

Karan Narula

San Diego, CA | +1-858-319-5650 | kanarula@ucsd.edu | linkedin.com/in/knarula03

SUMMARY

Systems and GenAI Engineer with a proven track record of architecting, securing, and deploying enterprise-grade AI applications and pipeline infrastructure. Experienced in embedding with clients as an AI Consultant and technical lead to bridge frontier AI models with complex corporate networks. Expert in agentic orchestration (LangGraph, MCP), production LLM evaluation at scale (1M+ traces/month), and solving enterprise authentication and cloud security friction to accelerate product adoption.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, Java, C/C++, C#, SQL, Bash

AI & LLM Ecosystem: Vertex AI, Google Gemini, OpenAI, Anthropic Claude, Azure OpenAI, Model Context Protocol (MCP), LangGraph, LangChain, RAG, GraphRAG, Prompt Engineering, Fine-tuning, PyTorch, TensorFlow, LLM Evaluation & Observability

Cloud & Infra: GCP, Azure, AWS, Docker, Kubernetes, FastAPI, GitHub Actions, CI/CD, MongoDB, Snowflake

Enterprise Integration & Security: SAML SSO, OAuth, NextAuth, MFA, Cloud Security Compliance, Identity-Provider Actions

PROFESSIONAL EXPERIENCE

Audible Sight, Inc. | Remote

Lead Systems Engineer (CTO) | Feb 2026 - Present

- Architected and shipped a production multi-stage AI audio description pipeline (frame selection, multimodal prompt orchestration, post-processing, and TTS integration) from prototype to enterprise-grade, serving paying B2B customers.
- Engineered the platform's enterprise integration layer—implementing SAML SSO, conditional MFA, account linking, and consumer domain blocking via identity-provider actions—successfully unblocking procurement with public-sector and enterprise prospects.
- Owned full cloud security compliance posture; completed formal security assessments for industry registries, authored technical policy documentation, and diagnosed/resolved critical production authentication failures caused by container build configurations.
- Designed a subagent-based development architecture across company repositories using agentic frameworks, accelerating team delivery of interactive studio interfaces, voice-selection tools, and captioning export features (SRT/VTX).
- **ML Systems Intern (Feb 2025 - Aug 2025):** Deployed a scalable vision-language evaluation pipeline benchmarking open-source and commercial VLMs (including Gemini and Claude), improving caption accuracy by 25% and driving a 20% increase in subscription revenue.

Netomi | Remote

Data Science Intern | Sep 2025 - Dec 2025

- Built an end-to-end LLM evaluation and observability framework processing 1M+ traces monthly; tracked system accuracy, latency, cost-per-request, and safety signals (hallucination scoring, prompt injection detection, RAG metrics) across model versions.
- Designed automated LLM-as-a-judge pipelines alongside human-annotation loops, surfacing actionable prompt optimization insights that improved model response accuracy by 18% and reduced latency variance by 25%.

Flectere | Remote

AI Consultant & Data Science Intern | Jun 2023 - Sep 2025

- Embedded with an enterprise property-management client (100+ users) to lead technical discovery, mapping their legacy workflows to architect an AI-powered Outlook plugin using FastAPI; automated email classification and cut processing time by 45%.
- Built a robust validation layer achieving 85%+ intent accuracy for the client's production deployment, and engineered cloud CI/CD pipelines with precise token tracking that drove a 20% operational cost reduction.
- Developed an NLP-driven expert network graph using GraphRAG and a Streamlit interface, boosting client operational efficiency by 25%.
- Led an ensemble machine learning project achieving 75%+ accuracy on sports outcome predictions, supported by end-to-end CI/CD infrastructure.

EDUCATION

University of California, San Diego | B.S. Mathematics-Computer Science | GPA 3.9/4.0, Provost's Honors | Sep 2021 - Dec 2025

Relevant Coursework: Machine Learning, Recommender Systems, Web Mining, Data Science, Design & Analysis of Algorithms, Theory of Computation, Software Engineering, Systems Programming.