

Hyowon Bernabe | Full Stack Developer + AI Engineer (Master Resume)

Email: contact@hyowonbernabe.me

LinkedIn: linkedin.com/in/hyowon-bernabe

Contact: +639948640169

GitHub: github.com/hyowonbernabe

Portfolio: hyowonbernabe.me

PROFESSIONAL SUMMARY

AI engineer who takes LLM products from first idea to live release: agents, RAG, and fine-tuned models inside full-stack apps people actually use. Works across the whole stack, from model and prompt design to Next.js front ends and low-level Rust and C#.

EXPERIENCE

Chief Technology Officer (CTO) | FLH Media (flhmedia.com) (2026 - Present)

- Built **Jarvis, an AI cold-email agent** that researches each lead's website and socials, writes a personal email about something real in their business, sends two timed follow-ups, sorts replies, and refuses instructions hidden in the pages it reads.
- Gave it guardrails for a cold-email account: it sends only in the lead's working hours, handles bounces and unsubscribes, and raises its own daily send limit only after a clean day.
- Built the lead pipeline behind it: Python crawlers over Google Maps and business directories across **808 Florida cities**, turning **381,000+ listings into 275,000+ deduplicated leads**, 77,000+ with emails.
- Shipped **10+ client websites**, hand-coded in **Next.js, Tailwind, and GSAP** and deployed on Vercel.

Chief Technology Officer (CTO) | ZuraLog (zuralog.com) (2026 - Present)

- Led the build of an **AI health coach live on the App Store and Google Play**, writing ~80% of 4,850 commits across Flutter (iOS and Android), FastAPI on Railway, and Next.js.
- Designed the coach as a **bounded tool-use loop over 17 MCP tool servers** (workouts, nutrition, sleep, memory, charts), with safety checks before every model call and a cross-vendor fallback (Gemini to Claude Haiku) behind a circuit breaker.
- Wrote native **Swift HealthKit** and **Kotlin Health Connect** platform channels instead of a third-party health package, and gated **Pro subscriptions** on the server through RevenueCat.
- Ran the backend as 5 Railway services (FastAPI, Celery workers, Redis, Postgres on Supabase) covered by **360 backend test files**.
- Built an image pipeline that made one AI illustration per exercise for an **839-exercise library**, using real photos as the base so the form stays correct.

Full Stack Developer + AI Engineer | Globit Transient (2023 - Present)

- Reduced manual booking errors by **~60%** by building a **full-stack booking platform with Next.js, PostgreSQL, and Prisma**, with conflict-proof availability checks, PayMongo payments, and auto-expiry of unpaid bookings.
- Handled **~70%** of customer inquiries and cut response times by **~50%** with an **AI concierge** on the **Vercel AI SDK** that answers from the live knowledge base and real-time unit availability.
- Built an admin dashboard tracking **occupancy and RevPAN**, plus calendar, agent commissions, reviews, and an audit log.
- Added push notifications and 14 unit and end-to-end test files, including a double-booking test.

PROJECTS

HuntPeople (AI People-Search Agent, huntpeople.ai) | Python, FastAPI, Next.js, Docker, Postgres, Stripe (2026)

- Built and launched a **live, paid OSINT product** where an AI agent takes one clue (an email, phone, handle, or domain) and chains lookup tools until it identifies the person, for one search or an uploaded list of hundreds.
- Ran every search in its **own throwaway Docker container** with a real shell, so pages the subject controls can never touch the server, and **capped each run by spend, steps, and time** no matter what the agent decides.
- Kept record merging in code, not the model: profiles only join on a **matching phone, email, or domain**, so one wrong guess can't spread to every later customer. Shipped teams, a public API, and Stripe billing on top.
- Replaced the first hand-written search loop with **opencode running in a per-search container**, behind Caddy, a worker queue, and Resend email, all on one server (410 commits).

Stonecutter AI Platform (AI Analyst for Amazon Brands, [GitHub](https://github.com)) | Next.js, Vercel AI SDK, Orama, Transformers.js, PromptFoo (2026)

- Built an **AI assistant that picks its own tools** (SQL queries, document search, listing compliance checks) and chains them for mixed questions, with a **27-test PromptFoo eval suite** written before the UI.
- Made **text-to-SQL safe**: SELECT only, every table and column checked against an allowlist, a read-only connection, and one retry that feeds the error back to the model.
- Built **hybrid keyword and vector search** over local embeddings, with heading-aware chunks, a context summary on each chunk, and parent chunks returned when a smaller one matches.

ShadowPrompt (Portable Discrete Academic Interface) | Rust, Win32 API, OpenRouter (Claude), Chrome DevTools Protocol (2026)

- Wrote a **tool-calling agent loop from scratch in Rust**, with parallel tool calls and live web search the model triggers itself, shipped as a **single zero-install file that runs from a USB drive** (84 downloads).
- Automated **multi-page Google Forms** through the user's own Chrome: it reads every question type and fills only unanswered ones.
- Cut repeat-question cost **~10x** by sending study notes as a **cached prompt block**, backed by 73 Rust tests and a PromptFoo eval suite for the system prompt.

Kuroko (Stealth AI Interview Assistant) | C#, WPF, Whisper.net, SQLite, Semantic Kernel (2025)

- Built a **real-time speech-to-text pipeline** with **local Whisper** and silence detection that feeds live answer suggestions during calls.
- Built **local RAG over personal documents**: embeddings stored in SQLite with cosine similarity search written in C#.
- Made the overlay **invisible to screen sharing** using the Win32 display-affinity API to exclude it from capture.

AI TOS Risk Summarizer (*AI-Powered Legal Risk Detector*) | Python, Llama 3.1 8B, Unsloth, LoRA, Ollama, Gradio (2025)

- Fine-tuned **Llama 3.1 8B with LoRA** on a free Colab T4 to flag fair, risky, and unfair Terms of Service clauses, publishing three model versions on **Hugging Face** with loss graphs.
- Built the training set by **knowledge distillation**: a larger teacher model (Gemini) labeled and explained **698 balanced examples**.
- Shipped it as a **160 MB adapter instead of a 16 GB model**, loaded at runtime by Ollama for private, on-device analysis.

Messenger Z (*Privacy-First Messenger Modification*) | Kotlin, Xposed API, LSPatch, DexKit, GitHub Actions (2025)

- Built a **root-free Messenger privacy mod** (~211 downloads) by injecting the **Xposed framework** into the APK with **LSPatch**, with a GitHub Actions pipeline that rebuilds and releases a patched build every two weeks.
- Built **"Ghost Mode" for end-to-end encrypted chats** by reverse-engineering Messenger's bridge to its native code and dropping read-receipt commands, plus story download and an unsent-message logger.
- Added per-chat ghost mode, link-tracking removal, an in-app debug console, and a build that installs next to the Facebook app.

Phantom (*Agentic AI for Gaming PCs, entry, The Next Gen 2026*) | Tauri v2, Rust, React, TypeScript, Supabase, OpenRouter (2026)

- Built a desktop AI agent that finds out why a gaming PC runs badly, solo in five days (281 commits): a **Rust recorder** for frame timing, system state, network and input, plus a Tauri and React app with 12 working screens.
- Ran the agent on **opencode as a separate process** with a custom plugin, 9 skills (League of Legends, VALORANT, full diagnosis, player memory), and PowerShell scripts that inspect the machine.
- Routed model calls through a Supabase edge function to OpenRouter (Gemini Flash), falling back to free models. The first full diagnosis ran **63 tool calls in 3.1 minutes for \$0.50** and returned 11 findings.
- Backed it with 22 Supabase migrations, **95 Rust tests**, and a survey of 111 players.

CTF Solver (*Autonomous Capture-the-Flag Agents for Practice CTFs*) | Claude Code, Python, Discord API, browser automation (2026)

- Built a router and solver setup on Claude Code for practice CTF challenges: the router starts one background solver agent per challenge link.
- Each solver reads the challenge through browser automation, downloads the files, opens its own Discord channel, works the challenge (web, crypto, forensics, reverse engineering, pwn, OSINT), then commits a write-up and posts the result. Wrote 54 of 63 commits.

Presence (*WiFi-Sensing Elder Care, entry, AI Ready ASEAN Youth Challenge 2026*) | Next.js, TypeScript, Ollama (Llama 3.1 8B), WebSockets (2026)

- Built a prototype that watches over an elderly person through WiFi signals instead of cameras: a live dashboard of body position, vital signs, and an activity timeline fed by an open-source WiFi sensing service.
- Wrote alert rules for falls, inactivity, and abnormal vitals, with a **local Llama 3.1 8B** explaining each alert, writing a daily summary, and answering family questions from live data. Runs on simulated sensor data.

Polter (*Desktop Ghost That Reads Your Work Rhythm, [GitHub](#)*) | Tauri 2, Rust, React, SQLite, OpenRouter (2026)

- Built a pixel ghost for the Windows desktop that reacts to typing rhythm (never the keys), mouse movement, and app switching, sorting the user into seven moods with a **Rust classifier**.
- Learns a personal baseline over about 30 days, flags unusual stretches, and shows short AI-written observations, with a 7-day chart and a privacy mode. Built solo for DevKada's CodeKada online hackathon, placing 2nd (172 commits, 11 test files).

HyoBot (*Personal-Voice Messenger Reply Drafter*) | Chrome Extension (MV3), JavaScript, OpenRouter (2026)

- Built a Chrome extension that drafts Messenger replies in the owner's texting style from a **voice profile built from 2,160 real messages**.
- Added per-contact draft or auto-send modes with guards (reply delay, quiet hours, daily cap) and per-contact memory, with drafting as the default.

SAMCIS OJT Report Hub (*Internal Tool, Saint Louis University*) | Next.js, TypeScript, Supabase, Drizzle, OpenRouter, Playwright (2026)

- Built an internal tool for SAMCIS faculty that reads student internship reports (PDF or Word), pulls out structured fields, and lets staff review and edit them, with search, charts, and export to 5 formats.
- Extracts with fixed rules first and calls an LLM only when the rules aren't confident. Admin and faculty roles enforced on the server.

Kintsugi (*Dependency Repair Agent*) | FastAPI, PyGithub, E2B, Docker, Next.js, OpenRouter (2025)

- Built an agent that scans a project for outdated or vulnerable dependencies, applies the fix inside an **E2B sandbox**, and opens a pull request with the change.

Cortex (*AI Right-Click Menu for Windows, [GitHub](#)*) | Python, Groq, OpenRouter, Ollama, PyInstaller (2026)

- Built a Windows right-click extension that sorts files into folders, names them from their contents, and summarizes text files, with undo and a choice of Groq, OpenRouter, or local Ollama models. About 64 pytest tests and a GitHub Actions installer build.

Sikap-Salita (*Filipino Sign Language Recognition, team thesis of 7, [GitHub](#)*) | Python, PyTorch, MediaPipe, FastAPI, edge-tts (2026)

- Built the team's real-time demo: MediaPipe reads body and hand points from the webcam, a transformer model predicts the sign over a rolling window of frames, and edge-tts speaks it in English or Filipino.
- Served predictions through FastAPI to a browser front end with a skeleton overlay and video upload. Earlier wrote the landmark extraction and preprocessing pipeline.

Natively-Z (*Personal build of Natively, an open-source AI meeting assistant, [GitHub](#)*) | TypeScript, Electron, ElevenLabs Scribe (2026)

- Added Filipino and Taglish live transcription, spoken mock interviews with an end review, quick opening lines while the main model is still thinking, automatic model fallback, and Windows audio fixes.

- Portfolio Site** (*Motion Film Portfolio, hyowonbernabe.me*) | *HyperFrames, GSAP, Three.js, Lenis, Vercel* (2026)
- Built a scroll-driven portfolio made from a motion film (a 3D intro scene, a real loader, and a skip-to-section menu), with a Playwright performance suite across simulated devices.
- i-tril** (*Lab Reservation System, Saint Louis University*) | *Next.js, TypeScript, PostgreSQL, Prisma, NextAuth* (2025)
- Built a lab booking system with live availability, clash warnings, and QR entry passes that lab custodians scan, plus admin stats and an action log.
- DisProfile** (*Discord Member Filter*) | *Chrome Extension, JavaScript, Vercel* (2026)
- Built a Chrome extension that filters Discord server members by account age and badges, with a small license server that checks, expires, and revokes keys.
- Rift Helper** ([GitHub](#)) | *Java* (2025)
- Built a desktop tool that automates League of Legends queueing through the game client's local API, with no code injection or screen reading (23 downloads).
- Hide Root Guide** ([GitHub](#), 42 stars) | *Android, Magisk* (2025)
- Wrote a step-by-step guide to hiding root from apps that detect it, including banking apps like GCash, tested on Android 15.
- Legion Mode Theme** ([GitHub](#)) | *PowerShell* (2026)
- Built an add-on for Lenovo Legion Toolkit that changes the Windows accent color, wallpaper, and keyboard lighting when the power mode changes, with no background process.
- Messenger-X** (*Fork of Messenger-Revanced*) | *Shell, ReVanced* (2026)
- Added a pre-patch step that renames clashing permissions so the patched Messenger installs next to the Facebook app.
- TAGriculture** (*Cattle Farm App*) | *Kotlin, Android* (2025)
- Built an Android app for cattle farms: livestock records with photos, weight graphs, health alerts, feed and body-condition scoring, and a market value estimate.
- False Position Solver** (*Numerical Methods*) | *Next.js, GSAP, Math.js, KaTeX* (2025)
- Built a web app for the False Position root-finding method, with an animated theory page and a step-by-step solver.
- Other:** PowerPulse (Kotlin, Android, team of 3, 2024), Huffman Tree compressor with a Swing visualizer (Java, 2024), Graph data structure demo (Java, 2024), t1-campus-forms (Next.js, Prisma, 2025).

RESEARCH

- Phantom: Reinforcement Fine-Tuning of a Small Language Model for Autonomous Computer Maintenance** (*CS 411 thesis proposal, team of 6*) (2026)
- Proposes training a 4B model (Qwen3.5, Unsloth, QLoRA) to fix broken Windows machines: fine-tune on a larger teacher model's successful fixes, then GRPO reinforcement learning in deliberately broken VMs, scored by how healthy the PC is afterward. Proposal stage.

TECHNICAL SKILLS

AI/ML: AI Agents, Tool Calling, MCP, RAG, Evals (PromptFoo), LLM Fine-Tuning, LoRA/Unsloth, Knowledge Distillation, Prompt Caching, OpenRouter, Vercel AI SDK, Hugging Face, Ollama, Whisper

Languages & Stack: Python, TypeScript, Rust, C#, Kotlin, Swift, Java, C++, Next.js, React, FastAPI, Flutter, PostgreSQL, Prisma, Supabase, Redis, Docker, Stripe, Vercel, Railway, Git

EDUCATION

Saint Louis University | Bachelor of Science in Computer Science (BSCS) (2023 - 2027)

AWARDS & CERTIFICATIONS

- Champion** | Hack4Gov 5 Cordillera Regional Qualifier (Capture the Flag) (2026)
- 2nd Place** | CodeKada Online Hackathon by DevKada, with Polter (desktop AI companion in Rust and Tauri) (2026)
- Harvard University** | CS50's Artificial Intelligence with Python (2026)
- Databricks** | Generative AI Fundamentals Accreditation (2026)
- Hugging Face** | AI Agents Course, AI Agents Fundamentals (2026)