

Solomon Mithra

Software Engineer

Stockton, CA | (209) 373-5858 | solomonmithra99@gmail.com
solomon-mithra.dev | github.com/solomon-mithra | linkedin.com/in/solomon-mithra

PROFESSIONAL SUMMARY

Software engineer with an M.S. in Computer Science (GPA 3.9, Phi Kappa Phi) and 3 years across full-stack development, AI systems, and automation. Strong CS fundamentals — data structures, algorithms, and object-oriented design — paired with a live production product built end-to-end (150+ automated tests) and open-source tools (1,000+ npm downloads).

TECHNICAL SKILLS

AI / LLM: LLMs, RAG, prompt engineering, AI agents / agentic workflows, embeddings, vector search, pgvector, vector databases, LangChain, evals, PyTorch / Hugging Face (familiar)

Languages: Python, JavaScript, TypeScript, SQL, Java

CS Foundations: Data structures, algorithms, object-oriented programming, system design

Web & Backend: React, Vite, Next.js, Node.js, Fastify, FastAPI, REST APIs, Hono.js, Streamlit

Databases & Queues: PostgreSQL, Redis, BullMQ, MongoDB, AWS S3, Cloudflare R2, Supabase

Cloud & DevOps: AWS, Vercel, Docker, Caddy, Git, CI/CD, AWS EC2, AWS Lambda, Cloudflare Workers, Cloudflare Pages, Inngest

Practices: Agile/Scrum, unit testing

CERTIFICATIONS

AWS Serverless Demonstrated (AWS), Sep 2026: hands-on lab exam on Lambda, API Gateway, Step Functions, DynamoDB

EXPERIENCE

Founding Engineer

Budgeting Parent (full-time since Jul 2026)

Mar 2026 – Present

Remote — US

- **Worked directly with the CEO** to turn product goals into shipped features, owning each one from scoping and architecture to production as the sole engineer (**AI money copilot**).
- Built a durable Plaid bank-data integration reaching 12,000 financial institutions (incremental cursor-based sync, webhooks) with edge-enforced per-user rate limiting and a freemium billing gate (Clerk + Supabase).
- Re-architected the app into a React + Vite frontend on Cloudflare Pages and a Fastify API on a self-hosted VPS, load-balanced by Caddy across 3 Docker instances with health checks, migrating every endpoint (auth, Plaid, Stripe webhooks, streaming AI chat) with no API contract changes.
- Cut **p95 latency from 1.7s to 177ms at 200 req/s** in a load test, with all 12,000 requests succeeding, while dropping host CPU from 74% to 53% and **monthly infrastructure cost by 47%**.
- Moved data to self-hosted PostgreSQL 17 with per-tenant row-level security, using row locks and idempotency keys so concurrent API instances and repeated webhooks never double-count usage or double-apply a payment.
- Used Redis for per-user rate limiting and a shared cap on concurrent AI model calls across all 3 API instances.
- Moved bank sync, billing reconciliation and account deletion onto BullMQ queues in a separate worker process, writing each job to a Postgres outbox in the same transaction as the data change so jobs survive duplicate deliveries, Redis loss and worker crashes, including a test that kills a worker mid-job.

Software Engineer

Cher (real-estate technology startup)

Oct 2025 – Jun 2026

Remote — US

- Built CRM and loan-workflow systems (FastAPI, PostgreSQL, Cloudflare R2); shipped an AI loan agent in the borrower flow with prompt caching that cut repeated LLM input cost by 90%.

Cybersecurity Research Intern (Part-time)

Central Valley Technology

July 2025 – August 2025

Remote — US

- Rebuilt the company landing page under a tight deadline, raising Lighthouse performance from 27 to 98 and SEO to 100.
- Wrote Python automation for testing and reporting across internal security tooling.

Software Development Engineer

Kennect Inc.

Mar 2023 – Aug 2023

India

- Modernized a Vue 2 monolith into a Nuxt 3 micro-frontend (~25% bundle); built reusable UI components on REST APIs (+18% adoption).

Software Engineer — Automation (RPA)

Wipro

Aug 2021 – Mar 2023

Bengaluru, India

- Built automation processing 5,000+ daily financial transactions; cut error rate from 11% to under 3% and processing time by 35%.

PROJECTS & OPEN SOURCE

- **mem01 — Self-Hosted Memory Service for AI Agents (open source):** Dockerized FastAPI + PostgreSQL/pgvector service that stores agent memory as beliefs with lifecycle operations (supersede, invalidate, merge) and deterministic, token-budgeted retrieval. Scored 80.5% on the LoCoMo long-conversation memory benchmark using about a fifth of the retrieval context of published memory systems, with zero LLM calls on reads.
- **Gateia — AI Verification SDK (open source):** Deterministic validation layer for LLM systems enforcing schema correctness in production AI workflows. 1,000+ npm downloads.

- **Relevect — Local Context Engine for AI Agents (open source)**: Local-first retrieval with semantic search (embeddings, pgvector), automatic re-indexing for PDF/Markdown/text, and a FastAPI interface for AI agents.
- **LiteLLM — OSS contribution (Python)**: Authored PR #31942 fixing the proxy's projected-spend budget alerts to respect daily/weekly budget reset windows instead of calendar month-end, with regression tests.
- **OpenMemory (mem0) — OSS contribution (TypeScript)**: Authored PR #4 adding Grok CLI support to mem0's cross-agent session-porting CLI — bidirectional adapters, round-trip tests, and session-discovery fixes.

EDUCATION

M.S. Computer Science — University of the Pacific, Stockton, CA | Aug 2023 – May 2025 | GPA 3.9 | Phi Kappa Phi (Top 10%)

B.E. Computer Software Engineering — St Joseph Engineering College, Mangaluru | 2017 – 2021