

PRACHI RAVI SAIBEWAR

Orlando, FL | Open to Relocation | +1 386-456-8740 | prachi.saibewar.cs@gmail.com | [LinkedIn](#) | [GitHub](#)

PROFESSIONAL SUMMARY

Software Engineer with 3.5+ years of experience building and owning production AI-enabled, backend, and full-stack systems end to end, from system design and API development through implementation, testing, deployment, observability, and production support. Experienced with Python/FastAPI, C#/.NET, React/TypeScript, SQL, and AWS, with hands-on work across Generative AI, LLMs, RAG, agentic workflows, real-time decisioning, distributed systems, evaluation, performance optimization, and reliability.

TECHNICAL SKILLS

Languages & Development: Python, C#, TypeScript, JavaScript, SQL, FastAPI, .NET, ASP.NET Core, Node.js, Express.js, React, REST APIs, Microservices

AI / LLM: Generative AI, LLMs, RAG, LangGraph, Embeddings, Vector Search, Structured Tool Calling, Prompt Engineering, LLM Evaluation, Guardrails, Human-in-the-Loop

Data & Cloud: PostgreSQL, pgvector, Redis, SQL Server, Oracle SQL/PLSQL, AWS (S3, SQS, Batch), Event-Driven Architecture

DevOps & Observability: Docker, CI/CD, Jest, Playwright, Postman, OpenTelemetry, Prometheus, PromQL, Chronosphere

AI-Assisted Development: Cursor, Claude Code, GitHub Copilot

EXPERIENCE

AI Software Engineer, Rokt

Sep 2025 – Aug 2026 | Boca Raton, FL

- Built Python/FastAPI backend services for real-time next-best-action ranking, combining first-party, cart, order-value, behavioral, and transaction signals to score eligible products and offers; improved NDCG@10 from 0.74 to 0.82 across 50K+ offline evaluation impressions.
- Integrated identity-resolved and predictive customer signals from Rokt mParticle through C#/.NET (ASP.NET Core) REST APIs, AWS S3, PostgreSQL, and event-driven data workflows, increasing eligible-signal coverage by 23% while enforcing GDPR/CCPA consent and data-governance controls.
- Developed Generative AI/LLM experimentation workflows for Rokt's Adaptive Content Engine using foundation-model APIs, prompt engineering, and structured outputs; automated LLM evaluation and guardrail checks for quality, brand safety, and schema compliance reduced rejected generations from 12.4% to 4.8% across regression test sets.
- Reduced p95 real-time decisioning latency from 185 ms to 132 ms through Redis caching, asynchronous parallel signal retrieval, and API-call consolidation while maintaining established ranking-quality thresholds.
- Built a React/TypeScript AI decision-debugging console surfacing identity signals, eligibility rules, ranked experiences, decision scores, latency, and outcome telemetry; instrumented workflows with OpenTelemetry and Prometheus/PromQL in Chronosphere, reducing investigation time by 30%.

Research Engineer, University of Central Florida

Jun 2025 – Sep 2025 | Orlando, FL

- Modernized the Chartalist research platform by migrating legacy PHP functionality to reusable React/TypeScript components, restructuring shared frontend logic to improve maintainability, testing, and iterative feature development.
- Implemented and optimized AlphaCore graph-analysis algorithms in Python, NetworkX, and NumPy, reducing repeated computation through vectorized operations and caching and improving experimental runtime by ~30% while validating changes with automated tests and pull-request review.
- Engineered Python feature-engineering pipelines for Bitcoin chainlet-orbit analysis, transforming raw transaction structures into 48-dimensional daily feature vectors and reproducible time-series datasets for downstream graph analytics and machine-learning experiments.

Senior Systems Engineer (Software Engineering), Infosys

Jun 2021 – Jul 2023 | Remote, India

- Developed enterprise full-stack financial applications using React, TypeScript, Node.js/Express, REST APIs, and Oracle SQL/PLSQL, supporting production workflows processing 3,000+ daily customer and account transactions.
- Improved API response times by 25% by profiling high-traffic service paths, optimizing Oracle SQL/PLSQL queries and indexes, and refactoring Node.js data-access logic that introduced unnecessary database round trips and processing overhead.
- Built REST-based healthcare integrations for member, benefits, and eligibility data using SQL validation, source-to-target mappings, and data-transformation rules, supporting 5,000+ daily transactions and reducing integration/data-quality defects by 20%.
- Reduced regression defects by 18% by expanding Jest unit coverage, Postman API suites, and end-to-end integration testing across new features and business-critical existing workflows.
- Reduced production issue-resolution time by 40% through root-cause analysis across React clients, REST services, application logic, and Oracle databases, tracing data inconsistencies and coordinating validated fixes through production releases.

PROJECTS

IncidentIQ - Full-Stack Agentic Production Support Platform

React, TypeScript, Python, FastAPI, LangGraph, RAG, pgvector, PostgreSQL, Langfuse, Docker

- Built a full-stack production-support agent combining pgvector-based RAG, stateful LangGraph orchestration, structured tool calling, and human approval workflows.

EDUCATION

Master of Science in Computer Science | University of Central Florida • Orlando, FL | GPA: 3.9

Aug 2023 – May 2025

Bachelor of Engineering in Computer Science | Savitribai Phule Pune University • Pune, India

Aug 2016 – May 2020