ADR-004 : Vision inline (lue au moment Q) plutôt que pré-ingestion

ADR-004 : Vision inline (lue au moment Q) plutôt que pré-ingestion

> *Date : 2026-04-11 — Statut : accepté*

Contexte

Les règles de jeux contiennent souvent du contenu visuel important : icônes, schémas, plateau, tuiles, couleurs. Le retrieval texte seul rate ces aspects.

Premier prototype 2026-Q1 : vision **à l'ingestion** — chaque page était envoyée à un modèle vision qui produisait une description textuelle, ingérée comme un chunk supplémentaire.

Problèmes constatés :

- **Coût élevé** : 1 appel vision par page × N pages × tous les jeux ingérés = facture importante
- **Sur-description** : le modèle décrivait des détails inutiles, pollution du retrieval
- **Latence d'ingestion** : ajoutait 5-10 min par jeu

- **Description figée** : si la question demande "que représente la zone rouge en haut à gauche", la description ingérée n'a peut-être pas mentionné cette zone (subjective)

Décision

Vision inline au moment de la question. Claude lit le PNG de la page la plus pertinente directement via son outil `Read` côté VM SSH, **uniquement** quand la question a une dimension visuelle.

Détails dans `pipeline-rag/vision-inline.md`.

Conséquences

Bonnes

- **Coût zéro à l'ingestion** : juste `pdftoppm` qui rend les PNG localement
- **Précision** : Claude regarde l'image avec la question en tête → réponse ciblée, pas de description générique
- **Latence d'ingestion réduite** de 5-10 min
- **Fonctionne grâce à** `--append-system-prompt` qui préserve le contrat outils Claude Code
- **Pas de dépendance à un modèle vision tiers** : Claude Code suffit

Mauvaises

- **Latence par question +20-30s** quand l'image est lue (Read tool sur PNG 300 DPI)
- **1 image max** : si la question concerne 2 pages, on en perd une. Compromis assumé (passer à 2-3 doublait/triplait la latence pour gain marginal)
- **Granularité** `(livret, page)` **obligatoire** : oubli a déjà coûté un bug — chunk page 4 base + chunk page 4 extension donnent la mauvaise image
- **Dépend du chemin SSH côté oracle** : `permissions.additionalDirectories` doit pointer sur le volume PNG côté VM
- **Si le PNG manque** (rendu raté à l'ingestion), Claude ne peut pas lire — silently dégradé

Alternative envisagée

- **Vision pré-ingérée + cache** : décrit chaque page une fois, stocke la description, ré-utilise. Coût initial élevé, problème de description figée non résolu.
- **Vision par chunk** : décrit la zone autour du chunk plutôt que la page entière. Trop fin, pas implémentable simplement avec PDF natif.
- **OCR seulement** : voir ADR sur OCR (Phase 1 livrée 2026-05-06). Complémentaire mais pas substitut — l'OCR récupère du texte, pas l'analyse visuelle.

Variante future (Phase 2 OCR)

Roadmap : Phase 2 OCR (cf. mémoire `project_ocr_phase2.md`) prévoit Claude vision en fallback de Tesseract (page-by-page si confidence basse). Différent du flow actuel : fallback OCR vs lecture à la question. Les deux peuvent coexister.

Revision #1
Created 2026-05-10 15:19:52 UTC by thymon
Updated 2026-05-10 15:19:52 UTC by thymon