

Sriharan Manogaran

Lead Software Engineer – Machine Learning · NLP · GenAI · Distributed Systems

+91 95662 41052 | sriharan72@gmail.com | linkedin.com/in/sriharan16 | Chennai, India

PROFESSIONAL SUMMARY

Machine Learning & AI engineer with 7+ years of experience building production-grade ML systems, spanning traditional NLP, statistical ML, and modern GenAI. Deep expertise across the full ML spectrum: from classical anomaly detection, distributional profiling, and embedding-based classifiers to LLM-powered applications, RAG pipelines, and agentic systems. Strong distributed systems foundation (Apache Spark, Kafka, SQS) and hands-on cloud infrastructure experience (AWS, Kubernetes). Co-architect of Freshworks' Freddy AI Insights platform, processing 200M+ ML evaluations daily at 99.99% SLA compliance. Currently driving the GenAI evaluation framework for all Copilot AI features across Freshworks products. Proven track record of leading high-performing ML teams, filing patents, and delivering systems that scale from prototype to millions of users.

PROFESSIONAL EXPERIENCE

Lead Software Engineer – Machine Learning
Freshworks, Chennai, India

Apr 2023 – Present

Freddy AI Insights - Key Co-Architect (AI/ML Systems Lead)

- Co-architected and implemented Freddy AI Insights, a distributed intelligence engine that proactively surfaces anomalies, trends, and root cause analysis (RCA) for enterprise customers, processing ~32M data-generation queries and ~200M ML evaluations daily across tens of thousands of tenants.
- Coordinated delivery across 6 cross-functional teams to ship Insights, aligning on data contracts, schema, and phased rollouts without production impact, with strong observability to track data across systems.
- Led the design and implementation of the AI/ML layer: statistical profiling algorithms (distributional skew, monotonicity detection, longest-contiguous-subsequence analysis, change-point detection) executed as in-memory Spark batch jobs within a strict 1-hour SLA.
- Designed a multi-stage async architecture (SQS + Kafka + Spark Streaming + Spark Batch) that decouples I/O-heavy data fetching from CPU-intensive ML computation, achieving 99.99% SLA compliance and reducing pipeline cost to ~5% of legacy spend while processing 10X the data volume; identified insights are asynchronously pushed via SQS to trigger automated RCA and natural language summary generation.
- Built a configuration-driven ML pipeline (YAML config → dynamic Databricks Spark DAG) enabling new Freshworks products (Freshservice, Freshdesk) to onboard new insight types in days rather than weeks.
- Extended the platform post-launch: Added Customer Experience specific metrics (April 2025), customer-level insight customisation, and CXO-level account-wide insights, scaling from Employee Experience to full Customer Experience customer base.
- Presented an architectural deep-dive at the Freshworks Architects' Forum and co-authored a subsequent feature on the [Freshworks Engineering Blog](#) (March 2026) alongside system and infrastructure architects.

Conversational Analytics - Agentic AI

- Built a natural language to chart/graph system for the Freshworks Analytics Platform, enabling users to query data in plain English and receive dynamically generated visualisations, reports, and export schedules, serving 17K+ requests across 4K+ accounts in production.
- Designed and implemented a fully custom agentic orchestration system, predating frameworks like LangChain, LangGraph, and Flowise, with custom-built tool routing, state management, and multi-turn conversation handling, demonstrating deep first-principles understanding of agentic system design.
- Architected and implemented the LLM orchestration layer handling multi-turn dialogue, intent disambiguation, dynamic metric add/remove, and small-talk using a combination of LLMs and classical NLP techniques; extended to support system actions including report creation, email scheduling, and other platform operations.
- Implemented RAG pipelines to ground the system in product-specific data schema, table attributes, and metric definitions, enabling accurate query resolution against dynamic, customer-specific data models without hardcoding domain knowledge.
- Mentored team members on MLE best practices and drove key architecture decisions across the system.

GenAI Evaluation Framework - Current Initiative

- Architecting and implementing a standardized evaluation pipeline, supporting both online and offline triggers for all Copilot GenAI features across Freshworks products, unifying the AI Agent Platform with the Data Intelligence Platform (Databricks/MLflow) via Arize Phoenix as the observability bridge.

- Designed a Medallion data strategy (Bronze → Silver → Gold Delta tables) normalising both live production traces and offline simulation data, enabling apples-to-apples LLM-as-a-Judge evaluation across environments.
- Eval results are published to two Delta table destinations: A Gold metrics table for Databricks SQL executive dashboards, and MLflow experiment evaluations persisted to Delta tables for long-term trace-level developer debugging, overcoming MLflow's native experiment retention limits.
- Enabled offline experimentation using curated Golden Datasets against sandboxed test agents, proving variant quality before production promotion.

MLOps & Infrastructure

- Managing AWS infrastructure for the Neo Analytics AIML team across VPC, Kubernetes, RDS, SQS, S3, EFS, and Lambda; automated disaster recovery (DR) processes ensuring high availability and operational continuity.
- Optimised infrastructure and pipeline costs by ~24% through a combination of Karpenter-based node autoscaling, Nitro instance migration, VPC endpoints for SQS, instance recalibration, and targeted code optimisations.
- Managing Databricks workspaces and Unity Catalog across 5 regions and staging environments, enforcing access policies and ACLs for data governance and compliance.

Senior Software Engineer - Machine Learning

Oct 2021 – Mar 2023

Freshworks, Chennai, India

- Led a team of 4 engineers to design and deliver a multi-product, multi-tenant custom intent detection system, enabling customers to configure bot, workflow, and internal-application use cases across Freshworks products.
- Architected and implemented the full E2E system from scratch: microservices for CRUD, async model training and embedding generation (FastAPI + Celery workers), real-time inference, and active learning integrated into a batch processing pipeline backed by MongoDB, Redis (data), S3 (model storage), and Elasticsearch (semantic search) and deployment of the same.
- Deployed as the first-layer query handler across multiple Freshworks products, the system handles 1M+ requests per day from 10K+ active bots in production.
- Used LaBSE multilingual sentence embeddings with a classification layer for intent detection, enabling cross-language support without per-language model variants and significantly reducing training overhead.
- Coordinated with 8 cross-functional teams to build and deliver the E2E system, and collaborated with 3 additional product teams for feature integration and rollout, ensuring alignment across requirements, data contracts, and release timelines.
- Built Freshworks' first version of MLOps platform on Databricks covering experiment tracking, feature store, and model registry; integrated KServe for production model serving with a continuous improvement feedback loop.
- Co-managed AWS infrastructure for the AIML Platform team using Kubernetes; automated DR processes.

Software Engineer - Machine Learning

Apr 2020 – Sep 2021

Freshworks, Chennai, India

- Extended the core bot platform to Freshchat and Freshservice, adapting the data sync pipeline to each product's requirements while reusing downstream microservices, scaling the system to 60K+ customers and 200K+ bots across products.
- Fine-tuned a BERT model for an IT Support Bot Service Item Suggester on Freshservice, reducing redundant support queries for 30K+ customers and 100K+ bots; deployed via TensorFlow Serving for optimised inference.
- Developed a V2 Solution Article Suggester using MUSE (Multilingual Universal Sentence Encoder) embeddings with a similarity and preprocessing layer, improving accuracy and cross-language relevance.
- Ensured GDPR and HIPAA compliance for bot and model account lifecycles by automating data cleanup via cron jobs.

Graduate Trainee Software Engineering

Jun 2019 – Mar 2020

Freshworks, Chennai, India

- Built the core multi-tenant model training and data sync pipeline for Freshdesk bots, using the product's knowledge base to automate L1 query handling, serving 42K+ customers and 90K+ bots. Stack: Java, Spring Boot, Kafka, MySQL, Redis.

PATENT

Customer Service and Employee Experience Entity Management Using AI-Generated Insights for Efficient Observability and Planning

USPTO Application No. 19/400,946 · Filed: November 25, 2025 · Status: Pending

PUBLICATIONS

Freddy AI Insights: Tech Behind the Scenes

March 2026

Freshworks Engineering Blog | [Read Article](#)

Architecture deep-dive covering async + Spark distribution for mining 182M+ ML evaluations daily under a 1-hour SLA. Co-authored with system and infrastructure architects.

TECHNICAL SKILLS

Programming Languages: Python, Java, SQL

ML & AI: NLP, Intent Detection, Multilingual Embeddings, Anomaly Detection, Statistical Profiling

LLMs & GenAI: RAG, Agentic Systems, LLM orchestration, LLM-as-a-Judge, prompt engineering, fine-tuning

MLOps: Databricks, MLflow, KServe, Arize Phoenix

Frameworks: FastAPI, Spring Boot, Spark (Streaming & Batch), Celery

System Design: Distributed systems, microservices, event-driven architecture, scalability

Databases: MySQL, MongoDB, Redis, Elasticsearch, PostgreSQL, OLAP (Snowflake, E6)

Infrastructure: AWS (VPC, K8s, SQS, Kafka, Lambda, RDS, S3, EFS), CI/CD (Jenkins, AWS CodePipeline, GitHub Workflows)

RECOGNITION & AWARDS

- Freshworks Gold Award - highest-tier quarterly engineering recognition, awarded for exceptional impact on the Insights & AI platform.
- Multiple Freshworks Silver Awards and Spot Awards across tenure.
- Invited presenter at Freshworks Architects Forum, presenting Insights distributed architecture to all senior architects up to CTO level.

EDUCATION

Bachelor of Engineering – Computer Science

Sathyabama Institute of Science & Technology, Chennai

2015 – 2019