

TIRTH SHAH

Software/AI Engineer (2+ YOE) | M.Tech in AI @ IIT Jodhpur

+91-9377461119 | ✉ shahtirth0212@gmail.com |  [shahtirthhh](https://github.com/shahtirthhh) |  [shahtirthhh](https://www.linkedin.com/in/shahtirthhh) |  Portfolio |  [void-snippets](https://www.npmjs.com/package/void-snippets)

TECHNICAL SKILLS

Microservices & Messaging: Apache Kafka (KRaft & Zookeeper), Redis, Docker, Kubernetes, GitHub Actions CI/CD

Backend: NestJS, FastAPI, Node.js, WebSocket, REST

Frontend: React, TypeScript, TanStack Query, Socket.IO

AI & LLM: OpenAI, ElevenLabs, Vapi, Retell AI, Ollama (self-hosted LLMs), Gemini, Ideogram

VoIP & Telephony: Twilio, Telnyx, FreeSWITCH, A2P Compliance

Cloud & Databases: AWS (EC2, ECS, S3, Lambda), MongoDB, PostgreSQL, Firebase

Open Source: [void-snippets](https://www.npmjs.com/package/void-snippets) — typed HTTP service layer, TanStack Query hook factory, Socket.IO hooks & routing contracts for React (Creator & Maintainer)

Automation & Integrations: Make, Zapier, Jobber, Google Workspace

WORK EXPERIENCE

Digipine Infotech LLP (*Sep 2024 - Present*)

Full Stack | Microservices | AI

Software Engineer

Apr 2025 – Present

- Led zero-downtime migration from Zookeeper-based Kafka to KRaft mode across 6 microservices, eliminating the Zookeeper dependency and reducing cluster recovery time by ~45% during fault injection tests.
- Architected a multitenant BYOC SaaS platform on AWS ECS with per-tenant Kafka topic isolation, sustaining 120K+ events/day with zero cross-tenant data leakage across 15+ active business accounts.
- Owned end-to-end design of an AI-powered debt recovery engine integrating OpenAI function-calling with Ollama-hosted fallback LLMs, reducing manual agent intervention by ~65% on repetitive recovery workflows.
- Led a 4-person team across 2 parallel product tracks, conducting weekly architecture reviews and driving on-time delivery across all milestones for 2 consecutive quarters.
- Mentored 3 interns through structured code reviews and pair sessions, cutting average onboarding time from 3 weeks to 10 days.

Associate Software Engineer

Sep 2024 – Apr 2025

- Built a real-time conversational AI pipeline connecting Twilio telephony to locally hosted LLMs via Ollama over WebSocket, reducing voice response latency from ~3.1s to ~320ms under 50 concurrent calls.
- Developed a FastAPI service layer exposing Ollama LLM inference over REST, handling 200+ concurrent inference requests at p95 latency under 420ms. Which was containerized and deployed on AWS ECS alongside the NestJS microservice fleet.

OPEN SOURCE

[void-snippets](https://www.npmjs.com/package/void-snippets) (Creator & Maintainer) (Live NPM Package)

PNPM Monorepo | TypeScript | React | TanStack Query | Socket.IO

- Developed and published a 3-package TypeScript monorepo ([@void-snippets](https://www.npmjs.com/package/@void-snippets)) reaching 200+ weekly NPM downloads within 3 months of launch through entirely organic growth with zero paid promotion.
- Engineered a typed HTTP service layer, TanStack Query v5 hook factory, Socket.IO hooks, and a type-safe routing contract — eliminating ~3 hours of per-feature scaffolding time per developer and standardizing patterns across 4 internal production codebases at Digipine.

KEY PROJECTS (END-TO-END OWNERSHIP)

Omni-channel Communication & BYOC Platform

MERN | Kafka | Twilio | WhatsApp Business API | AWS ECS | GitHub Actions

- Architected a multitenant BYOC SaaS platform unifying WhatsApp Business API, Twilio SMS, and voice channels into a single interface — onboarding 15+ business tenants with full telecom provider isolation and zero config cross-tenant bleed.
- Built a Kafka-backed event routing layer processing 50K+ real-time communication events/day across all three channels, achieving sub-500ms end-to-end message delivery under concurrent tenant load.

Event-Driven Game Backend Architecture

NestJS | Kafka | Redis | OpenAI | Ideogram | RevenueCat | GitHub Actions

- Created a high-concurrency NestJS backend with Kafka event streaming and Redis caching, sustaining 300+ concurrent users with p99 response times under 180ms — validated under load testing with simulated burst traffic.
- Integrated OpenAI and Ideogram APIs with dynamic real-time prompts to generate personalized stories and images on-demand, serving 1,000+ AI-generated content pieces/day with no manual intervention.
- Implemented RevenueCat in-app purchase and subscription flows, reducing payment integration time from 2 weeks to 3 days via standardized webhook handlers and automated CI/CD via GitHub Actions.

Bulk Call Forwarding Platform

MEAN | Kafka | Redis | GitHub Actions

- Engineered a FreeSWITCH-based bulk call forwarding system sustaining 500+ simultaneous calls with sub-200ms routing latency and even load distribution across forwarding targets.
- Designed Kafka consumer groups for real-time call event processing and Redis-backed routing tables, achieving a 99.5% successful call connection rate under peak traffic conditions.

AI-Powered Debt Management Operations

React | OpenAI Functions | Ollama | FastAPI | GitHub Actions

- Designed a human-in-the-loop AI agent system using OpenAI function-calling and locally hosted Ollama LLMs via FastAPI. Where agents autonomously triaged 1,200+ debt accounts/month, surfacing recommended actions for human approval and reducing case review time by ~70%.
- Built a React dashboard tracking full recovery lifecycle per account — status, contact history, AI recommendations, and agent decisions, giving operations teams a single pane of glass across all active recovery workflows.
- Implemented a fallback routing layer that switched between OpenAI and Ollama at runtime based on latency thresholds, keeping inference costs predictable while maintaining ~98% agent uptime during OpenAI outages.

COLLABORATIVE PROJECTS (MODULE OWNERSHIP)

VoIP Customer Support System

Twilio | ElevenLabs | NestJS | AI Workflow Automation

- Automated missed-lead callbacks using Twilio and ElevenLabs voice AI, recovering ~30% of leads that would otherwise go uncontacted — eliminating manual follow-up for the sales team entirely.

Automated Lead Qualification System

Twilio | ElevenLabs | NestJS

- Built an AI-driven IVR with NestJS that collected lead details conversationally before routing to a human agent, reducing average pre-qualification handle time by ~40% and improving handoff quality.

Multi-Lingual Estimation AI

LLM Fine-tuning | Jobber API

- Fine-tuned an LLM for French-language pricing estimates with direct Jobber CRM sync, cutting estimation turnaround from 2 days to under 10 minutes and eliminating manual data entry for the estimations team.

EDUCATION

Indian Institute of Technology Jodhpur

Master of Technology in Artificial Intelligence

Jodhpur, India

2026 – 2028

- Coursework in machine learning, deep learning, NLP, and MLOps — applied directly to production AI pipelines in concurrent a industry role.

LDRP Institute of Technology and Research

Bachelor of Engineering in Computer Engineering

Gandhinagar, India

Graduated 2024