L'oracle est en pause quota

L'oracle est en pause quota

> *Dernière mise à jour : 2026-05-10*

Symptômes

- Une ingestion en cours s'arrête net et passe en `ingestStatus='scheduled'`
- L'UI affiche un panneau "L'oracle se repose, reprise à HH:MM"
- Les questions `/play` peuvent renvoyer une erreur ou rester bloquées en "thinking"

Pourquoi

Le compte Claude utilisé par la VM `oracle` a atteint son quota d'usage (Pro / Team plan, fenêtre 5h glissante). Le CLI `claude` renvoie alors un message du genre :

> Usage limit reached, your limit will reset at 21:00 PM

Le service `claude-quota.ts` détecte ce message (multiples patterns supportés : `reset at`, `try again at`, `available again at`, formats 12h/24h, timestamp Unix) et lève un `ClaudeQuotaError` au lieu de renvoyer la réponse vide.

Ce que fait l'app automatiquement

Pendant une ingestion

1. Les workers SSH parallèles (contextual-llm, conflict-detect) arrêtent de spawn de nouvelles tâches dès qu'une `ClaudeQuotaError` est détectée
2. Les caches JSON sont **flushés sur disque AVANT** de propager l'erreur (critique pour la reprise sans perte)
3. Le `coordinator.ts` rattrape l'erreur, bascule la ligne SQLite en `ingestStatus='scheduled'` avec `ingestScheduledAt = resetAt + 2 min`
4. Émet un event SSE `quota_pause` avec `{ resetAt, retryAt }` → le wizard UI affiche la copie dédiée
5. À l'heure de reset, le scheduler relance `runIngestion` qui re-rentre dans les stages, lit les caches JSON, et ne re-traite que les chunks manquants

Pendant une question /play

L'oracle peut hallu silencieusement quand le quota est dépassé (le CLI renvoie parfois la quota notice comme une réponse assistant normale). Le détecteur scanne aussi le contenu streamé pour ce cas. Si détecté, l'erreur est exposée à l'UI qui affiche un message clair.

Ce que tu peux faire

- **Patienter** : laisser tourner, ça reprendra tout seul à `resetAt + 2min`
- **Vérifier l'heure de reset** : event SSE `quota_pause` ou logs `/app/data/logs/server.log` (grep `quota`)
- **Si urgent** : passer sur un autre compte Claude (changer les credentials dans `/home/oracle/.claude/.credentials.json` côté VM oracle, restart container backend)

Si la détection ne marche pas

Si Claude change le wording du message de quota et que `claude-quota.ts` ne le reconnaît plus, l'erreur sera générique ("answer empty") au lieu de pause planifiée.

Étendre `QUOTA_MARKERS` dans `src/services/claude-quota.ts` :

```
const QUOTA_MARKERS = [
  'usage limit reached',
  'you've hit',
  'rate limit reached',
  // ... ajouter le nouveau marker
];
```

Et tester avec un mock de stream qui contient le nouveau message.

Revision #1
Created 2026-05-10 15:19:44 UTC by thymon
Updated 2026-05-10 15:19:44 UTC by thymon