

Erik Hill

AI Evaluation & Testing Engineer

Summerville, SC (Charleston area) · erik.hill0987@gmail.com · 808-765-0240
linkedin.com/in/erik-hill-98895575 · egnaro9.github.io · Open to remote (US)

SUMMARY

Self-taught AI evaluation and testing engineer. In under a year, career-changing from professional kitchens, I built a **verifiable-evaluation suite**: seeded-defect task families with provable answer keys, a **certification lab for coding agents**, a **published mutation-tester for eval graders**, and an evidence protocol under which **eleven published evaluation claims are independently replayed in CI** — deterministic, no LLM judges, failures published by design. Also **shipped a game to Google Play**; **twenty-plus public repos**, **two PyPI packages**, **one npm package**, **a merged fix to a 12-year-old compiler**, and **an open fix to a widely-used eval framework**.

SKILLS

Agentic / AI: Multi-agent orchestration · LangGraph · Model Context Protocol (MCP) · LLM application development · Tool-calling & guardrails · RAG · Evals & regression gating (nDCG) · Prompt & context engineering · Safety (OWASP LLM Top 10, human-in-the-loop) · Differential-oracle, property-based & mutation testing · No-LLM-judge grading
Backend & data: Python · FastAPI · Pydantic · SQLAlchemy 2.0 · PostgreSQL · pgvector · Alembic · REST / OpenAPI

Frontend / mobile: TypeScript · JavaScript · React · Next.js 15 · Java · Capacitor · Android (adb) · WebAssembly (Pyodide, TeaVM)

Ship & verify: pytest · Vitest · JUnit · jqwik · Playwright · PyPI / npm packaging · GitHub Actions (matrix CI) · Docker · Render · Neon

SELECTED PROJECTS — all public at github.com/egnar09

- **agent-certlab** — **Certification lab for coding agents**: three **seeded-defect task families** with provable answer keys (incl. **coordinated two-file defects** no single-file patch can pass), a **precondition prover** (clean passes, every seed fails — it refused one of my own defects), and **calibration agents** the harness must separate before certifying. Two real agents certified; every verdict **regradeable from artifacts**.
- **reference-fleet** — **An answer key for eval suites**: six models each broken **one documented way at an exact seeded rate**, so detection rates are measurable against ground truth — a naive suite detects **1 of 6** (paired protocol). Plus a **LoRA-trained native defect**: mixture 0.507 → realized **0.200** — training's rate control is lossy, measured.
- **vac-protocol** — **Verifiable Agent Claims**: an evidence-bundle format + **offline verifier**; a registry where **eleven real claims are hash-checked and re-earned in CI**, with a **public record of it rejecting my own evidence**. Plus **vac-gate**: replayable evidence or no green check.
- **evalmut** — **Mutation testing for eval graders (PyPI, with its gradecore engine)**: inject known defects, report which checks stayed green. Found **3 holes in its own engine's suite**, **6 in ports of a widely-used framework's assertions**. **No LLM judge**; every operator tied to a documented real failure. *90 tests*.
- **rag-eval-lab** — deterministic RAG **eval harness** that catches a planted hallucination in CI; from-scratch **BM25 matching the published SciFact baseline (0.664 vs 0.665)**. *38 tests*.
- **Also: crashkit** — live (crashkit.onrender.com) deterministic adversarial stress-test · **model-drift** — public board re-running a **frozen, hash-fingerprinted suite** across **16 models / 5 labs**, with a derived minimum-detectable-regression floor · **eval-history** — deployed FastAPI/Postgres eval store · **mcp-tools** — MCP server from the spec, no SDK · **llm-wire-stub** (npm) · a merged **TeaVM compiler fix** · an **open PR to promptfoo**.

EXPERIENCE

Independent AI Evaluation & Testing Engineer — Solo Practice · Remote · Mar 2026 – Present

- Designed a five-stage agent loop (Strategy → Execution → Critic → Evaluation → Ops) that plans, builds, reviews, and validates changes to a real React / Capacitor product, human-in-the-loop on every consequential step.
- Built per-role model routing and hook-based automation that hands work between stages with no manual relaunch; enforced rules with **property-based invariants** (jqwik) over a native engine and **adb-driven on-device validation**. Case study: github.com/egnar09/agentic-dev-harness

Indie Game Developer — SeraphLight Studios · Remote · Nov 2025 – Present

- **Shipped Tap Dodge Rush to Google Play** (Kotlin / Jetpack Compose, AdMob) — closed testing with 12+ testers, then store review, then launch. Building a second, larger game (React / Capacitor) — the product the harness above operates on.

Kitchen Manager / Shift Lead — Big Island Brewhaus & other high-volume kitchens · Hawaii · 2018 – 2025

- Led shifts and trained staff across island restaurants; owned prep, inventory, and food-safety under sustained pressure — the reliability and discipline I bring to software now.

EDUCATION

freeCodeCamp — self-directed software curriculum.