# Rohan Surana

San Diego, CA | 510-265-4239 | rohansurana2810@gmail.com | linkedin.com/in/rohansurana28/ | github.com/rohan2810

## Education

**Masters of Science in Data Science** — *March 2026*
University of California, San Diego (UCSD) GPA: 3.88/4.00. Advisor: Julian McAuley — San Diego, CA

**Bachelor of Science in Software Engineering** — *May 2022*
San Jose State University, GPA: 3.87/4.00 **(Summa cum laude)** — San Jose, CA

## Experience

**Dell Technologies** — Hopkinton, MA
AI Research Intern — June 2025 – Sept 2025

- Built multi-agent LLM system (LangGraph/LangChain + vLLM) with request batching and KV-cache, cutting p95 latency by 40%
- Designed scalable RAG pipelines integrating LLMs with vector databases using hybrid search and intelligent chunking strategies
- Engineered short/long-term memory with retrieval caching to reduce serving cost & exposed APIs for integration with existing workflows
- Developed agent-based monitoring using MCP and A2A protocols to automate telemetry collection, anomaly detection, and remediation

**Dell Technologies** — Santa Clara, CA
*Software Engineer II* — Mar 2024 – Aug 2024

- Architected TOSCA-based framework to orchestrate resources, cutting provisioning time 30% and improving infrastructure flexibility
- Built real-time infrastructure digital twin with predictive analytics using graph databases to enable drift detection and safe remediation
- Optimized intent workflows with graph-based scheduling to improve response time by 25% under peak load and increase throughput
- Developed drift-detection & reconciliation engine and partnered with the telemetry team to enhance analytics and predictive maintenance

*Software Engineer I* — Jul 2022 – Mar 2024

- Accelerated cluster time-to-ready by 20% through automated Kubernetes operators supporting GPU workloads, KServe, and TorchServe
- Enhanced service performance by redesigning API and proto services with gRPC, leading a 4-intern team to boost efficiency 15%
- Expanded system-wide tracing and observability coverage 30% by deploying OpenTelemetry with Prometheus, Jaeger, and Grafana
- Implemented multi-tenant gRPC middleware in Golang with policy-based authn/authz, enabling granular access control policies

**Confluxsys LLC** — May 2020 – Aug 2020
Software Developer Intern — Folsom, CA

- Built modules using Spark, Scala, & GNNs improving pipeline throughput 25% unlocking analytics for healthcare & finance clients
- Created data-mining utilities using DataStax libraries; reduced job runtimes 45% and standardized ETL for downstream ML features

## Selected Publications

- **In-context Ranking Preference Optimization**
  J. Wu*, **Rohan Surana***, Z. Xie, Y. Shen, Y. Xia, T. Yu, R. Rossi, P. Ammanabrolu, J. McAuley — *COLM 2025*
- **Traceable and Explainable Multimodal Large Language Models: An Information-Theoretic View**
  Z. Huang, J. Wu, **Rohan Surana**, R. Jain, T. Yu, R. Addanki, D. Arbour, S. Kim, J. McAuley — *COLM 2025*
- **Image Difference Captioning via Adversarial Preference Optimization**
  Z. Huang, J. Wu, **Rohan Surana**, T. Yu, D. Arbour, R. Sinha, J. McAuley — *EMNLP 2025*
- **MASS-DPO: Multi-negative Active Sample Selection for Direct Policy Optimization**
  **Rohan Surana***, J. Wu*, X. Li, Y. Shen, C. Wang, T. Yu, P. Ammanabrolu, J. Shang, J. McAuley — *Under Review - ICLR 2026*
- **Active Data Distillation for Zero-shot LLM CRS**
  **Rohan Surana***, J. Wu*, Z. Xie*, Y. Xia, H. Steck, D. Liang, N. Kallus, J. McAuley — *Under Review - WSDM 2026*
- **MusiCRS: Benchmarking Music-Centric Conversational Recommendation**
  **Rohan Surana***, A. Namburi*, G. Mundada*, A. Lal*, Z. Novack, J. McAuley, J. Wu — *Under Review - ICASSP 2026*

## Projects

**Privacy-Preserving LLM Training with Synthetic Data** | *CrewAI, Unsloth, PyTorch, Llama-2* — Sept 2024 – Dec 2024

- Designed a multi-agent data-generation framework with CrewAI for persona generation and PII masking to produce PII-safe corpora
- Fine-tuned 2 OSS LLMs via PEFT (Unsloth) to 99% PII removal and built 270-Q eval benchmark for model assessment

**Autonomous Transportation Computer Vision System** | *YOLOv5, TensorFlow, OpenCV* — Jan 2022 – May 2022

- Built end-to-end real-time vehicle detection & tracking (YOLOv5+BiLSTM+OpenCV) with predictive analytics; achieved 0.45 RMSE
- Optimized YOLOv5 and TensorFlow for Raspberry Pi to meet on-device memory limits and boost throughput and system efficiency

## Technical Skills

- **Languages:** Python, Golang, Java, Scala, SQL
- **ML/DL:** PyTorch, TensorFlow, scikit-learn, Transformers, OpenCV, NumPy, Pandas
- **LLM/NLP:** LangGraph, LangChain, vLLM, Unsloth, CrewAI, Hugging Face
- **Infrastructure/Frameworks:** Kubernetes, Docker, FastAPI, Apache Spark, gRPC, Git, MLflow, Ray, AWS, GCP, CI/CD, Postman
- **Databases:** ChromaDB, MySQL, PostgreSQL, DGraph, Neo4j, MongoDB