

Raymond Huang

Senior IT leader with 20+ years in cloud operations, information security and governance, now a hands-on AI builder. I run my own travel media, membership and travel advisory business as a one-person operation, on an AI automation stack I designed and built. Open to contract and project-based AI automation work.

Email me · jwhuang@hotmail.com · raymondhuang.ca · linkedin.com/in/rjwhuang

60+

Claude Code skills in use

100+

scheduled workflows

4,300+

articles behind a RAG assistant

20+ yrs

IT, cloud & security leadership

Selected work

All of these are running in production for my own business. The numbers come from my own logs and records.

01 One-person AI operations platform

Problem. Four business lines and one person: a content site, a paid membership community, a travel advisory practice and a lead-generation site.

- 60+ Claude Code skills, each encoding one repeatable job: drafting articles, researching prices, handling inquiries, renewing members, reporting.
- 100+ scheduled workflows run those jobs unattended. Slack is the operations console for results, diffs, failures and items waiting for approval.
- Each job goes to the cheapest model that can do it: Claude Opus for heavy reasoning, Sonnet for standard execution, OpenAI Codex for read-only research, and plain scripts where no model is needed.

Day-to-day research, publishing, monitoring and reporting run without me. I spend my time on judgment calls and on clients.

Claude Code

Python

PowerShell

Task Scheduler

MCP

Slack

02 Members-only RAG assistant, and the security incident that improved it

An on-site AI assistant answers member questions from 4,300+ articles (14,000+ vector chunks) and five community chat archives. Answers are scoped to each member's tier.

The incident. Paid articles were protected only where the web page is displayed. The assistant searched the vector index directly, and the index did not know which articles were paid. Non-paying accounts could therefore get paid content by asking the assistant.

The fix.

- Every chunk now carries a permission tag, and the assistant filters by permission before any text reaches the model.
- Tags are re-applied after every content refresh. If an article fails to fetch, the job does not default it to public.
- Role-based regression tests replay the same questions as different member tiers.

Audit after the fix: 0 unprotected chunks across 84 gated articles. Average response time tuned from 14.3s to 11.1s on a fixed 8-question benchmark, then streaming added.

FastAPI

Qdrant

OpenAI embeddings

Claude

WordPress plugin

HMAC sessions

Cloudflare Tunnel

03 Live pricing intelligence

Problem. Travel prices sit behind authenticated trade portals and change daily. Looking them up by hand did not scale, and published price pages went stale.

- Browser-automation pipelines run inside an already-authenticated session. Where possible they call the portal's own endpoints from inside the page instead of clicking through the UI.
- They feed price lists that refresh automatically: 600+ cruise sailings, 20 flight routes, all-inclusive packages and resort deals. The oldest entries refresh first, so the pipelines stay within rate limits.
- A bookability check follows each search result through to the payment page, so prices that show in search but cannot be booked get caught.

One portal rewrite cut a 20-hotel search from over a minute to about 26 seconds. Every booking flow stops at the payment page for a human to approve.

Chrome DevTools Protocol

Playwright

Python

Static site generation

04 Change-detection sentinels

- **Offer monitor:** checks 20 credit-card issuer pages every day, compares them with the published articles, drafts updates and posts the diff to Slack. If a page cannot be read reliably, it goes to a human instead of being guessed.
- **News monitor:** scans 7 industry RSS feeds every 3 hours for promotions that affect existing articles.
- **Award-seat report:** a daily report on a third-party API capped at 1,000 calls a day. Queries are batched, and new features reuse data already fetched, so they add zero extra calls.

Python

RSS

REST APIs

Slack

05 Membership and community automation

- **Renewals:** a spreadsheet is the source of truth. Reminders go out once per member, sorted by tier, through an email API with SMTP fallback. Status is written back to the sheet and summarized in Slack.
- **Community mining:** every day an LLM screens five member chat groups (a 316K-message archive) and surfaces content candidates. A human decides what gets published.
- **Knowledge archive:** 722 member-reported data points consolidated into a searchable archive, updated with new entries every two weeks.

Google Sheets API

Resend

LLM classification

06 Web products and prototypes

- **cpptool.ca:** a CPP retirement calculator in 6 languages. Static site with an SVG break-even chart and zero backend cost.

- **Product-review site:** static site with a Python build pipeline, GA4 analytics and affiliate links.
- **Chrome extension (working prototype):** Manifest V3. Converts WebP/AVIF images to JPEG/PNG from the right-click menu, entirely in the browser.
- **Generative video prototype:** text to image (FLUX) to video (Kling).

How I build

Twenty years of security and audit work shapes how I let AI act on a business.

Human sign-off where it matters

Payments, client emails and anything published go to a person for approval. Agents prepare the work; they do not make those decisions.

Verify before publishing

Every fact and number gets checked against its source before it goes public. A wrong price published once costs more than a slow pipeline.

Least privilege for agents

Secondary models run read-only. API tokens are scoped to one site. Agents archive files instead of deleting them.

Cost is a design input

Each job goes to the cheapest tool that can do it. Rate-limited APIs get batched queries and shared data.

What I can do for your team

Contract or project-based. Remote, or on site in Metro Vancouver.

Automate repetitive work

Map where a team's time goes, then automate the highest-value pieces with LLM agents and scheduled workflows.

RAG assistants with real access control

Assistants over your internal documents, with permissions enforced at retrieval and tested role by role.

Data collection without an API

Browser and API automation for portals and sites that offer no export, with monitoring and alerts.

AI governance

Usage policy, approval checkpoints and audit trails. I wrote a corporate AI policy in my last IT leadership role.

Background

B.Sc. Computer Science · ISO 27001, SOX and GDPR programs · AWS multi-region operations

2026 – present	Founder & Applied AI Builder Travel media, membership and travel advisory business, run on my own AI automation stack
2022 – 2026	Senior IT Manager Health research funder. Led the corporate AI policy, Azure and Microsoft 365, vendor risk and IT budget
2021 – 2022	Governance, Risk & Compliance Manager Global water-quality group. SOX program, vendor security assessments, privacy
2017 – 2020	Technical Operations Manager Cloud operations and information security for a global SaaS: AWS in the Americas, EMEA and APAC; 99.95% uptime; 100+ external audits
2001 – 2017	IT Manager · Operations Team Lead · Systems Administrator Production environments, disaster recovery, team leadership

Let's talk

Have a process that eats your team's time, or an AI pilot that needs to go to production safely? Send me a short note about it.

jwhuang@hotmail.com · linkedin.com/in/rjwhuang

© 2026 Raymond Huang · Vancouver, BC