PRASAD NIMBAJI CHAUDHARI

AI Engineering Lead | Full Stack Architect | Generative AI & Multi-Tenant SaaS

+91 7798917132 | Pune, Maharashtra, India | pnchaudhari1996@gmail.com
LinkedIn: linkedin.com/in/prasad-chaudhari1996 | GitHub: github.com/prasad98765

PROFESSIONAL SUMMARY

AI Engineering Lead & Full Stack Architect with 5+ years building production-grade, cloud-native systems — with a deep specialization in **Generative AI, LangChain, RAG pipelines, and agentic workflows**. Designed and shipped enterprise AI platforms (WhatsApp & Web) that cut support workload by **70%** and improved response accuracy by **60%** using retrieval-augmented generation and vector search.

Proven architect of **event-driven microservices, async queue-based pipelines (SQS/Celery), and multi-model AI systems** across Node.js, Python, Django, and AWS — designing for high concurrency, fault tolerance, and 99.9% uptime. Experienced in **LangGraph multi-agent orchestration, LangSmith observability, OpenRouter model routing, and vector DB integration** for context-aware AI responses. Leads cross-functional engineering teams, translating complex AI/product requirements into reliable, scalable, zero-rollback production deployments.

Owned a **multi-tenant university admissions SaaS platform** end to end — from client requirement discussions and project setup to architecture, payment gateway integration (BillDesk, Easebuzz), omnichannel communications (SMS, WhatsApp, Email, Voice), and admin analytics — serving multiple national-level entrance exam programmes on a single platform.

EDUCATION

Bachelor of Engineering (B.E.), Pimpri Chinchwad College of Engineering, Pune — 2015 – 2019

TECHNICAL SKILLS

AI & Generative AI	LangChain, LangGraph, LangSmith, RAG, Prompt Engineering, Agentic Workflows, OpenRouter APIs, Vector DB (Pinecone / pgvector), AI Chatbots, Hallucination Reduction, MCP (Model Context Protocol)
AI Integrations	OpenAI, Claude (Anthropic), Gemini, Whisper, AWS Rekognition, Multi-model Routing, Computer Vision (Face Detection)
Frontend	React.js, Next.js, Redux, HTML5, CSS3, Tailwind CSS, Responsive Design
Backend & APIs	Node.js, Express.js, Python, Django, Django REST Framework, REST APIs, Microservices, Serverless Architecture, Celery, WebSockets, Webhooks
Databases & Caching	MongoDB (Atlas), PostgreSQL, Redis, Valkey, Database Indexing, Connection Pooling
Cloud & DevOps	AWS (EC2, S3, Lambda, SQS, EventBridge, Rekognition), Docker, Kubernetes, Jenkins, CI/CD, Gunicorn, Sentry, k6 Load Testing
Payments & Messaging	BillDesk, Easebuzz, Stripe, WhatsApp Business API (Meta), Smartping SMS, Bulk SMS Gateways, Elastic Email, Click-to-Call (VoIP), n8n
Security & Access	JWT Authentication, OTP Verification, Role-Based Access Control (RBAC), PII Masking, Multi-Tenancy, Data Isolation
Architecture	Event-Driven Architecture, Async Queue Pipelines, Multi-Agent Systems, Multi-Tenant SaaS, System Design (HLD/LLD)
Programming Languages	JavaScript (ES6+), TypeScript, Python
Engineering Practices	Agile (Scrum), TDD, Performance Optimization, Observability & Monitoring, Client Requirement Gathering, Stakeholder Management

EXPERIENCE

Engineering Team Lead (Full Stack / AI) — Jun 2023 – Present
Geta AI (SRV Media), Pune — *Pune, India*

- Architected and led end-to-end development of an **enterprise AI chatbot platform** (WhatsApp & Web), reducing manual support workload by **70%** across production deployments.
- Designed **LangChain + RAG pipelines** with vector search and context-window management, improving response relevance and contextual accuracy by **60%** and significantly reducing hallucinations.

- Built **LangGraph-based multi-agent orchestration** for complex, multi-step AI workflows with tool-calling, memory, and state management.
- Implemented **LangSmith observability** for tracing, evaluation, and monitoring of LLM pipelines in production, enabling data-driven prompt optimization.
- Designed **OpenRouter-based multi-model routing** to dynamically select optimal LLMs (OpenAI, Claude, Gemini) based on task complexity and cost constraints.
- Integrated CRM systems and third-party APIs, reducing average query resolution time by **45%**.
- Architected **event-driven microservices** using Node.js, Python, and async queue pipelines supporting high-concurrency AI workloads with fault tolerance.
- Led **client discussions and requirement gathering** with university admissions teams for an online application form platform; converted business needs into technical specifications, sprint plans, and delivery milestones.
- **Set up the project from scratch** — system architecture, repositories, staging and production environments, Docker/Kubernetes deployments, CI/CD, and monitoring — and took it live for multiple entrance exam programmes.
- Acted as the primary technical point of contact for the client: ran demos and UAT, handled change requests during live admission cycles, and resolved production issues against tight application deadlines.
- Mentored and managed a **5-member engineering team**, improving release stability and reducing post-release issues by **20%**.

Software Engineer Jan 2021 – Jun 2023
Turtlemint, Pune *Pune, India*

- Developed core modules for insurance comparison and renewal platforms serving **5M+ active users**.
- Built scalable APIs and advisor tools (**TurtlemintPro, Turtlefin**), contributing to a **30% increase in lead-to-policy conversion**.
- Engineered a **schema-driven Form Builder Framework** (React, TypeScript, JSON Schema) used across multiple insurance products, reducing feature development effort by **60%**.
- Integrated **Signzy KYC APIs** for automated identity verification, reducing customer onboarding time by **40%**.
- Optimized backend services and database queries, reducing page load time and improving system throughput.
- Actively led system design discussions, code reviews, and performance optimization initiatives.

Software Engineer (Fellowship Program) Nov 2019 – Dec 2020
BridgeLabz, Mumbai *Mumbai, India*

- Delivered full-stack client projects using **Agile and TDD practices**; built **CI/CD pipelines** for stable, predictable release cycles.

PROJECTS

Geta AI Chatbot Platform (WhatsApp & Web)
LangChain, LangGraph, LangSmith, RAG, OpenRouter, Node.js, Python, React, AWS, Vector DB

- Designed and built an **enterprise-grade multi-channel AI chatbot platform** with RAG pipelines, vector search, and LangGraph agentic workflows for context-aware, hallucination-reduced responses.
- Implemented **LangSmith tracing & evaluation** to monitor LLM quality metrics (relevance, faithfulness) in production and drive continuous prompt improvements.
- Engineered **OpenRouter-based model routing** to switch between OpenAI, Claude, and Gemini models dynamically based on cost and task type.
- Built secure REST APIs and webhook integrations for WhatsApp Business API and CRM systems; reduced support workload by **70%** and query resolution time by **45%**.

GETA AI-Driven Application Form Platform (University Admissions SaaS)
Next.js, React, Python, Django, Node.js, MongoDB Atlas, Redis, Celery, AWS (S3, SQS), Docker, Kubernetes, MCP (MongoMCP), DigiLocker API, BillDesk, Easebuzz, WhatsApp Business API, Smartping SMS, Matomo, Sentry, k6

- Architected an **end-to-end AI-driven admissions platform** handling the full applicant lifecycle — signup with OTP verification, form submission, document verification, test and course selection, payments, analytics, and drip marketing — supporting **3,000+ TPS** via horizontal scaling, worker pools, MongoDB connection pooling, and Redis-backed rate limiting, validated through bulk load testing.

- Built a **multi-tenant architecture** running several entrance exam programmes on one backend, with a dedicated MongoDB database per workspace, domain-based tenant routing, and per-tenant configuration cached in Redis.
- Integrated **BillDesk and Easebuzz payment gateways** with split settlement, per-tenant credentials, a transaction ledger, refund handling, and asynchronous post-payment processing.
- Built a standalone **omnichannel communications service** (SMS, WhatsApp, Email, Voice) on Celery, AWS SQS, and Redis — bulk broadcasts to **100K+ recipients**, ~**2,400 messages/sec** ingest, delivery-callback tracking, and per-tenant API keys for data isolation.
- Designed a **credit-based billing system** for messaging with atomic debits, idempotent refunds, and an audit ledger, plus a billing page for balance and expense tracking.
- Built the **admin panel**: role-based access control, server-side PII masking, 14+ analytics reports (source/UTM, payments, city and course reports), bulk filters, and background CSV export to S3 for large datasets.
- Integrated **DigiLocker API** for real-time government-verified document fetching (Aadhaar, marksheets, certificates), eliminating manual uploads.
- Built an **AI profile-photo validation** pipeline combining an in-browser face detection model with a server-side LLM check to reject non-compliant photos before submission.
- Built an **AI screen-share chatbot** (RAG-based) guiding applicants step-by-step through the form — detecting errors, suggesting corrections, answering queries contextually in real time.
- Implemented **MongoMCP (Model Context Protocol)** for natural-language admin access to MongoDB — enabling record queries, status filters, and data insights without writing queries.
- Built **UTM + Matomo** self-hosted analytics and an **AI-driven drip marketing engine** (WhatsApp, SMS, Email) with behavior-triggered sequences based on application stage and inactivity signals.
- Diagnosed production latency with k6 load tests and Sentry — tuned MongoDB connection pooling, added indexes, and fixed worker timeouts, removing a **2–3 second** delay on admin APIs.

Billing & Notification Automation Engine

Python, React, AWS Lambda, SQS, RDS, CloudWatch, Stripe API

- Architected an **event-driven serverless automation engine** for billing, notifications, and reporting using AWS Lambda and SQS; integrated **Stripe APIs** for subscription and payment management.
- Automated email, SMS, and WhatsApp notifications via real-time event triggers, reducing manual operational effort by **75%**.

Omnichannel Marketing & Campaign Management Platform

Python, React, MongoDB, AWS, Queue-based Processing

- Built a scalable omnichannel campaign platform (WhatsApp, SMS, Email) with **queue-based async delivery** handling **10K+ daily messages** and real-time analytics dashboards for delivery and engagement tracking.

CRM Contact Data Management System

Python, React, MongoDB, REST APIs, AWS

- Built a centralized CRM with **data validation, deduplication, and normalization pipelines**; designed secure REST APIs for lead ingestion and third-party integrations.

System Reliability & Performance Optimization (Cross-Projects)

- Achieved **99.9% system uptime** through proactive monitoring, caching strategies, DB indexing, and safe-rollback CI/CD pipelines.

ACHIEVEMENTS

- Recognized **3x as a top performer** for innovation and high-impact AI product delivery.
- Successfully delivered **15+ production deployments** with zero rollbacks across AI and full-stack systems.
- Improved team onboarding efficiency by **25%** through structured mentoring and technical documentation.
- Reduced production release bugs by **20%** using **TDD** and peer code reviews.
- Praised for leadership, cross-functional collaboration, and consistently meeting delivery deadlines.