KALERU ABHIRAM

Full Stack AI Engineer

+91 7993119262 • abhiramkaleru@gmail.com • Hyderabad, India
https://www.linkedin.com/in/abhiram-kaleru-a81281246 • https://abhiramkaleru.web.app

PROFESSIONAL SUMMARY

Full Stack AI Engineer (Backend-focused) with 2.8+ years building scalable fintech, distributed, and AI-powered systems. Currently architecting a multi-tenant AI marketing SaaS platform with LLM-powered agents and production RAG pipelines. Previously led system design for a live crypto platform with 10,000+ users and 80,000+ daily events, and shipped a production RAG-based AI support chatbot (LangChain, Gemini, Qdrant) live on jaimax.com. Core expertise: Node.js, Fastify, FastAPI, BullMQ, Redis, PostgreSQL, MongoDB, and AWS. Proven track record shipping end-to-end products from architecture to production.

KEY ACHIEVEMENTS

- Architecting Business AI, a multi-tenant AI marketing SaaS platform with an LLM-powered agent (Claude, Vercel AI SDK) spanning RAG, content generation, social publishing, and campaign analytics.
- Implemented output validation, structured JSON schemas, and function-calling guardrails to ensure reliable, production-grade LLM responses.
- Reduced transaction failure rate by 30% — designed event-driven architecture with BullMQ, retry logic, and idempotency.
- Improved API response time by 50% and cut DB load by 40% — implemented Redis caching layer across high-traffic endpoints.
- Scaled platform to 10,000+ active users processing 80,000+ daily events — highly available AWS infrastructure (EC2, ALB, VPC, CloudFront).
- Eliminated manual KYC verification entirely by building an OCR-based automated KYC validation system using Tesseract OCR / Google Cloud Vision across multiple verification workflows.
- Engineered three-phase JMC Mining System (Normal, Recovery, Resurrection) — zero revenue loss during any outage cycle.
- Sole engineer for JAIMAX FORCE HRMS — full backend and frontend delivered independently, used daily by the entire organization.
- Built production RAG AI support chatbot (LangChain, Gemini, Qdrant) — deployed live on jaimax.com, eliminating manual support queries with 24/7 instant AI responses.

TECHNICAL SKILLS

Core:	Node.js, Fastify, FastAPI, BullMQ, Redis, PostgreSQL, MongoDB, AWS, React.js, Next.js
Languages:	JavaScript (ES6+), Python, HTML5, CSS3
Backend:	Express.js, Fastify, REST APIs, MVC, Microservices, Event-Driven Systems, SOLID, ACID
Frontend:	React.js, Next.js, Tailwind CSS, Bootstrap, Redux Toolkit, RTK Query
Databases:	PostgreSQL, MongoDB, Redis, Cloud Firestore — Polyglot Persistence, H/V Scaling
Msg Brokers:	BullMQ, Redis Queues, Worker Threads, Cron Jobs, Event-Driven Architecture (DLQ, Retry, Idempotency)
Cloud/DevOps:	AWS (EC2, VPC, ALB, S3, CloudFront, SSM), Docker, CI/CD, Git
Security:	JWT, OAuth 2.0, AES Encryption, DigiLocker KYC, OWASP Top 10, RBAC, Rate Limiting, Multi-Tenant Isolation, Advisory Locks
Payments:	Razorpay, PhonePe, PayU, OCR Verification (Tesseract, Google Cloud Vision), Webhooks
Blockchain:	Smart Contracts, Custodial Wallets, Token Swaps, Web3.js, Ethers.js, MetaMask, Binance BEP20
Observability:	Distributed Tracing, Application Logging, Prometheus, Grafana, Loki, Promtail
Secondary:	WebSockets, Firebase (Auth, Firestore, Cloud Functions, Push Notifications), Formik, Yup, Mailgun, SendGrid, Amazon SES, AdMob
AI & LLM:	RAG (Retrieval-Augmented Generation), LangChain, Claude, Vercel AI SDK, Gemini API, OpenAI, fal.ai, Pinecone, Qdrant (Vector DB), Prompt Engineering, Embedding Pipelines, AI Chatbot Development, AI Image/Video Generation, Structured Output / JSON Schema, Function Calling, Model Evaluation & Regression Testing, Guardrails
Integrations:	Meta, Google, LinkedIn APIs — Advertising, Leads, and Analytics
Tools:	Postman, Swagger, VS Code, GitHub, GitLab

PROFESSIONAL EXPERIENCE

Full Stack AI Engineer | **GALAXY TECH SOLUTIONS PRIVATE LIMITED** *Jul 2026 – Present*
AI Marketing & Business Agency Platform | Hyderabad | On-site

- Architected Business AI, a multi-tenant AI marketing SaaS platform using Next.js, Fastify, Node.js, PostgreSQL, MongoDB, Pinecone, Redis, and BullMQ, with services spanning API Gateway, Auth, Billing, Admin, Design, and Experts for web/mobile clients.

- Built an LLM-powered AI Agent with Claude and the Vercel AI SDK, supporting RAG, knowledge search, content generation, social publishing, lead analytics, and campaign workflows.
- Engineered production RAG pipelines with document/website ingestion, embeddings, Pinecone tenant isolation, async processing, idempotency, and change detection.
- Developed AI image/video generation using fal.ai, OpenAI, and Gemini with provider abstraction, webhooks, retries, and usage metering.
- Integrated Meta, Google, LinkedIn, and Razorpay for leads, advertising, subscriptions, and billing.
- Implemented defense-in-depth multi-tenancy — advisory locks, idempotent callbacks, webhook isolation, and BullMQ workload fairness.
- Built the Next.js frontend consuming Fastify-based backend microservices, with each service dedicated to a specific platform capability.
- Implemented individual and agency workspace switching — enabling users to move seamlessly between personal and agency-managed contexts.
- Implemented structured output handling and function calling (JSON schema-based tool calls) across LLM Agent workflows for reliable, machine-parseable responses.
- Built evaluation and regression testing harnesses to measure LLM output accuracy and catch prompt/model regressions before deployment.

Tech Stack: Next.js, Fastify, Node.js, PostgreSQL, MongoDB, Pinecone, Redis, BullMQ, Claude, Vercel AI SDK, RAG, fal.ai, OpenAI, Gemini, Structured Outputs, Function Calling, Prompt Versioning, Microservices, REST APIs

Full Stack Engineer | **JAISVIK SOFTWARE SOLUTIONS PRIVATE LIMITED** *Apr 2025 – Jun 2026*

Crypto Fintech | On-site | Promoted from Associate Web Developer

- Promoted to lead all system design, AWS infrastructure, and architecture decisions for the platform.
- Designed event-driven architecture with BullMQ, DLQ & retry logic → reduced transaction failures by 30%.
- Built Redis caching layer across APIs → cut DB load by 40%, response time by 50%.
- Engineered JMC Mining System (3-phase: Normal → Recovery → Resurrection) → zero revenue loss during outages.
- Designed J-Wallet custodial wallet — AES-encrypted key storage, BEP20 deposits, token swaps.
- Built OCR-based payment verification across Razorpay, PhonePe, PayU → eliminated manual reconciliation.
- Deployed highly available AWS stack (EC2, ALB, VPC, S3, CloudFront, SSM).
- Built real-time internal community chat system using WebSockets — enabling live peer messaging across the platform.
- Integrated Firebase push notification system for mining status updates and real-time platform announcements — delivered instantly to all active users.
- Owned full product lifecycle — design, deployment, monitoring, and incident resolution.
- Built a production RAG-based AI customer support chatbot (LangChain, Gemini, Qdrant vector DB) — developed as a live chat widget on jaimax.com for 24/7 automated user support.
- Sole engineer — owned full backend (REST APIs, DB schemas, auth, RBAC, payroll, PDF pipelines) and frontend for JAIMAX FORCE HRMS.
- Built React.js + Tailwind CSS UI for HR, Admin, Manager, and Employee roles with granular access control.
- Integrated Firebase push notifications for real-time logout alerts and session-expiry warnings across all employee roles.
- Delivered: Attendance, Leave management, Payslip generation, dynamic letter PDF export, Feedback & Document modules.

Tech Stack: Node.js, FastAPI, BullMQ, Redis, PostgreSQL, MongoDB, AWS, React.js, WebSockets, Firebase, Ethers.js, Web3.js, LangChain, Gemini API, Qdrant

Associate Web Developer | **JAISVIK SOFTWARE SOLUTIONS PRIVATE LIMITED** *Apr 2024 – Mar 2025*

Crypto Fintech | On-site

- Delivered end-to-end KYC, payments, dashboards, and backend automation for a live crypto product.
- Built DigiLocker KYC automation → real-time identity verification, Indian regulatory compliance.
- Integrated Razorpay & PhonePe — secure transaction handling, webhook processing, automated reconciliation.
- Designed BullMQ job queues with dedicated workers for mining ops, background tasks, and scheduling.
- Applied Redis distributed locking → eliminated race conditions in concurrent mining workflows.
- Migrated payment & order services from monolith to microservices with secure proxy routing.
- Built RBAC role dashboards, RTK Query API integrations, and dynamic PDF reporting UI.

Tech Stack: Node.js, Express.js, BullMQ, Redis, MongoDB, PostgreSQL, React.js, RTK Query, DigiLocker API

KEY PROJECTS

▸ Business AI – Multi-Tenant AI Marketing SaaS Platform (GALAXY TECH SOLUTIONS PRIVATE LIMITED)

- Multi-tenant architecture across API Gateway, Auth, Billing, Admin, Design, and Experts microservices for web/mobile clients.
- LLM-powered AI Agent (Claude, Vercel AI SDK) — RAG, knowledge search, content generation, social publishing, lead analytics, campaign workflows.
- Production RAG pipelines — document/website ingestion, embeddings, Pinecone tenant isolation, async processing, idempotency, change detection.
- AI image/video generation via fal.ai, OpenAI, and Gemini with provider abstraction, webhooks, retries, and usage metering.
- Meta, Google, LinkedIn, and Razorpay integrations for leads, advertising, subscriptions, and billing.
- Defense-in-depth multi-tenancy — advisory locks, idempotent callbacks, webhook isolation, BullMQ workload fairness.

Tech Stack: Next.js, Fastify, Node.js, PostgreSQL, MongoDB, Pinecone, Redis, BullMQ, Claude, Vercel AI SDK, fal.ai, OpenAI, Gemini API

▸ Jaimax Coin – Crypto Ecosystem & Automation Platform (JAISVIK SOFTWARE SOLUTIONS PRIVATE LIMITED)

- 10,000+ active users • 80,000+ daily events • Razorpay, PhonePe, PayU multi-gateway payments.
- J-Wallet: AES-encrypted custodial wallet, BEP20 deposits, token swaps • failures reduced by 30% via event-driven BullMQ architecture.
- Real-time internal community chat via WebSockets • Firebase push notifications for mining updates and platform announcements.
- Full observability: Prometheus, Grafana, Loki, Promtail • Highly available AWS deployment.

Tech Stack: Node.js, BullMQ, Redis, PostgreSQL, MongoDB, AWS (EC2, ALB, VPC, S3), React.js, WebSockets, Firebase, Web3.js
Live: https://www.jaimax.com

▸ Jaimax AI Customer Support Chatbot — RAG Pipeline (JAISVIK SOFTWARE SOLUTIONS PRIVATE LIMITED)

- Built a production RAG-based AI customer support chatbot deployed as a live chat widget on jaimax.com — handling user queries on platform FAQs, wallet usage, mining, and payment support.
- Architected the full RAG pipeline using LangChain + Gemini — document ingestion, chunking, embedding generation, semantic retrieval via Qdrant vector DB, and response synthesis.
- Built FastAPI backend to serve the RAG pipeline with streaming responses — integrated as a React.js chat widget live on the platform, available 24/7 to all users.
- Eliminated repetitive manual support queries — instant AI-powered answers grounded in verified Jaimax documentation with zero hallucinations via Qdrant semantic retrieval.

Tech Stack: Python, FastAPI, LangChain, Gemini API, Qdrant, React.js, Prompt Engineering, Embedding Pipelines
Live: https://www.jaimax.com

▸ Jaimax Force – Internal HRMS Portal (JAISVIK SOFTWARE SOLUTIONS PRIVATE LIMITED)

- Sole engineer (backend, frontend) — REST APIs, RBAC, payroll engine, PDF generation — entire org uses it daily.
- Modules: Attendance, Leave, Payslip, Salary, Letter generation (PDF), Feedback, Document verification.

Tech Stack: Node.js, Express.js, React.js, Tailwind CSS, MongoDB, PostgreSQL, Redis, JWT
Live: https://www.force.jaimax.com

ADDITIONAL PROJECTS

▸ College Circle – Student Collaboration Platform

- Cross-institution MERN platform: resource sharing, communities, real-time messaging, event discovery.
- Firebase Auth, Cloud Functions, RBAC (student/contributor/admin), Redux/RTK, optimized MongoDB schemas.

Tech Stack: React.js, Redux/RTK, Node.js, Express.js, MongoDB, Firebase Auth, Firebase Cloud Functions
Live: https://www.collegecircle.cc

▸ Thrivewell Healthcare – Full Stack Web Application

- Next.js healthcare platform — appointment booking, SEO-optimized interface, real-time Firestore admin dashboard.
- Fast load times, clean UI/UX, scalable architecture for reliable digital healthcare delivery.

Tech Stack: Next.js, Firebase Firestore, Tailwind CSS, Firebase Cloud Functions
Live: https://www.thrivewell-healthcare.in

EDUCATION & LANGUAGES

B.Tech – Computer Science & Data Science Engineering *2020 – 2024*
Narsimha Reddy Engineering College, Hyderabad | CGPA: 7.27 / 10
Languages: English (Professional) • Telugu (Native) • Hindi (Conversational)

SOFT SKILLS

Problem Solving • System Thinking • Communication • Time Management • Adaptability • Leadership • Active Listening