

Anup Sedhain

npsedhain@gmail.com | npsedhain.com | linkedin.com/in/npsedhain | github.com/npsedhain

Senior fullstack AI engineer and tech lead (6+ years) who takes LLM products from spec to production in a regulated clinical setting: retrieval, evals, safety, and integration. Founded Airvet's Care Portal and Parent Portal codebases; lead its two-engineer AI team and roadmap.

Experience

Airvet Inc. – Los Angeles, CA

2021 – Present

Sr. Fullstack AI Engineer

Jan 2026 – Present

- Took AtlasVet, a cited clinical-answer assistant for veterinarians, from spec to production in 5 months as primary engineer and release owner (AWS Bedrock, Next.js, Supabase, Langfuse); 180+ vets have asked 2,000+ questions.
- Architected hybrid retrieval across Bedrock, Exa web search, and a partner corpus with dual-query Reciprocal Rank Fusion, Cohere reranking, and per-lane fail-open; delivered 6 s p50 TTFT and <21 s p95 for end-to-end cited answers.
- Made answer faithfulness measurable: a deterministic claim-to-citation validator flags 28% of drafts, backed by Langfuse LLM judges and a shadow-mode Bedrock grounding scorer over 590+ production answers.
- Ran a blinded, server-randomized A/B that streamed two retrieval pipelines side by side to clinicians; the KB+web arm won 2.5:1 (420 votes) and web evidence became the default.
- Shipped patient-document Q&A under a PHI privacy contract that keeps record text out of Bedrock, Exa, and Langfuse; replaced static AWS keys with 15-minute OIDC credentials and least-privilege Terraform IAM.
- Built AI Consult end to end (Go, Next.js, admin) for markets without telehealth: LLM fact extraction feeds a 23-rule deterministic S1–S5 triage classifier, with an LLM reviewer guarding against under-triage.
- Led Phase 2 of the Go call-transcription pipeline (Twilio, SQS, S3, ffmpeg, AssemblyAI), now 100+ calls/day feeding SOAP notes and saving vets 15+ minutes per case; root-caused an ffmpeg bug generating 125 GB of silence per call.
- Created @airvet/event-stream, a shared SSE and push package that replaced 9,200 lines across two portals and sustains 99.96% stream success over 65K production connections.

Software Engineer

May 2025 – Dec 2025

- Founding engineer of the Parent Desktop Portal (Next.js on Go APIs) and sole feature author through v1.0: vet chat with media, SSE and FCM push; now 10.8K logged-in users and 7,900 web signups.
- Built config-driven partner signup with employer SSO/SAML, email-domain, and unique-ID eligibility so partners such as Allstate launch by configuration; 100+ partner organizations now onboard members on web.
- Co-led Care Portal's SSE reliability redesign (backoff, Last-Event-ID replay, cache reconciliation), cutting missed and duplicate message reports by over 95%.
- Built MCP-based developer agents (Claude Code, Codex) that drive JIRA, Figma, GitHub, and Playwright test generation, cutting release cycles from 4 to 2 weeks.

Software Engineer & Tech Lead

May 2021 – Jul 2024

- Founded Care Portal (created the repo; 130 PRs in its first year), the real-time console 20+ vet techs use to work 400+ chat cases a day (Next.js, Go, SSE, Redis Pub/Sub).
- Designed its event-driven architecture, replacing polling with SSE badging that sustains 1,000+ messages/sec at peak; rewrote Twilio in-browser video to fix reconnect failures, rolled out per hospital via Remote Config.
- Tech lead and release owner for three clinical portals: drove MUI v5 and App Router migrations, built the web membership funnels (employer group codes, Stripe subscriptions), and mentored engineers.

Software Engineer, Freelance, Polydelic AS – Oslo, Norway (6+ SaaS platforms, 500K+ users)

Aug 2020 – Jun 2021

Software Engineer, Leapfrog Technology Inc. – Kathmandu, Nepal

Jul 2019 – Dec 2020

Education

University of Florida, M.S. Computer Science, GPA 3.93

Aug 2024 – Dec 2025

- Graduate Research Assistant: built a multistage NLP pipeline (BERT, K-means, TF-IDF) on the HiPerGator HPC cluster that detected AI-generated content across 2M+ BlueSky posts with 92% accuracy and tracked attitude shifts over time.

Kathmandu University, B.E. Computer Engineering, GPA 3.63

Sep 2015 – Jul 2019

Technical Skills

Applied AI: RAG and hybrid retrieval, reranking, structured outputs, guardrails, LLM-as-judge and human-labeled evals, agents and tool use, MCP, context engineering, Langfuse, Braintrust, AWS Bedrock, OpenAI, Gemini, Exa

Engineering & Tooling: TypeScript, Go, Python, Node.js, React/Next.js, SSE, WebRTC, PostgreSQL/Supabase, MongoDB, Redis, AWS (SQS, S3, IAM/OIDC), Terraform, GCP Cloud Run, Docker, Kubernetes; Claude Code (custom skills, plugins, MCP servers), OpenAI Codex, QRSPI spec-driven development, ADRs