9700, 2 x TP1 | W3A4 / W3Rot | DFlash7 | KV KV4 | image 245c71d5

B | Swift1.5 Qwen3.8-27B / Paition | 2 R9700, TP2 | Quark RTN-MXFP4 | MTP3 (tête BF16) | KV FP8 | image 245c71d5

C | Qwen3.8 Flash Next original / Radiance natif | 2 R9700, TP2 | experts IQ4R + tronc INT8 | MTP3 | KV FP8 | image ef8cecc8

D | Swift1.5 Flash Next / Radiance natif | 2 R9700, TP2 | experts IQ4R + tronc INT8 | MTP3 | KV FP8 | image ef8cecc8

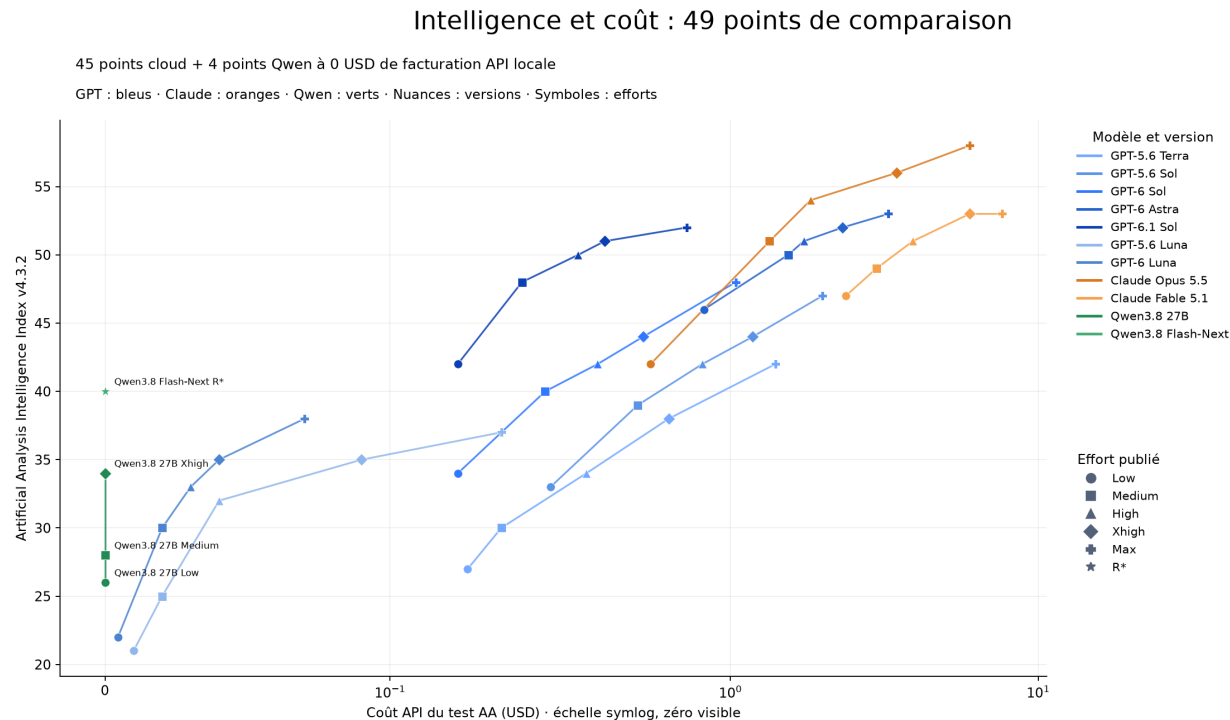
Deux R9700 · entrée 127632-127674 tokens · cache naturel. 0 USD = tokens API uniquement, matériel/électricité exclus.

**Combien de personnes ?** Avec Swift Flash Next à ~128K / low : une tâche, **174,7 tokens/s par réponse** ; deux tâches, **269,9 tokens/s cumulés** ; quatre, **463,7** ; huit, **855,5** et **110,4 tokens/s par réponse**. Huit personnes lançant chacune une demande correspondent au point C8 soumis ; une équipe plus grande peut partager la machine avec des usages intermittents. La file et le temps de réponse dépendent de la charge.

**384 groupes réussis, 1 440 requêtes, deux répétitions** : Qwen3.8-27B, Swift1.5 Qwen3.8-27B, Qwen3.8 Flash Next et Swift1.5 Flash Next. Fenêtre configurée : 262 144 ; entrée réellement mesurée : jusqu'à ~128K ; sortie ≤4 096. Le débit cumulé inclut préremplissage et attente ; la génération par réponse les exclut.

# Intelligence : situez le local face au cloud

**Le choix du local peut associer qualité des modèles ouverts et absence de facturation des tokens API.** Comparez les niveaux de raisonnement et le coût : les 49 points du graphique d'origine sont conservés, dont 45 références cloud et quatre points Qwen.



Source : Artificial Analysis, relevé du 12 octobre 2026. Scores de référence des modèles ; qualité des recettes quantifiées non remesurée. Le coût du test AA n'est pas un tarif par million de tokens. Electricité et matériel exclus du coût API local.

**Des repères chiffrés :** dans l'archive Artificial Analysis du 12 octobre, Qwen27 obtient **26 / 28 / 34** en low / medium / xhigh ; Qwen Flash Next atteint **40** dans le réglage publié. Sur la sélection LiveBench, les scores globaux sont **76,19 pour Qwen Flash Next**, **75,27 pour Qwen27** et **75,67 pour Claude Sonnet 5.5 max**. Les modèles cloud les mieux classés restent plus hauts : Opus 5.5 max **83,22**, GPT-6 Astra max **82,16**. Ces indices distincts ne s'additionnent pas et le classement varie selon la tâche.

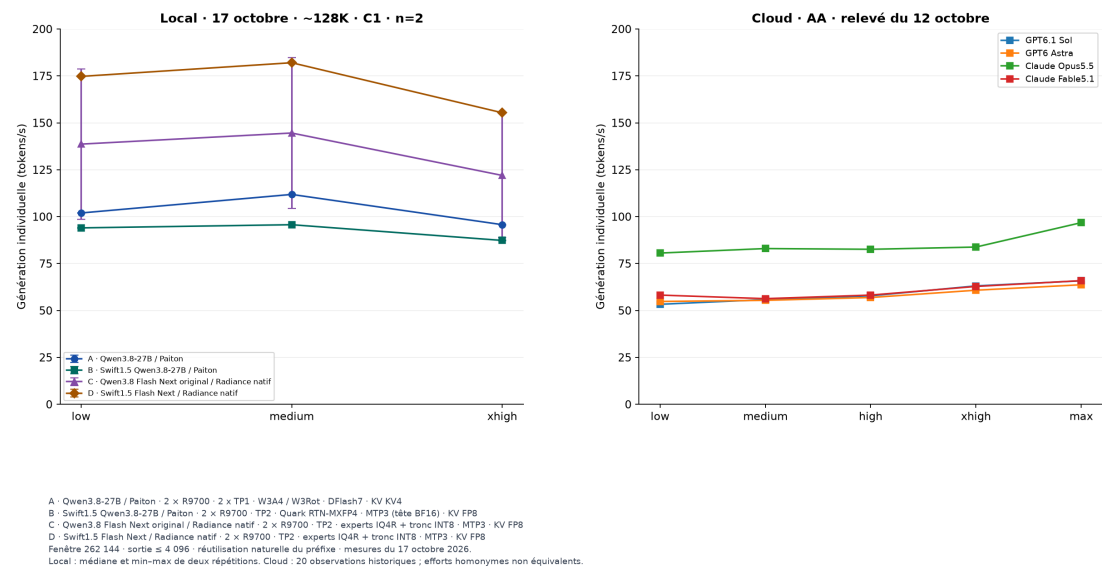
**Lecture du graphique :** plus haut = meilleur indice ; plus à droite = coût API par tâche Artificial Analysis plus élevé. Le coût local API est **0 USD**. Les scores décrivent les modèles de référence publiés ; la qualité des variantes locales quantifiées et Swift n'a pas été mesurée ici. Sources : [Artificial Analysis](#) et [LiveBench](#). Détails, tests de code, GPQA et tableaux complets dans le [document de base](#).

# Vitesse et coût : une alternative locale concrète

Une réponse à 174,7 tokens/s au point ~128K / low / C1, jusqu'à huit demandes soumises ensemble : dimensionnez le serveur selon vos usages. Le graphique conserve les 20 références cloud et les vitesses locales, avec leurs protocoles distincts.

Vitesses locales et cloud : comparaison documentée

Deux sources et protocoles distincts : pas un classement à conditions identiques.



| Exemple : 128 000 tokens d'entrée + 4 096 de sortie |  | Facturation API, USD |
|-----------------------------------------------------|--|----------------------|
| Quatre configurations locales                       |  | 0                    |
| GPT-6.1 Sol                                         |  | 0,29696              |
| Claude Opus 5.5                                     |  | 0,59392              |
| GPT-6 Astra                                         |  | 1,4848               |

**0 USD signifie zéro achat de tokens API locaux**, hors matériel et électricité. Exemples cloud calculés aux tarifs publics archivés du 18 octobre 2026, sans cache, outils, taxes ni remises. Les mesures cloud externes et la campagne locale utilisent des tâches et contextes différents : le graphique situe les ordres de grandeur. Sources tarifaires : [OpenAI](#) et [Anthropic](#). Toutes les conditions, répétitions, comparaisons de coût et sources restent dans le [document complet](#).