

Anthony Windey

Développeur IA agentique & IBM i

Cette année, j'ai conçu **un harnais pour que Claude Code développe en sécurité, un dispositif de développement assisté par IA pour IBM i et un assistant multi-agents auto-hébergé** que j'exploite. Chaque chiffre est mesuré.

43 · 23

skills · hooks conçus pour Claude Code

7 836

commits en 2026 sur 5 dépôts, tous écrits par Claude Code

1 880

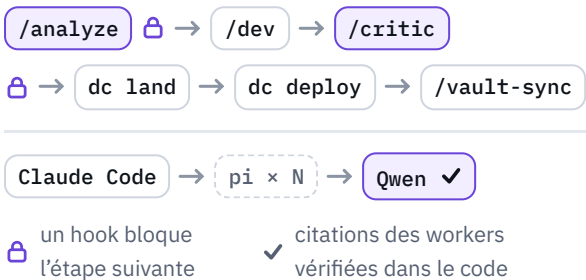
notes dans le vault du projet (sa base de connaissances), dont 482 décisions

2 × RTX 5090

inférence locale, génération d'images et de vidéos, workers locaux de Claude Code

01 Harnais pour Claude Code

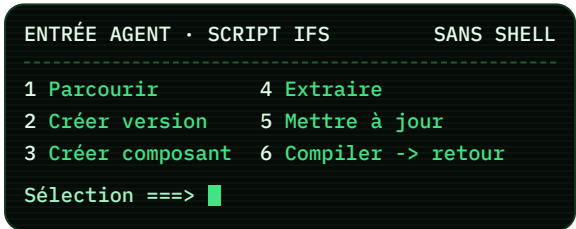
Plugin de skills et de hooks · CLI de session · serveurs MCP · workers locaux



- **La méthode en skills.** /analyze planifie, /dev implémente, /critic relit le diff, dc (un outil maison) fusionne et déploie, /vault-sync enregistre les leçons.
- **MRS Memory.** Mémoire de type wiki : embeddings + BM25, RRF, cross-encoder ; rappel testé sur jeux de référence.
- **Imposé par des hooks.** Pas d'édition sans plan, pas de fusion sans relecture ; une worktree par session.
- **Workers locaux.** Les grosses lectures vont à des agents locaux (pi) sur Qwen, une question par tâche, sans coût d'API.

02 Développement assisté par IA sur IBM i

Agent dans VS Code sur le poste, qui joint l'IBM i par SSH · MCP

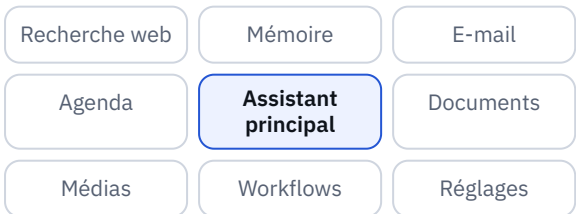


ILLUSTRATION

Tout le cycle, derrière **une seule clé SSH**

- **La carte d'abord.** Un skill maison documente chaque concept métier qu'une tâche touche et guide la navigation dans le code.
- **Du contexte.** Les sources IBM i sont exportées dans l'IFS ; la documentation du projet est synchronisée sur le poste.
- **Un script, toute la boucle.** Il enchaîne les étapes jusqu'à la compilation, versions et extractions ARCAD comprises, et renvoie les erreurs ; l'agent corrige jusqu'à ce que ça compile. Écrans verts 5250 pilotés par MCP, avec un miroir dans le navigateur.

03 MRS Studio Assistant multi-agents auto-hébergé · conçu et dirigé seul · depuis mars



Qwen sur 2 GPU locaux · Claude au choix

- **Un spécialiste par domaine.** Chaque demande va à l'un des huit spécialistes, avec ses outils et son contexte.
- **Un « brain » qui consolide.** L'assistant extrait des faits des conversations et des e-mails ; une étape de consolidation retient le durable et repère les doublons.
- **Sur mon matériel.** Modèles Qwen quantifiés (vLLM, NInfer, Strata) derrière une passerelle compatible OpenAI.