

Shreyas M Kaushik

Email: shreyasmk.mathur@gmail.com · Phone: +91 8867229568

LinkedIn: linkedin.com/in/me-shreyasmk · GitHub: shreyasY2k.github.io · Medium: @shreyasmk.mathur

SUMMARY

AI Engineer with 4+ years building production LLM systems, multi-agent platforms, and secure cloud-native infrastructure. Designed and scaled an on-premises agentic platform to 500+ deployed agents serving 1M+ requests/day for BFSI clients, with built-in AI red-teaming (prompt injection, jailbreak, goal-hijacking) and OPA-backed policy governance. Strong in Python, LLM gateway design, RAG, multi-agent orchestration, and AI security. Startup-forged generalist who owns systems end-to-end.

TECHNICAL SKILLS

- **Agentic AI & LLMs:** Multi-agent orchestration, LLM gateway design, RAG pipelines, tool use / function calling, prompt engineering, OpenAI / Anthropic / Gemini APIs, LiteLLM
- **AI Security & Governance:** AI red-teaming, prompt injection testing, jailbreak & goal-hijacking detection, PII-aware routing, OPA policy, OWASP guardrails, immutable audit logging, multi-tenant isolation
- **Agent Frameworks & Orchestration:** AgentScope, LangChain, LangGraph, Temporal, Netflix Conductor, event-driven architecture
- **Languages:** Python, TypeScript, Golang, JavaScript, SQL, Bash
- **Web & APIs:** FastAPI, Node.js, Django, Express, React.js, RESTful APIs, NGINX
- **Cloud & DevOps:** AWS (EC2, S3, SQS, EKS, ALB, API Gateway, Cognito), Docker, Kubernetes, Helm, GitHub Actions
- **Data & Infrastructure:** PostgreSQL, MongoDB, Redis, Redis Sentinel, Kafka (Redpanda), TypeORM
- **Observability & Auth:** Grafana, Loki, Tempo, Prometheus, Promtail, Keycloak, OAuth 2.0, JWT, MFA

EXPERIENCE

- **Mphasis** Bangalore, India
Feb 2026 - Present
 - **ATOM – Enterprise Agentic Platform:** *AgentScope, TypeScript, Golang, FastAPI, Kubernetes, OPA, PostgreSQL, Redis, Grafana Alloy, Loki, Tempo*
Architected and shipped an on-premises, multi-tenant agentic platform for BFSI clients, now running 500+ production agents serving 1M+ requests/day. Every LLM call flows through GATE (JWT auth, OPA policy, rate limiting, immutable audit log); agents run as isolated containers on Docker/Kubernetes via a custom runtime. Teams author agents through an SDK and deploy with a single command.
 - **AEGIS – AI Red-Teaming & Agent Governance:** *Python, OPA*
Built the platform security layer (Agentic Enterprise Governance & Information Security) that continuously red-teams every deployed agent – automated prompt-injection, jailbreak, and goal-hijacking test suites – and enforces PCI-DSS / GDPR / DORA guardrails and tenant isolation before agents reach production.
 - **NexusAI – Enterprise LLM Gateway:** *LiteLLM, FastAPI, Python, PostgreSQL, Redis, Docker*
Architected a multi-tenant LLM gateway serving BFSI clients with API-key management, per-tenant budget governance, PII-aware routing, and audit logging across multiple model providers.
- **IHX – A Perfios Company** Bangalore, India
Jul 2024 - Feb 2026
 - **Document Intelligence Platform:** *TypeScript, FastAPI, AWS EKS, MongoDB, S3*
Built the document-processing platform for India's largest health benefits administrator, lifting first-instance complete claim submissions from 23% to 77% within 2 weeks. The platform now classifies 4M+ medical images every month.
 - **Workflow Orchestration Engine:** *Netflix Conductor, Zookeeper, Redis Sentinel, TypeScript*
Architected orchestration for document-processing pipelines (extraction, OCR, classification, QC) with dynamic definitions and parallel processing, cutting processing time 40% and improving reliability.
 - **Persistence Engine:** *TypeScript, TypeORM, PostgreSQL, Redis*
Built a single-layer persistence solution that reduced average query time from 1–2s to 10–15ms across 10 tables, materially improving throughput and scalability.
 - **Status Notification Service:** *TypeScript, AWS SQS, AWS SNS, MongoDB*
Built a high-throughput system processing 10M+ messages/day to track state across AI/ML entities, surfacing unified real-time status to clients.
 - **Platform Infrastructure & Security (End-to-End Ownership):** *Kubernetes, Helm, NGINX, AWS, Grafana, Loki, Prometheus, Promtail, Keycloak, OAuth 2.0, JWT*
As an early engineer at a startup, owned infrastructure and access control end-to-end: designed Kubernetes/Helm scaling and hardening for AI-model serving, stood up the full observability stack, and built a Keycloak-based MFA/SSO solution with custom plugins, cutting login time 50%.

PROJECTS

- **Praxis – Open-Source Agent Platform (TypeScript, Golang, FastAPI):** Personal open-source platform for running compliant AI agents on-premises: SDK-authored agents, single-command deploy, a policy gateway enforcing auth / OPA / rate-limits / immutable audit log, isolated agent containers on Docker/Kubernetes, and a Studio UI for lifecycle and human-in-the-loop management.
- **Collaborative Compiler – Real-Time Collaborative Code Editor (JavaScript, WebSockets):** Web-hosted collaborative code editor supporting real-time multi-user editing and autocomplete, architected and built by hand before AI coding assistants existed. github.com/shreyasY2k/Collaborative-Compiler
- **Contra – AI-Assisted 3D Game (Three.js):** Built a complete 3D run-and-gun shooter in 2 hours using AI tooling (GitHub Copilot, Claude), with enemy AI, vehicle physics, and weapon/particle systems. shreyasy2k.github.io/contra
- **Python Spawn (Node.js, npm package):** NPM package bridging Node.js and Python, enabling execution of Python scripts and functions from Node.js applications. github.com/shreyasY2k/python-spawn

EDUCATION

- **Sapthagiri College of Engineering** Bangalore, India
• *B.E., Information Science; CGPA: 8.5* *2018 - 2022*

TECHNICAL WRITING & CERTIFICATIONS

- **Technical Writing:** Multi-part series on Keycloak MFA/SSO (custom plugins, SMS/Email/TOTP/Telegram, centralized user management) – medium.com/@shreyasmk.mathur
- **Certifications:** Architecting with Google Compute Engine (Coursera Specialization); The Modern JavaScript Bootcamp (Udemy)