

Shubham Gupta

+91-9050216555 | shubhamynr22@gmail.com | linkedin.com/in/shubhamynr22 | github.com/shubhamynr22

Profile

AI-focused Backend Engineer with 3+ years of experience building and scaling distributed systems, now specialized in designing multi-agent AI architectures, tool-calling agent harnesses, and LLM-powered product features. Strong foundation in Node.js, NestJS, event-driven pipelines, and AWS cloud infrastructure, combined with hands-on experience building agent orchestration, RAG pipelines, and AI automation for real-world SaaS, Fintech, and consumer AI products. Proven track record in providing systems that support 100k+ concurrent users and shipping production agentic AI systems used by real users.

Technical Skills

- **Languages:** Python, TypeScript, JavaScript, SQL, Solidity
- **Backend & APIs:** Node.js, NestJS, Express.js, REST APIs, GraphQL, Microservices, API Gateway, WebSockets, React.js, Next.js
- **Databases & Caching:** PostgreSQL, MongoDB, Redis
- **Messaging & Events:** Apache Kafka, RabbitMQ, BullMQ, AWS EventBridge, AWS SQS, Event-Driven Architecture
- **Auth & Security:** JWT, OAuth2, RBAC, SSO, Multi-Tenant Access Control
- **Cloud & DevOps:** AWS (EC2, S3, RDS, Lambda, EventBridge, SQS, ALB/NLB, CloudFront), Docker, GitHub Actions, Jenkins, CI/CD, Linux, Git
- **AI & LLM Tools:** OpenAI API, Multi-Agent Orchestration, Tool Calling, Agent Harness Design, Agentic Workflows, Prompt Engineering, LangChain, Retrieval-Augmented Generation (RAG), GitHub Copilot, Cursor AI
- **Integrations:** WhatsApp Business API, Email APIs (Gmail/SMTP-based flows)

Experience

Memorae

April 2026 – Present

Backend AI Engineer

Delhi

- Rearchitected the message delivery infrastructure for Memorae, a memory-focused AI assistant, replacing a cron-based system with an event-driven pipeline using AWS EventBridge and SQS, cutting delivery latency by 50% and achieving near-100% delivery reliability at scale.
- Designed and built a multi-agent architecture from scratch to replace a single prompt-response model, implementing a custom agent harness with an orchestrator, specialized sub-agents, and tool-calling, enabling reliable multi-step task execution and reducing hallucinations.
- Built email as a first-class channel for Memorae, including AI-powered auto-drafting and auto-labelling of incoming emails, and integrated WhatsApp Business API as a primary conversational channel for user interactions.
- Built collaborative boards and lists within Memorae fully operable through natural language, allowing users to create, update, and manage tasks conversationally via the agent system.
- Contributed broadly to product direction and backend architecture, working closely with the founding team on system design decisions as the product scaled.

Simplr

August 2025 – December 2025

Software Engineer – Backend

Gurugram

- Built and maintained backend microservices for identity and authorization modules using Node.js and NestJS, supporting enterprise SSO and multi-tenant access flows for a SaaS platform.
- Designed and enforced fine-grained RBAC policies across distributed services, hardening platform security and reducing unauthorized access incidents.
- Developed ETL pipelines to migrate large-scale enterprise data from Salesforce to AWS, handling schema transformation, data validation, and error recovery across millions of records.
- Resolved critical PostgreSQL performance bottlenecks through targeted indexing and query optimization, reducing average query latency by 40%.

NE Group

June 2024 – July 2025

Software Development Engineer – Backend

Hyderabad

- Rearchitected the API Gateway for a high-traffic gaming platform, improving scalability and reducing latency by 35%, enabling support for 100k+ concurrent users.
- Designed real-time event pipelines using Kafka and Redis, achieving sub-100ms game state updates and processing 10k+ events per second at peak load.
- Built a WebSocket-based communication layer for live multiplayer interactions and real-time score updates, reducing client polling overhead by 60%.
- Developed ERC-20 and ERC-1155 smart contracts for in-game asset management and on-chain betting logic using Solidity.
- Optimized PostgreSQL and MongoDB schemas through denormalization and compound indexing, improving write throughput by 25% and read latency by 30%.

Credain

Software Development Engineer – Backend

Feburary 2023 – April 2024

Mumbai

- Architected the full backend from scratch for a Web3 banking platform, defining service boundaries, API contracts, and data models for a secure multi-chain financial product.
- Built wallet management, transaction processing, and multi-chain asset handling modules, enabling users to interact with multiple blockchains through a unified REST API.
- Implemented async transaction workflows using RabbitMQ, ensuring reliable, idempotent, and fully auditable processing of financial operations with zero data loss.
- Established CI/CD pipelines with automated smart contract testing via GitHub Actions, reducing deployment time by 50% and release risk.

Idea Usher

Blockchain Developer Intern

January 2022 – December 2022

Mohali

- Developed and audited Solidity smart contracts (ERC-20, ERC-721, ERC-1155) for NFT marketplace and DeFi applications, focusing on gas optimization and contract security.

IIT Hyderabad

Blockchain Research Intern

January 2022 – June 2022

Hyderabad

- Researched parallel transaction scheduling in Hyperledger Sawtooth, benchmarking smart contracts to demonstrate measurable throughput improvements under concurrent workloads.

Projects

QueryLens – Natural Language Slow Query Explainer

Docker

Node.js, NestJS, OpenAI API, PostgreSQL,

- Built a schema-aware developer tool that uses OpenAI to explain why a PostgreSQL query is slow, grounding explanations in real EXPLAIN ANALYZE execution plans, table structure, and index availability for non-hallucinated diagnosis.
- Generates runnable index migration snippets and query rewrites with before/after cost comparisons, exposed as a REST API and CLI tool for integration into developer workflows and CI pipelines.
- *GitHub*

IdleWatch – Cloud Resource Waste Detector

Docker

Node.js, AWS SDK, PostgreSQL, React.js, Terraform,

- Built an AWS cost intelligence tool that identifies genuinely idle resources by analyzing CloudWatch utilization metrics over time, scoring waste by estimated monthly cost impact across RDS, Lambda, EBS, and S3.
- Implemented scheduled metric aggregation for trend-based idle detection and auto-generates Terraform snippets for right-sizing or terminating flagged resources.
- *GitHub*

Education

Indian Institute of Information Technology, Sonapat

B.Tech in Information Technology | CGPA: 7.9/10

2019 – 2023

Sonapat

Relevant Coursework: Data Structures & Algorithms, DBMS, Operating Systems, Computer Networks, Object-Oriented Programming