Vimal D

AI Engineer — Production RAG, Agents & LLM Applications

Bengaluru, India · Available to join immediately · Open to full-time remote roles

+91 97391 97492 · vimalring@gmail.com · vimal.design · linkedin.com/in/d-vimal · Live product: app.haznox.in, media.haznox.in

Professional Summary

AI engineer with 1 year of hands-on AI engineering experience, built on 10 years as a Senior UI/UX Designer and Frontend Engineer at Epicor. Successfully launched 2 AI products live for real users: Hazzy AI, a multi-tenant RAG and tool-calling platform (Python/FastAPI, Postgres + pgvector, hybrid search, cross-encoder re-ranking, cited answers) at app.haznox.in, and Haznox Social AI, an AI content and publishing platform. Owns the full lifecycle: architecture, retrieval quality, agent safety, observability, cost and latency tuning, CI/CD and Linux operations. Completed Generative AI training in Python, vector databases, RAG, re-ranking, LLM evaluation and guardrails. Comfortable explaining and defending design decisions with client technical teams, and seeking a full-time role building production LLM systems at enterprise scale.

Career Highlights

- **Successfully launched 2 AI products live for real users** — Hazzy AI (RAG chatbot and AI agent) and Haznox Social AI (AI content and publishing).
- **1 year of AI engineering** across RAG, agents and tool calling, LLM operations and production deployment.
- **4 live SaaS products** designed, built and deployed solo, end to end.
- **10 years at Epicor** as Senior UI/UX Designer and Frontend Engineer across 4+ enterprise products; closed 100% of the UX Tech Debt epic and cut handoff time by about 40%.

Core Skills

- **RAG & retrieval:** chunking and ingestion, hybrid dense + BM25 retrieval, query rewriting and multi-query, parent-document retrieval, cross-encoder re-ranking, citation-grounded answers, pgvector, self-hosted sentence-transformers embeddings

- **Agents & tool calling:** tool registry and connector framework, end-user-scoped execution, human-approval gates on write actions, append-only audit of every retrieval and tool call; builds with Claude Code and custom Claude Skills
- **Guardrails & quality:** answers restricted to approved, tenant-scoped sources; citation enforcement; isolation enforced at the data and tool layer, not the prompt; evaluation test sets, hallucination checks and guardrail design (Generative AI training)
- **LLM operations:** model and provider selection (Anthropic, OpenAI), prompt caching, per-tenant token and cache-hit tracking, Langfuse tracing, Prometheus + Grafana metrics, latency profiling and capacity tuning
- **Security & data isolation:** server-verified tenant isolation, domain allow-listing, service-token internal APIs, no public runtime port, security reviews covering SSRF, webhook verification, auth and secrets
- **Backend & data:** Python, FastAPI, PHP/Laravel, Node.js, REST APIs, PostgreSQL, MySQL, Redis, MinIO, queue workers (Horizon), n8n automation
- **Deployment:** Docker / docker-compose, GitHub Actions CI/CD, Caddy, Linux VPS operations, production monitoring
- **Testing & quality:** Playwright, functional and workflow testing, production validation
- **Product & UX engineering:** BRDs and requirements, architecture planning, design systems, accessibility, React/Vue/Inertia frontends

Professional Experience

Freelance AI Engineer — Independent Projects · Bengaluru, India · May 2026 – Present

- Successfully launched 2 AI products live for real users (Hazzy AI and Haznox Social AI), alongside 2 business SaaS products, owning requirements, architecture, implementation, testing, deployment and production support.
- Architected **Hazzy AI**, a multi-tenant RAG and agent platform: Laravel control plane (tenancy, billing, staff console) plus a Python/FastAPI runtime for ingestion, embeddings, retrieval, LLM calls and tools, linked by a token-authenticated internal API.
- Built the retrieval pipeline end to end: hybrid dense + BM25 search on Postgres/pgvector, query rewriting and multi-query expansion, parent-document retrieval and cross-encoder re-ranking, with answers restricted to tenant-scoped, versioned sources and returned with citations.
- Enforced grounding and safety in code rather than prompts: server-verified tenant ID on every query, embedding and tool call; per-end-user scoping; no write action without explicit human approval; append-only audit trail.

- Built a connector and tool-calling framework (WHMCS, WHM, WooCommerce, Shopify, custom APIs) with a central tool registry; approved write actions run as queued jobs, never inline in the request path.
- Instrumented production LLM usage with Langfuse tracing, Prometheus/Grafana dashboards, prompt caching and per-tenant token and cache-hit reporting for cost control.
- Diagnosed a production out-of-memory failure that dropped a live tenant mid-request; profiled embedding latency (≈1s single request vs ≈4s at 3 concurrent under a 2 GB cap) and resized containers and infrastructure accordingly.
- Moved embeddings from a hosted API (Voyage) to self-hosted sentence-transformers, removing a third-party data hop and per-call cost.
- Ran a production architecture review and repo-wide security scan (SSRF in connectors, payment-webhook verification, auth/session, secrets) and prioritised the gaps found: eval harness, per-tenant rate limiting, deletion propagation and agent prompt-injection defences.
- Shipped **Haznox Social AI** (FastAPI + React): AI captions and poster generation, brand profiles, auto-scheduling, multi-platform publishing and an AI blog pipeline (brief → outline → draft → humanise check → SEO → WordPress publish).
- Built **Haznox Hosting**, a custom WHMCS-style hosting and billing platform, and **Haznox Invoice**, a GST invoicing and billing app.
- Automated client business workflows with LLM-powered n8n flows (AI email auto-reply, enquiry tracking); deployed all services with Docker and GitHub Actions CI to Linux VPS hosts.

Senior UI/UX Designer & Frontend Engineer — Epicor Software · Bengaluru, India · 2016 – May 2026

- Owned end-to-end UX and frontend delivery for enterprise Retail ERP, Manufacturing ERP (BisTrack), e-commerce and payment-gateway products used by global customers.
- Built production frontend in HTML, CSS and JavaScript: responsive components, centralised CSS and reusable layout patterns that reduced frontend rework.
- Designed payment-gateway configuration, transaction, dashboard and live-preview flows, simplifying complex financial workflows for merchants.
- Created reusable components, interaction specs and an icon library; cut design-to-development handoff time by about 40% and scaled patterns across 4+ products.
- Closed 100% of the UX Tech Debt epic, the first full closure in the product line's history.
- Applied AI-assisted development and generative AI tools in frontend delivery and prototyping during the final year.

- Worked daily with product owners, business analysts and engineering teams to translate requirements into shippable specifications.

Selected Production AI Systems

Product	What it does	Stack
Hazzy AI	Multi-tenant RAG chatbot and account-aware agent: hybrid search, re-ranking, citations, guardrails, approval-gated tools	Python, FastAPI, Laravel, Postgres + pgvector, Redis, MinIO, Langfuse, Prometheus/Grafana, Docker
Haznox Social AI	AI captions, posters and blogs with brand profiles, scheduling and multi-platform publishing	FastAPI, React, OpenAI, Playwright rendering
Haznox Hosting	Custom WHMCS-style hosting and billing platform with client management and automation	Laravel, MySQL, WHM/WHMCS integrations
Haznox Invoice	GST invoicing and billing with payment reminders and GST reports	Laravel, MySQL
n8n LLM automations	AI email auto-reply and enquiry tracking for client businesses	n8n (self-hosted), OpenAI, Google APIs

Earlier Experience

- **UI Designer — ZMENIT Technologies** · Oct 2013 – Mar 2015 · Web and mobile interfaces built as responsive HTML/CSS templates.
- **Web Designer — MASNOP Advertising** · Jan 2013 – Aug 2013 · Client websites aligned to brand and business goals.
- **Web / Graphic Designer — ASMAN Technologies** · Dec 2010 – Dec 2012 · Frontend build in HTML, CSS and JavaScript; maintained live sites and handled domain, hosting and server tasks.

Education & Certifications

- Generative AI Program — 2026 · Python, vector databases, RAG, re-ranking, LLM evaluation and guardrails
- Bachelor of Computer Applications (BCA), Bharathiar University — 2010

- Master in Graphical Program — World Multimedia Association, 2010
- Design Thinking — SkillUp, 2021

Languages

English · Tamil ·