

Anthony Windey

Agentic AI & IBM i developer

This year I designed a harness that lets Claude Code develop safely, an AI-assisted development setup for IBM i, and a self-hosted multi-agent assistant that I run. Every number below is measured.

43 · 23

skills · hooks I designed for Claude Code

7,836

commits in 2026 over 5 repositories, all written by Claude Code

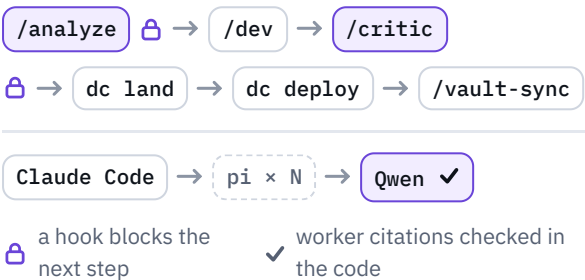
1,880

notes in the project vault (its knowledge base), 482 of them decisions

2 × RTX 5090

self-hosted inference, image and video generation, local workers for Claude Code

01 Claude Code harness Plugin of skills and hooks · session CLI · MCP servers · local workers



- **Method as skills.** /analyze plans, /dev implements, /critic checks the diff, dc (an in-house tool) merges and deploys, /vault-sync records the lessons.
- **MRS Memory.** A wiki-style memory engine with hybrid search: embeddings + BM25, RRF fusion, cross-encoder rerank; recall tested on golden sets.
- **Enforced by hooks.** No edit before the plan, no merge before the review; one git worktree per session.
- **Local workers.** Big reading jobs are split into one-question tasks for local agents (pi) on Qwen, at no API cost.

02 AI-assisted development on IBM i

Agent in VS Code on the workstation, reaching the IBM i over SSH · MCP

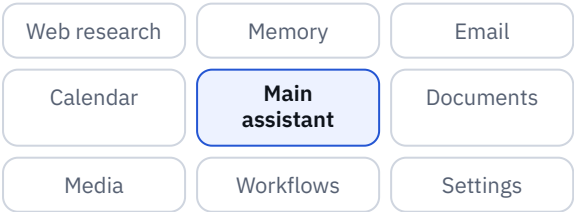
```
AGENT ENTRY · IFS SCRIPT      NO SHELL
-----
1 Browse                     4 Check out
2 Create version             5 Update source
3 Create component           6 Compile -> feedback
Selection ==> █
```

ILLUSTRATION

The whole change cycle, behind one SSH key

- **Map first.** An in-house skill documents each business concept a task touches and guides navigation to the code.
- **Context.** IBM i sources exported to the IFS; project documentation synced to the workstation.
- **One script, full loop.** It runs every step through compilation, ARCAD versions and check-outs included, and returns the errors; the agent fixes until it compiles. 5250 green screens driven through MCP, mirrored live in a browser.

03 MRS Studio Self-hosted multi-agent assistant · designed and directed alone · since March



Qwen on 2 local GPUs · Claude if chosen

- **One specialist per domain.** Each request goes to one of eight specialists, each with its own tools and context.
- **A brain that consolidates.** The assistant extracts facts from conversations and mail; a consolidation step keeps the lasting ones and flags duplicates.
- **On my own hardware.** Quantized Qwen models (vLLM, NInfer, Strata) behind an OpenAI-compatible gateway.