Sanmukh Sain Karri

San Francisco, CA · +1 332-254-6972 · shanmukhsain@gmail.com · linkedin.com/in/shanmukhsain · github.com/samshanmukh

Authorized to work in the U.S. on H-1B (transfer-eligible, not cap/lottery-subject).

SUMMARY

AI/ML Engineer with 6+ years shipping production software and applied ML/GenAI systems in Python. Hands-on builder who owns features end-to-end — problem framing, prototyping, evaluation, rollout, and production monitoring — and partners closely with product, data, and platform teams. Deep experience with LLMs and RAG, defining evaluation criteria and safety guardrails, A/B testing and experimentation, and operating AI features at scale for reliability, latency, and cost. Comfortable across the stack and translating model outputs into clear narratives for non-technical stakeholders.

CORE SKILLS

- **GenAI / LLM:** LLMs, Retrieval-Augmented Generation (RAG), prompt engineering, fine-tuning, embeddings & vector search, Vectara, LLM APIs, agentic workflows
- **AI Quality & Trust:** evaluation criteria & offline/online evals, A/B testing, guardrails & safe-use patterns, hallucination reduction, human-in-the-loop, monitoring, PII / sensitive-data handling
- **Experimentation:** experimental design, hypothesis testing, causal inference (diff-in-diff, propensity matching), statsmodels, SciPy.stats
- **Machine Learning:** TensorFlow, Scikit-Learn, deep learning, classification/regression/clustering, model development & deployment, Pandas, NumPy
- **Languages:** Python (advanced), SQL, JavaScript/TypeScript, PHP
- **Backend / APIs / Cloud:** FastAPI, Flask, Django, Node.js/Express, REST, GraphQL; AWS (EC2, S3, RDS, CloudFormation, Auto Scaling), PostgreSQL, MongoDB, CI/CD, Git
- **Visualization / BI:** Power BI, Matplotlib, Seaborn, Plotly

FEATURED PROJECTS

Truth Whisperer (MoonrakerFactCheck) — Real-Time Audio Fact-Checking Hackathon

- Lead developer for the LLM/RAG implementation: built an agentic, multi-agent Retrieval-Augmented Generation pipeline that fact-checks transcribed audio against verified sources in real time and returns source citations.
- Integrated speech-to-text with the RAG fact-checker and tuned retrieval to balance latency and precision across diverse accents and dialects.

VISA-VANTAGE — LLM/RAG Assistant Feb 2024

- Built an LLM-powered RAG application on Vectara answering U.S. immigration questions from real-time USCIS data, owning the feature from prototype through evaluation.
- Defined evaluation criteria and grounded responses over retrieved sources to reduce hallucination and keep answers current — an applied guardrail/quality pattern.

EXPERIENCE

Clinicom Healthcare Inc — AI/ML Engineer / Software Developer Feb 2024 – May 2026

Dothan, AL (Remote)

- Owned data/ML features end-to-end on an enterprise healthcare platform — from problem framing and prototyping through deployment and production monitoring.
- Designed evaluation criteria and A/B tested adaptive assessment workflows; applied causal inference to measure the effect of changes on patient-reported outcomes against business goals.
- Built and operated production data/ML pipelines in Python (acquisition, cleaning, feature engineering, statistical analysis) handling sensitive patient data with PII-aware, access-controlled practices.

- Partnered cross-functionally on architecture and tooling decisions, set engineering best-practices, and translated model outputs into Power BI narratives for clinical and operations stakeholders.
- Built and deployed an LLM integration on IBM Granite (8B) to automate clinical report writing in production, sharply reducing per-patient report drafting time for clinicians and delivering major workflow-efficiency gains.

QQ Tech Inc — Data Scientist Intern Jun 2023 – Aug 2023
Tracy, CA

- Built an NLP application to parse legal documents, extract key clauses, and auto-generate summaries.
- Defined evaluation criteria and ran controlled experiments comparing model variants with hypothesis testing before adoption.
- Built data analysis and visualization (Python, ReactJS) to evaluate and iterate on model outputs.

Focus-N-Fly Inc — Senior Software Engineer Jan 2022 – May 2023
Burlingame, CA

- Owned end-to-end delivery of the Runcoach/Movecoach platforms, including Android release cycles, partnering with product and operating in production at scale.
- Designed and ran A/B tests on features and a patented adaptive-training algorithm; used causal analysis to measure impact on engagement and retention.
- Built and extended REST APIs (LEMP/Laravel) integrating Fitbit, Garmin, and Strava, and built Power BI analytics to drive product decisions.

Navtech — Software Engineer Nov 2019 – Jan 2022
Hyderabad, India

- Full-stack development (Node.js, Express, Angular, Flutter) and internal line-of-business tooling for international clients.
- Built an AWS DevOps automation app (CloudFormation, EC2, S3, VPC, RDS, Auto Scaling) that reduced deployment effort and cost — designing for reliability and operating cost.

Prahem Technologies — Web Developer Jan 2019 – Nov 2019
Hyderabad, India

- Built RESTful APIs and asynchronous backends (Node.js, Express, MongoDB) for the Pickkup delivery suite (user, agent, and admin apps).

EDUCATION

M.S., Data Science — University of the Pacific May 2024
San Francisco, CA · Coursework: Machine Learning, Advanced ML, Data Engineering, Data Analytics, Software Methods

B.Tech., Computer Science Engineering — JNTU 2015 – 2019
Hyderabad, India

CERTIFICATIONS

- Introduction to Generative AI — Google Cloud (2023)
- Attention Mechanism — Google Cloud (2023)
- Python Essentials for MLOps — Coursera (2023)
- Machine Learning with Python — Coursera (2023)