Hyowon Bernabe — Full Stack Developer + AI Engineer

Email: contact@hyowonbernabe.me | **LinkedIn:** linkedin.com/in/hyowon-bernabe |
Contact: +639948640169 | **GitHub:** github.com/hyowonbernabe | **Portfolio:** hyowonbernabe.me

PROFESSIONAL SUMMARY

Engineer specialized in building end-to-end AI applications, bridging the gap between advanced Model/Agent workflows and production-grade software. Experienced in designing RAG pipelines and fine-tuning LLMs using Python, while leveraging Next.js and TypeScript for scalable user interfaces. Polyglot developer with a strong foundation in systems programming (Rust, C++, Java), dedicated to creating efficient, high-performance AI solutions.

TECHNICAL SKILLS

AI/ML & Languages: AI Agents, LLM Fine-Tuning, RAG, LoRA/Unsloth, Hugging Face, Generative AI, Vector DBs, Python, TypeScript, JavaScript, Kotlin, Java, C++, C#, C, Rust, .NET
Full Stack & Tools: Next.js, React, Node.js, MySQL, PostgreSQL, Prisma, Firebase, Docker, Git

EXPERIENCE

Chief Technology Officer (CTO) | ZuraLog *(2026 - Present)*

- Supported **50+ fitness app integrations** by designing a **plug-and-play MCP server architecture in Python/FastAPI**, enabling new data sources to onboard without modifying the orchestration layer.
- Eliminated third-party API dependencies by writing native **Swift HealthKit** and **Kotlin Health Connect** platform channels, **enabling real-time background health sync** on both iOS and Android.
- Solo-built the **full cross-platform product** across **Flutter for iOS and Android, FastAPI on Railway,** and **Next.js,** achieving **~78% gross margin** through deliberate **Kimi K2.5** LLM selection.

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

- Reduced manual booking errors by **~60%** and resolved schedule conflicts by architecting a **full-stack booking platform with Next.js, PostgreSQL, and Prisma**, implementing automated validation logic and real-time availability checks.
- Autonomously handled **~70%** of customer inquiries and cut response times by **~50%** by deploying a **RAG-powered AI chatbot** integrated with **Vercel AI SDK**, enabling semantic search over the business knowledge base.
- Improved revenue visibility and enabled data-driven pricing decisions by developing a **comprehensive analytics dashboard** that tracks real-time KPIs like **occupancy rates and RevPAN** through interactive **Shadcn/ui** visualizations.

PROJECTS

ShadowPrompt *(Portable Discrete Academic Interface)* | *Rust, Win32 API, FastEmbed, Groq* *(2026)*

- Engineered a **zero-install, portable AI assistant in Rust**, allowing **execution directly from USB drives with sub-second latency** by implementing a custom flat-file vector store to bypass heavy database dependencies.
- Designed a **stealth-first user interface** using Win32 API hooks, **enabling 100% visual obscurity** by rendering 1×1 pixel-based status indicators and intercepting global key chords without stealing window focus.
- Implemented **local RAG and vision capabilities** with zero binary overhead, **achieving instant context awareness from on-disk documents** by integrating FastEmbed and leveraging native Windows Media OCR APIs.

Kuroko *(Stealth AI Interview Assistant)* | *C#, Whisper, SQLite Vector, Win32 API* *(2025)*

- Engineered a **real-time speech-to-text pipeline** that delivers context suggestions with **sub-second latency** by integrating OpenAI Whisper with continuous voice activity detection.
- **Improved answer relevance by ~40%** compared to baseline prompting by **building a local RAG system on SQLite with vector extensions** to semantically index and retrieve personal technical documentation.
- Developed a **stealth overlay interface** that remains **100% invisible to screen sharing software** by utilizing low-level Win32 API hooks to render graphics on a separate, uncaptured desktop layer.

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

- Fine-tuned an **8B parameter legal analysis model** on free-tier cloud hardware with **2x faster speeds and 60% less memory usage** by utilizing **Unsloth** and **LoRA** for efficient, resource-constrained training.
- Improved legal reasoning accuracy with a **40% reduction in false positives** by implementing a **Knowledge Distillation** pipeline where a larger teacher model annotated complex training data for the smaller student model.
- Enabled **real-time, private contract analysis** on local consumer devices by architecting a modular **adapter-only inference engine** using **Ollama** that dynamically loads specific legal skillsets without merging heavy model weights.

Messenger Z *(Privacy-First Messenger Modification)* | *Kotlin, Java, JADX, XPosed API, LSPatch* (2025)

- Developed a **root-free privacy modification** that allows **"install-and-forget" utility** on any Android device by using **LSPatch** to statically inject the **XPosed framework** directly into the APK's bytecode.
- Engineered a **"Ghost Mode" feature** for end-to-end encrypted chats by **reverse-engineering obfuscated Native C++ libraries** to intercept and silence read receipt signals before they reached the encryption layer.
- Built a **native-feeling settings UI** that integrates without resource ID conflicts by generating the entire interface programmatically in **Kotlin**, decoupling it completely from the host app's resource table.

EDUCATION

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

CERTIFICATIONS

Harvard University – CS50's Artificial Intelligence with Python (2026)
- Search, ML, neural networks, and optimization projects.

Databricks – Generative AI Fundamentals Accreditation (2026)
- LLM architectures and prompt engineering.

Hugging Face – AI Agents Course, AI Agents Fundamentals (2026)
- Built autonomous AI agents with tool integration.