SUJAL CHARAK

San Jose, CA | +1 857-919-6272 | sujalcharak20@gmail.com | LinkedIn | GitHub

SUMMARY

Applied AI engineer (MS Computer Science, Boston University) building production generative AI and ML systems: LLM evaluation and guardrail pipelines, RAG systems, conversational and agentic assistants, and on-device reinforcement learning. Two years of large-scale cloud data engineering in a regulated financial environment (Azure, tens of TB under strict SLA). Sole inventor on a patent-pending offline RL system; published author. Owns problems end to end, from problem framing through evaluation, safety, monitoring, and deployment on resource-constrained hardware.

EDUCATION

Boston University — *Master of Science, Computer Science (STEM Designated)* Sep 2024 – Jan 2026

Mumbai University — *Bachelor of Engineering, Electronics & Telecommunications* Jul 2017 – Jul 2021

EXPERIENCE

Data Engineer, Tata Consultancy Services Oct 2021 – Oct 2023

Client: ABN AMRO (regulated financial data / financial-crimes program), Mumbai, India

- Owned and operated production cloud data platforms (Azure Data Factory, Databricks), building PySpark and Delta Lake ETL pipelines processing tens of terabytes under strict SLA, reliability, and data-quality requirements in a regulated environment.
- Reduced data processing time by up to 80% and cut cloud compute cost ~20% through partitioning, caching, query tuning, and resource-aware pipeline design.
- Automated duplicate detection and cleanup across ~45 TB; built production pipelines with monitoring, logging, and failure-recovery for high availability, plus CI/CD and schema-change backfills.
- Worked Agile sprints with rotating on-call production support; performed validation and stakeholder issue resolution across raw and curated data layers.

Teaching Assistant, Boston University Jan 2025 – May 2025

Information Structures with Python, Boston, MA

- Graded labs and reviewed Python code quality for graduate students; ran weekly support sessions on debugging, data structures, and algorithmic fundamentals.

PROJECTS

PRI — Persistent Reinforcement Intelligence Apr 2025 – Present

Sole Inventor (Patent Pending, India 2025)

- Designed and built a fully offline reinforcement learning engine in C on Linux (~7,400 LoC, 22 modules) that ingests live procfs telemetry to autonomously optimize OS resource usage on aging hardware.
- Implemented tabular Q-learning over a discretized state space with persistent cross-reboot policy storage using an atomic A/B slot mechanism for crash-safe writes.
- Built a three-baseline evaluation harness (no-op, tuned heuristic, learned policy) with convergence diagnostics, plus a safety-aware action layer with reversible interventions and automatic rollback. Provisional patent filed in India (Jul 2025).

AIXT — LLM Security & Red-Teaming Toolkit Sep 2025 – Oct 2025

Security Researcher (Open Source) — github.com/SujalCharak/gemini-prompt-exfil

- Developed an LLM evaluation and guardrail pipeline (Gemini) for prompt-exfiltration and RAG context-leakage testing, using structured multi-turn attacks and automated scoring to surface alignment, robustness, and context-isolation failures.

Privacy Radar — Privacy-Aware Network Monitoring with AI Insights Sep 2025 – Dec 2025

Team Project — github.com/madhurdeepjain/PrivacyRadar/pull/40

- Built the Privacy AI subsystem for an Electron desktop network monitor: routed aggregated session telemetry (top apps, byte volumes, geolocation) through a single Gemini egress point returning a constrained four-section report (summary, key insights, risk, recommended actions) over a secure preload IPC boundary, with a deterministic local fallback when the model is unavailable.

Portfolio Voice Agent Mar 2025 – May 2025

Independent Project (Open Source) — github.com/SujalCharak/portfolio-voice-assistant

- Built a voice-first quantitative assistant on a two-process design: an LLM reasoning core (OpenAI query planning, dynamic response generation) and a continuous voice I/O loop (wake-word detection, speech recognition, text-to-speech).
- Combined 1000-scenario GBM Monte Carlo forecasts and NewsAPI sentiment into a spoken Buy/Hold/Sell recommendation with natural-language rationale, plus optional ffmpeg-rendered simulation videos.

Flight Price Optimization (Shortest-Path Algorithms) Sep 2024 – Dec 2024

Academic Project, Boston University (Team of 2)

- Co-developed a Python implementation of Dijkstra and Bellman-Ford over a 12-airport flight-cost graph (NetworkX) to find cost-optimal routes; contributed runtime benchmarking and complexity analysis.

SKILLS & CERTIFICATIONS

AI & ML: Generative AI, LLM Evaluation & Guardrails, RAG Systems, Prompt Engineering, Conversational & Agentic AI, AI Safety & Red-Teaming, Model Behavior & Robustness, Reinforcement Learning (Q-learning, policy persistence), Supervised Learning (Regression, Decision Trees, Random Forests, Gradient Boosting, XGBoost), Feature Engineering, Model Evaluation (Precision, Recall, F1, ROC AUC), TensorFlow, LangChain.js, Gemini (Google Gen AI SDK)
Cloud & Data Engineering: Azure Data Factory, Azure Data Lake, Databricks, Delta Lake, PySpark, ETL Pipelines, Data Modeling, Spark Optimization, Schema Evolution & Backfills, Pipeline Monitoring & Recovery, CI/CD, On-call Production Support
Programming & Tools: Python, SQL, C, Bash, Linux, Git, Node.js, TypeScript, React, Electron, IntelliJ, Agile/Scrum, DevOps
Certifications & Achievements: Python Specialization (Univ. of Michigan); AI for Everyone (DeepLearning.AI); IJCER Journal publication