Tanmay Sule

+91 9850188695 — suletanmay27+resume@gmail.com — linkedin.com/in/tanmay-sule-81a18019a
github.com/tanmaysule — tanmaysule.github.io

Summary

Lead Engineer with 3+ years at Eightfold AI, building and scaling core infrastructure and AI systems. Led the company's first agentic AI initiative (team of 4), owning agent frameworks, production triaging pipelines, and developer tooling. Concurrently own a horizontally-sharded Aurora MySQL platform (40+ clusters, 4 AWS regions) and drove the firm's AWS-to-Azure multi-cloud expansion. Independently architected a multi-tenant agentic RAG platform end-to-end.

Experience

Eightfold AI **Aug 2022 – Present**
Lead Engineer – Core Infrastructure / Platform

Agentic AI Developer Productivity Platform *Technical & Project Lead*

- Launched the company's first agentic AI initiative; recruited and mentored a team from 0 to 4 engineers while establishing org-wide standards for LLM integration, prompt management, and evaluation.
- Designed and built custom 4-layer agent framework on LangGraph/LangChain — reusable abstractions with observability middleware (per-invocation token tracking), multi-model LLM factory, and centralized tool registry.
- Sentry Production Triager — 3-stage pipeline: root cause analysis, code fix classification, and GitHub issue creation with automatic Copilot assignment, reducing time-to-resolution for production incidents. End-to-end AI-to-AI handoff where the agent triages the issue and assigns it to Copilot for the code fix. Redis distributed locking for concurrent issue prevention.
- Build Failure Triager — per-failure root cause analysis with cross-failure deduplication. Same root cause across multiple test failures produces a single GitHub issue, eliminating duplicate engineering effort. Multi-channel Slack reporting.
- Developer Chatbot — Slack-native conversational assistant that answers engineering questions by orchestrating specialized sub-agents. 8-step pipeline with intent classification, dynamic routing, concurrent tool execution, and structured Block Kit responses. Built-in multi-level error recovery.

OLTP Database Platform

- Led strategy and execution for a horizontally-sharded, globally replicated Aurora MySQL database across 40+ clusters in 4 AWS regions. Built internal CLI tooling (Python, Go) for cluster management, schema deployments, and horizontal/vertical scaling.
- Orchestrated zero-downtime major MySQL version upgrades via blue-green deployments.
- Delivered application-level BYoK (Bring Your Own Key) encryption, reducing RDS spend by $5,000/customer/year and strengthening enterprise compliance. Implemented real-time query cost analysis.

Multi-Cloud Architecture

- Core member of a year-long AWS-to-Azure initiative, unblocking enterprise customers with Azure-only mandates and enabling onboarding of the firm's largest client.
- Designed cloud-agnostic service abstractions (Redis, OLTP, Blob Storage/S3), ensuring seamless cross-cloud feature development for product teams.
- Extended in-house distributed async processing platform to Azure, orchestrating 100+ Docker containers via Azure Container Instances; reliably processing 700K+ operations daily.

Cloud-Native Analytics Platform

- Consolidated Redshift + Databricks into a unified StarRocks-based analytics layer, cutting query latency and simplifying BI reporting for product and data teams.
- Designed end-to-end event-driven ETL: streaming entity changelogs via Firehose to S3, orchestrating with SQS and async workers, routing via Airflow DAGs for high-throughput ingestion.

Independent Project

Multi-Tenant Agentic RAG Platform
Sole Architect & Engineer

- Designed and built a full-stack conversational AI platform end-to-end. Core challenge: answering cross-domain business queries that require joining structured database records with unstructured communications (Slack, Email, Notion) — resolving entity references across data sources and producing source-attributed responses.

- **Agent Orchestration** — two-level architecture: L0 orchestrator (9-node LangGraph, 29-field state) handles classification, planning, and evaluation with persistent working memory (entity tracking, pronoun resolution, topic continuity). L1 domain agents (plan-then-execute) decompose queries into DAGs of typed/SQL/search tasks with concurrent execution. Two-tier human-in-the-loop with crash-safe DynamoDB persistence.
- **Data Pipeline** — Go ETL with medallion architecture (Bronze, Silver, Gold). 5 event-driven Lambdas, 6 source adapters, self-registering mapper registry normalizing 40+ entity types. Content enrichment prepends source context before embedding for provenance-aware retrieval.
- **Retrieval Engine** — Corrective RAG with hybrid search: parallel pgvector cosine similarity and PostgreSQL full-text (tsvector, ts_rank_cd weighting) over 50 candidates each, fused via single-pass Reciprocal Rank Fusion SQL (k=60, FULL OUTER JOIN). Cross-source fan-out with global re-ranking. LLM-based relevance grading with automatic query reformulation when confidence is low. Separate JSONB metadata filter path for structured queries (channel, sender, date range).
- **MCP Tool Layer** — 30 tools across 3 Lambda backends: 24 typed ORM handlers (Go), 2 SQL tools with read-only guardrails (write blocklist, identifier validation, 500-row cap), 4 search tools (Python). JWT-based tenant authentication (Cognito). Bedrock AgentCore runtime with 13 SSE event types for real-time streaming.
- **Infrastructure** — 8 ARM64 Graviton Lambdas, Pulumi IaC (Go), per-tenant isolation (PostgreSQL schema-per-tenant, DynamoDB partitioning, dedicated MCP gateway instances). Provider-agnostic LLM factory supporting Anthropic and OpenAI, configurable per tenant.

Tech: Python, Go, LangGraph, Claude Opus, OpenAI, Aurora PostgreSQL, pgvector, DynamoDB, Bedrock AgentCore, MCP, Lambda, Cognito, Pulumi

Technical Projects

Opportunistic Transparent Huge Pages (M.Tech): Implemented OTHP in Linux kernel, optimizing memory and eliminating periodic scanning. Improved `khugepaged` performance; validated on production workloads.

Savaari – Distributed Cab Hailing: Designed SpringBoot microservices backend for cab dispatch; Docker/Kubernetes deployment.

Mbed-TLS Port (Rust): Ported Mbed-TLS from C++ to Rust, fixing security vulnerabilities and improving memory safety.

Eureka! – Asset Discovery: Built SSH-based tool to catalog network devices; real-time monitoring UI.

Skills

AI & LLM Engineering: LangGraph, LangChain, Agent Orchestration, MCP Servers, RAG Pipelines, Prompt Engineering, Pydantic, Embeddings, Vector Search, pgvector

Infrastructure & Cloud: AWS, Azure, Bedrock, Lambda, ECS, Aurora, DynamoDB, S3, SQS, Docker, Kubernetes, CloudFormation, Terraform, Pulumi, Grafana, Prometheus

Languages & Data: Python, Go, Java, Rust, C++, Bash, SQL, MySQL, PostgreSQL, Redis, StarRocks, Airflow

Architecture & Patterns: Distributed Systems, Event-driven Architecture, ETL Pipelines, Medallion Architecture, Multi-tenant Isolation, Blue-Green Deployments, Concurrency, Message Brokers

Education & Achievements

- Master of Technology, Computer Science & Automation, **Indian Institute of Science, Bangalore**
- **All India Rank 2** out of 100,000+ in Graduate Aptitude Test in Engineering (**GATE** CS, 2020)
- Selected as one of the **Top 25** individuals in company-wide annual performance recognition (Eightfold AI, 2024)
- JLPT N4 & N5 certification, Japanese-Language Proficiency Test