

Hyowon Bernabe | Full Stack Developer + AI Engineer

Email: contact@hyowonbernabe.me
Contact: +639948640169

LinkedIn: [linkedin.com/in/hyowon-bernabe](https://www.linkedin.com/in/hyowon-bernabe)
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.
 - 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.
- 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.

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.
- 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**.
 - 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.

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)