

Saurabh Wadhwa

AI Engineer — LLM agents, tool-calling, self-hosted inference

w.saurabh@protonmail.com | stringalong.net | github.com/saurabh | linkedin.com/in/wadhwa-saurabh | Phuket, Thailand (UTC+7)

SUMMARY

I build LLM agents that run real businesses end to end, and I host the models that run them. What I care about most is that they keep working for the people who use them: every change ships through a verification harness that drives the product like a real user, so an agent non-technical people rely on does not quietly break. Eight years in full-stack TypeScript, plus Python for the inference and evaluation side, six of them in DeFi and fintech where a wrong number is a lost order. Recent work: a managed AI-assistant platform taken from zero to paying customers, WhatsApp agents for a rental operator and a clinic, and self-hosted tool-calling inference.

INDEPENDENT AI PROJECTS

- **Sidekick — managed AI-assistant platform.** Took it from zero to paying customers. Built provisioning, streaming, and billing, with verification harnesses and end-to-end tests that gate every change. Deep Claude and MCP integration; contributor work on OpenClaw.
- **Lofty — WhatsApp guest-support agent.** Built a guest-support agent for a Phuket vacation-rental operator on the WhatsApp Cloud API: retrieval over the operator's knowledge base, tool use, and a clean handoff to a human when the agent is unsure. A test suite gates every change.
- **Drip Lab — clinic booking and payments agent.** Building a WhatsApp agent that books appointments and takes payment through Stripe for a clinic.
- **Self-hosted inference.** Run a 27B model on my own GPUs with vLLM (3x3090) for tool-calling agents. Cut time to first token from 19s to 9.2s and raised a tool-calling evaluation from 86 to 91.
- **NestAI — DeFi agent.** Eliza-based agent for conversational DeFi strategy execution and in-chat transactions, built at YieldNest.

TECHNICAL SKILLS

LLM and agents: Claude, Model Context Protocol, tool-calling loops, retrieval/RAG, evaluation harnesses, vLLM, self-hosted inference, OpenAI-compatible APIs

Languages: TypeScript (8 years), Python, Solidity (6 years), SQL

Frontend: React, Next.js, React Native, Expo, Tailwind, TanStack Query

Backend: Node.js, Nest.js, Hono, Express.js, GraphQL

Data and infra: Postgres, MongoDB, DynamoDB, Prisma, Drizzle, AWS (Lambda, Kinesis), Docker, Cloudflare Workers

WORK EXPERIENCE

Senior Full Stack Engineer | YieldNest Finance (Remote)

Jan 2025 – Present

- Built NestAI, an Eliza-based AI agent for conversational DeFi strategy execution and in-chat transactions
- Architected the full backend: multi-protocol DeFi integrations, strategy and token APY cron jobs, historical price tracking across chains
- Built the referral program end to end: event indexing, time-weighted rewards, partner multipliers, airdrop snapshot tooling
- Built the YND claim page, which reached a 40% stake conversion rate; contributed across a protocol managing \$500M+ in TVL

Senior Full Stack Engineer | Blueberry Protocol (Remote)

Jul 2024 – Dec 2024

- Contributed to the core position-management interface; built Trading Post (a PancakeSwap fork on Camp Network) and StakeUp Protocol's SDK and dapp

Senior Full Stack Web3 Engineer I Beezie (Contract, Remote)

Feb 2024 – Jun 2024

- Built an activity microservice tracking NFT sales, listings, and cancellations, supporting \$34M in swap volume; points and referral system

Senior Software Engineer I Mediverse (Remote)

Apr 2022 – Mar 2023

- Built and deployed a custodial-wallet key-infrastructure service and integrated it with a social app; Uniswap SDK token swapping

Web3 Integrations Engineer I Onramper (Contract, Remote)

Jan 2022 – Mar 2023

- Integrated the Uniswap SDK into serverless infrastructure (DynamoDB, Kinesis, Lambda), growing purchasable tokens from 20 to thousands of DEX tokens

Full Stack Web3 Developer I Tranche Finance (Remote)

Aug 2020 – Dec 2021

- Built the frontend and API from scratch (React, Express, Postgres); deployed upgradable Solidity contracts integrating Aave, Compound, and Yearn. Tranche grew from \$0 to \$30M in TVL during my tenure

HACKATHONS

- Winner, ETHGlobal HackMoney 2020. Bounties at ETH Online, ETH Bangkok, and ETH India.