

Yasaswi Raj Madari

+1 (206) 380-1891 | yasaswirajmadari@gmail.com | [linkedin.com/in/yasaswi-raj-madari-2b7028182](https://www.linkedin.com/in/yasaswi-raj-madari-2b7028182) | github.com/yasaswiraj | yashstudio.tech

PROFESSIONAL SUMMARY

Software engineer with 4+ years across backend, full-stack, and AI/ML work. Built C++ data pipeline components and AngularJS UIs at FactSet, a full-stack marketplace at SafetyKnights, and currently ship the AI pregnancy assistant and community features at numaa.ai (RAG, LLM routing and guardrails, user memory, voice). Comfortable owning a feature end to end: schema, API, frontend, deployment, and tests. Primary languages: Python, C++, TypeScript.

TECHNICAL SKILLS

Languages: Python, C++, TypeScript, JavaScript, SQL

AI and LLM: RAG, LLM Orchestration, Agentic Workflows, Prompt Engineering, Guardrails, Context Management, LLM Evaluation

AI Tooling: LangChain, LlamaIndex, HuggingFace Transformers, PyTorch, spaCy, OpenRouter, ElevenLabs, Text to Speech, Speech to Text

Backend: FastAPI, NestJS, Node.js, Express, REST APIs, WebSockets, Microservices, JWT Authentication

Frontend: React, Next.js, Angular, AngularJS, TailwindCSS, Vite

Databases: PostgreSQL, Supabase, pgvector, Row Level Security, MongoDB, Redis

Cloud and DevOps: AWS, EC2, S3, Docker, CI/CD, Git, Perforce, Linux, UNIX, gdb, DigitalOcean, Netlify

Testing: Playwright, Cypress, Protractor, Jasmine, Karma

PROFESSIONAL EXPERIENCE

ML/AI Engineer — numaa.ai, Incrivelsoft

Dec 2025 – Present

- Built Ask NuMaa, the AI pregnancy assistant on numaa.ai: a chat interface with multi-turn conversation history, saved chats, and a user Memory feature, backed by FastAPI and NestJS services with an LLM layer on OpenRouter.
- Built the RAG pipeline behind it with text embeddings and pgvector semantic search over curated clinical guidelines so answers are grounded in trusted medical content rather than model memory.
- Implemented question routing with a zero-shot classifier and explicit state machine: off-topic requests are declined, ambiguous ones get a clarifying question, in-scope ones go to the right handler, and users can be handed off to a real doctor when needed.
- Built the Memory and Add Context features on a MongoDB-backed user context store (profile, pregnancy stage, preferences, prior chats) with embedding-based lookup so the assistant personalizes responses across sessions.
- Built the Mom to Mom community feature (posts, replies, moderation) and the AI companion chat that runs alongside it, sharing the same auth, user profile, and context store.
- Integrated ElevenLabs for streaming speech-to-text and text-to-speech so users can talk to the assistant instead of typing.
- Containerized every service with Docker and deployed to AWS behind a reverse proxy; set up CI/CD so services build, test, and scale independently.
- Stack: FastAPI, NestJS, React (Vite) + TypeScript + TailwindCSS, Supabase (RLS), pgvector, MongoDB, OpenRouter, ElevenLabs, Docker, AWS.

Software Engineer — SafetyKnights

Jan 2025 – Jun 2025

- Built a bidding marketplace where EHS (environmental, health, and safety) professionals bid on client job postings; owned the feature work across the Angular 19 frontend and Node/Express REST backend.
- Designed the PostgreSQL schema for users, job postings, and bids, and added indexes for the listing and search queries that showed up as slow in early testing.
- Implemented JWT-based authentication with role-level access control for client and professional accounts, plus request validation on all write endpoints.
- Built the listing, bidding, and dashboard views in Angular 19 + TailwindCSS and wrote integration tests for the core API flows before handing the codebase off.

Software Engineer — FactSet

Jun 2022 – Jul 2024

- Developed and maintained C++ modules in a financial data pipeline that feed REST microservices consumed by client-facing dashboards; optimization work on the data processing paths improved throughput by 25% and cut end-to-end delivery latency.
- Built and maintained AngularJS 1.x UI components for client-facing financial products, working from JIRA tickets through code review and production release.
- Debugged production and runtime issues in UNIX environments with gdb and log analysis; managed source through Perforce with JIRA sprint workflows on a 15-engineer team.
- Migrated the frontend test suite from Protractor to Cypress and then to Playwright, cutting execution time by 35% and reducing flaky tests by 20%, which got CI feedback down to same-day.

Software Development Engineer — *coding.Studio()*, Hyderabad, India

Senior Software Development Engineer

SDE III

Jan 2020 – May 2022

Mar 2021 – May 2022

Jan 2020 – Feb 2021

- Led full-stack development of A.R.M.A and Relief Riders, two React/Node.js platforms that replaced paper-based request and approval workflows for VBIT and Relief Riders Hyderabad (details under Projects).
- Owned both codebases end to end, from UML system design and UI mockups through implementation, across two major A.R.M.A releases and 1,000+ commits on Relief Riders.
- Mentored junior club engineers as Senior SDE: set up code review practices and reusable frontend/backend patterns used by a rotating contributor team.

PROJECTS

YashStudio.tech — RAG-Powered AI Portfolio | *Python · Next.js · Supabase · pgvector · ElevenLabs*

- Built a recruiter-facing assistant that answers questions about my experience using RAG over structured resume and project data stored in Supabase with pgvector embeddings; retrieval is scoped so it does not invent experience that is not in the data.
- Implemented token-aware context management across multi-turn sessions to keep prompt size bounded instead of re-sending full history; cut average cost per conversation by 40% without losing coherence.
- Shipped three interaction modes (standard UI, chat, and voice via ElevenLabs TTS/STT) backed by a single Python LLM service that returns schema-validated JSON.

A.R.M.A — Automated Requests Management Application | *React · Node.js · PostgreSQL* | 2021–2022

- Built a request-management platform for VBIT (3,000-student institution) that digitized permission requests, letter generation, and hierarchical event approvals, replacing a manual paper workflow with a role-based pipeline; cut administrative processing time by 40% across 10+ campus forums and eliminated scheduling conflicts.
- Designed the database schema and REST API; built approval dashboards and real-time event statistics views used by 5 administrative departments, plus reusable frontend primitives (paginated/filterable tables, modal dialogs) shared across two major versions.

Relief Riders — Requests Management Application | *React · Node.js · JavaScript*

- Built a dual-portal platform for Relief Riders Hyderabad with separately deployable requester-facing and admin services so volunteers can intake, triage, and track relief requests without touching the public app.
- Drove the project from UML design and UI mockups to a working system; 1,000+ commits, 12 stars, and 16 forks on GitHub.

Debugging C — Competitive Coding Platform | *IEEE VBIT SB* | *Full Stack* | 2021

- Sole developer of a full-stack C programming contest platform used by 500+ participants at an IEEE VBIT SB event; built the backend, frontend, and database schema, deployed on a PaaS with CI/CD, structured logging, and monitoring, held 99% uptime during event peaks, and cut manual judging effort by 80%.

EDUCATION

University at Buffalo, SUNY

M.S. Computer Science & Engineering (ML/AI Focus)

Aug 2024 – Jan 2026

Vignana Bharathi Institute of Technology

B.Tech Computer Science Engineering

Aug 2018 – Aug 2022

ACHIEVEMENTS & LEADERSHIP

- Google Foobar Challenge — completed all 5 levels
- All India Rank 17 / 4,000 — IEEEExtreme 14.0
- Winner — IEEE Code Hustle | Runner-up — CodeCraft Competitive Programming
- Senior SDE — *coding.Studio()* | Web Master — IEEE VBIT SB | Design Lead — Odyssey Literature Festival