Suman Saurabh

+91-959158-3843 | sumans.saurabh92@gmail.com | linkedin/suman | github/sumansaurabh

Core Competencies

Agentic AI | CLI toolchains | Enterprise AI Infrastructure | Secure LLM Systems | Cross-Team Architecture Leadership | High-Scale ML Pipelines | Analytics and Observability

Summary

Led cross-functional teams to design and scale large-scale distributed systems, supporting 15M+ jobs/month and 200M+ users, with a strong focus on observability, cost optimization, and enterprise-grade compliance.

Skills

AI/ML Infrastructure:	LangChain | Transformers | RAG | Vector DBs | Tensorflow | PyTorch Distributed Quantization | KV Cache | PEFT | LoRA | QLoRA | RLHF
Cloud:	Azure | GCP | AWS | Kubernetes | EKS | GKS | Volcano | Terraform | Nomad
Frameworks:	NestJS | Fastapi | React | Gin | gRPC | Go-kit
Languages:	Python | Typescript | CSharp | Javascript | Java | Rust | Golang
Database:	Postgres | Redis | MongoDB | BigQuery | Cassandra | RDS | DyanmoDB
Messaging Systems:	RabbitMQ | Google Pub/Sub | Kafka | Azure Service Bus
Observability:	DataDog | Sentry | NewRelic | ClickHouse | Prometheus | OpenTelemetry | Kusto
Leadership & Strategy:	Technical Roadmaps | Team Mentorship | Hiring & Onboarding | Sprint Planning | OKRs | Product Thinking | Agile/Scrum | Stakeholder Alignment

Experience: 11 years

Principal Engineer **Sep 2025 - Apr 2026**
BlackBox *Remote*
No-Code AI platform

- Architected Golang-backed WASM sandbox plane, isolating 1M+ daily zero-shot code executions and unblocking Enterprise SOC-2 compliance for the core Copilot product.
- Led architecture for **agentic AI platform** with **6+ engineers**, designing **LangGraph/LangChain-based reAct agent runtimes** with DAG orchestration, tool-calling, and durable execution supporting 10K+ agent runs/day.
- Designed **graph workflow engine** (DAG execution, checkpointing, retry semantics) enabling **long-running, resumable agents** with memory persistence and fault-tolerant execution across distributed environments.
- Led **model router orchestration** (Claude, GPT, Grok) with capability-aware routing and context optimization, ensuring consistent behavior across heterogeneous LLM backends consuming 1B+ tokens per month.
- Institutionalized a high-throughput LLMOps telemetry mesh, ingesting 50M spans/day and managing 2.5TB+ of monthly trace data for deterministic replay; cut org-wide MTTR for complex AI logic anomalies by 60%.

Technologies Used: RAG, ReACT agent, Embeddings, VectorDB, Sandbox, Cross-encoder, Langfuse, Guardrails, LLM, Chain-of-thought, HNSW, bm25, Clickhouse, OpenTelemetry

Senior Software Engineer **Oct 2020 - Aug 2025**
Microsoft OpenAI, Azure Machine Learning *Bangalore, India*
Led platform initiatives across LLM training, AutoML, and secure ML infra contributing over $100M in value.

- Founding team member of AI Fine-tuning on IPP, leading engineering efforts to scale secure LLM training across VNet and Kubernetes, VLLM, processing 20B+ tokens annually and contributing to over $100M+ in revenue.

- Led design of secure multi-tenant ML infrastructure across Kubernetes and Azure, including **GPU scheduling (gang scheduling, bin-packing)**, cost-aware resource allocation, and isolation strategies for LLM workloads.
- Co-architected and led cross-org design reviews and roadmap planning for AutoML Job evolution, balancing SDK usability, compute cost optimization. The solution now supports **15M+ jobs per month**, reduces model development time by up to 90%, and is adopted by over **200K+ global users** via AI Studio and SDK.
- Mentored 8 engineers on secure protocol design; integrated CodeQL and GitHub Advanced Security into CI/CD pipelines; standardized threat modeling to ensure Microsoft compliance and eliminate recurring vulnerabilities.
- Led Scrum execution and 30+ architecture reviews for AI Fine-tuning and AutoML, collaborating with PMs, security, and infra teams to align sprint delivery with scaling, protocol, and compliance requirements.
- Co-developed TunDRA, a secure QUIC-based communication protocol in Rust powering over 1 million Compute Instances with 50% improvement in secure data transfer.

Technologies Used: Scikit-Learn, Transformers, LLMs, Embeddings, GPT, QLoRA, PEFT, Finetuning, DeepSpeed, Rust, Golang, QUIC, VectorDB, PyTorch, Qdrant, RAG, vLLM, Ray Train, BLEU, MMLU, Quantization, MLflow

Team Lead **Jul 2019 - Oct 2020**
ShareChat, Advertisement *Bangalore, India*

- Team lead for ShareChat-Ads team to build an ad-infrastructure from scratch to serve programmatic directs and Real-Time Bidding (RTB) ads resulting in revenue from 20M $ in span of one year.
- Built targeted advertisement platforms that segment over 40 million daily active users on 22 different user attributes to secure better reach of Direct-Deals.
- Introduced CTR Prediction and ad-pacing on direct-deals which improved CTR of ads by 100% .
- Architected Real-Time bidding infrastructure to facilitate auction between OpenRTB-compliant DSPs like AdMob, VMax, Criteo, AppNext, Facebook-Ad-Network, and others.

Technologies Used: Keras, Tensorflow, DeepCTR, MongoDB, Redis, PubSub, Kubernetes, Docker, Prometheus, Grafana, OpenTelemetry.

Senior Software Engineer **Apr 2016 - Jun 2019**
IQLECT, A high velocity real-time Analytic platform *Bangalore, India*

- Founding team member for Ampere, a high-throughput real-time PaaS analytics platform built on BangDB, capable of processing terabytes of streaming data in minutes for low-latency decisioning and interactive insights.
- Designed ML infrastructure for SVM, Regression, KMEANS, Information Extraction, and others. Multiple models can be trained asynchronously and the training can be scaled horizontally.

Software Engineer **May 2015 - Apr 2016**
HomeLane, Holistic furniture business *Bangalore, India*

- Developed a P2P communication platform between using WebRTC api and OpenTok service.

Independent Projects & Freelance

- **ClipboardHealth** - Led end-to-end migration of Clipboard Health's Payments service to microservices using NestJS, Terraform, and AWS, reducing deployment time by 34%, engineering effort by 20% .
- **Gpt-O-Matic** - Created a Hackathon project at Microsoft using OpenAI (GPT-3) and LangChain to generate an assistive tool for creating Azure resources using plain English.
- **Google Summer of Code(GSoc):** Built a Speech-To-Text engine for Apache Stanbol.

Education

B.Tech, Computer Science: LNM Institute of Information Technology, Jaipur(2015)