Henish Patel

harrypatel1214@gmail.com | 385-293-9015 | linkedin.com/in/henishpatel2004 | github.com/HenishPatel1214

Education

Bachelor of Science in Computer Science + Data Science **January 2024 – May 2027**
University of Utah, Salt Lake City, UT *GPA: 3.57*
Coursework: Operating Systems, Data Structures & Algorithms, Databases, Machine Learning

Skills

Languages: Python, Java, C++, C#, C, JavaScript, SQL, Bash, Assembly

Skills: Machine Learning (Regression, Clustering), Optimization, RAG, CI/CD, Data Analysis, Data Visualization

Technologies: Git, React, Docker, FastAPI, Ollama, ChromaDB, TensorFlow, PyTorch, Scikit-learn, Pandas, NumPy, Agentic AI Coding (Claude Code), AWS (EC2, S3), PostgreSQL, SQLite, Seaborn, Jupyter Notebook, PowerBI, Tableau

Certifications: Google IT Automation, Google AI Agents Course, Oracle Database SQL, IBM Data Science Professional

Experience

AI Systems Intern **May 2026 – Present**
Management & Training Corporation (MTC), Salt Lake City, UT

- Architecting a **self-hosted AI inference stack** with **Ollama** and **Docker**, deploying open-source LLMs for secure on-premise reasoning with zero external API exposure
- Designed evaluation harnesses benchmarking local model **latency, accuracy, and resource utilization** to accelerate retrieval and prompt iteration
- Built a **retrieval-augmented generation (RAG)** pipeline with **vector embeddings** to ground the self-hosted models in internal documents, improving answer relevance with full on-premise data privacy

Full Stack Software Engineer **Oct 2025 – Present**
mySpecSheet, San Francisco, CA (Remote)

- Architected an **AI-integrated dashboard** utilizing **VSCode OSS** and **TypeScript**; implemented custom **LSP extensions** to reduce developer context switching by **38%**
- Engineered a **"vehicle-as-a-repo"** architecture utilizing **Merkle-tree versioning** and **MCP Servers** for secure, immutable repair history and sandboxed data manipulation
- Developed a **WebSocket orchestration layer** for real-time telemetry synchronization with sub-**50ms** latency; implemented **OAuth2/RBAC** ready for **1,000+** concurrent service nodes

SUDO Software Platform Services Intern **July 2024 – April 2026**
University of Utah Faculty Information & Support, Salt Lake City, UT

- **Optimized GIS geospatial indexing** using **React, Node.js, and R-trees**, improving spatial data query performance by **27%** and reducing client-side rendering latency by **140ms**
- **Architected ServiceNow ETL pipelines** to automate performance auditing, reducing processing time from **4 hours to 15 minutes** and ensuring **100% data consistency** across relational schemas

Data Analyst Intern **May 2024 – August 2024**
University of Utah Health Facilities Management, Salt Lake City, UT

- **Led zero-downtime migration** of legacy reporting workflows to **AWS (S3, EC2)** with **MD5 checksum validation** across **500GB+** datasets, ensuring total data integrity
- **Engineered modular ETL pipelines** in **Python (Pandas)** automating reconciliation of **15,000+ monthly records**

Selected Projects

NVIDIA Grace-Blackwell DGX Spark Cluster
Python, CUDA, RAPIDS, RoCE v2, Docker

- Built a **DGX Spark Cluster** utilizing **ConnectX-7 200Gb/s** interfaces and **RoCE v2 (RDMA)** for peer-to-peer GPU memory access, creating a unified **256GB+** VRAM pool
- Utilized **CUDA accelerated libraries** to offload workloads to GPU, achieved a **50x** speedup in a small deployment

Stripe Agentic Billing Extension
Node.js, Redis, PostgreSQL, Stripe API

- Used the **Stripe V2 EventStream API** to enable real-time, token-based monetization for autonomous agents
- Implemented a **Redis ZSET-backed** event aggregator with **Lua scripting** to atomize credit burndown, reducing external API overhead by **85%** while maintaining **sub-ms** local latency

Leadership

JROTC Cadet Major, 4-Year Leadership Program **August 2019 – May 2023**

- Promoted to Cadet Major for leadership and dedication; led team exercises, improved unit operations, upheld standards