Teja Sunku

Redmond, WA (US Citizen, relocation ok) | +1 (206) 445-3399 | sunkut@outlook.com | /in/tejasunku

SUMMARY

Experienced AI Infrastructure Engineer with strong proficiency in Python and AWS, led the migration of a core LLM routing service to async architecture, enhancing performance. Managed the transition to Kubernetes for improved service orchestration. Combines data science expertise with production infrastructure experience to build scalable AI systems.

EXPERIENCE

Tumble (trytumble.app) — Oct 2025 - Present

Co-Founder & Tech Lead

- Built consumer planner app from scratch as technical cofounder, owning full-stack development with TypeScript, Next.js, and React patterns
- Led all architectural decisions including app structure, state management, and deployment pipeline
- Conducted user testing with early adopters to inform product direction, progressing from concept to MVP in 4 months

Ashvin AI — Jan 2024 - Jun 2025

Machine Learning, Operations (ML Ops)

- Led migration of the core LLM routing service to an async architecture, increasing throughput by 10x without additional compute. Optimized performance by resolving blocking AWS SDK bottlenecks, delivering an additional 3x throughput gain
- Designed multi-step document processing workflows using multimodal LLMs, converting PDF medical records to images and applying structured extraction/classification rubrics via prompt engineering
- Evaluated 100+ real customer documents to refine LLM prompts, systematically removing extraction errors and improving output quality; assessed cost/performance tradeoffs across models to optimize for accuracy and latency
- Spearheaded Kubernetes adoption across backend services by converting legacy Python scripts into containerized FastAPI microservices, completing migrations in about 1 week per service. Coordinated cross-functional onboarding to enable same-day deployments and production rollouts, speeding up development
- Redesigned CI/CD pipelines and implemented multi-repo integration testing, cutting build-to-deploy latency from ~10 minutes to under 3 minutes
- Built resilient data ingestion and export pipelines for legacy systems, normalizing untyped data into structured formats and enabling same-week client integrations

Cyentia Institute — Apr 2022 - Jan 2023

Data Scientist

- Analyzed datasets of 1 million+ rows to derive actionable insight and generate comprehensive reports, collaborating with Marketing to finetune visualizations
- Trained and implemented NLP machine learning models using SpaCy and BERT models to augment CVE-related data feeds with key words and phrases, improving data accuracy and relevance
- Extracted vulnerability descriptions from websites and APIs using R and Python, exporting structured tabular formats to AWS, which enhanced data accessibility and integration for security analysis
- Enabled reproducible testing environments by creating Docker scripts with detailed documentation, and by constructing AWS test environments and configs for validating pre-production code

EDUCATION

UC San Diego — 2023 - 2024

Machine Learning Bootcamp

University of Idaho — 2017 - 2021

B.S., Statistics & Philosophy, Minor in Computer Science (Graduating Senior of the Year, College of Science)

SKILLS

Languages: Python, SQL (Postgres), Bash, TypeScript

Cloud & Infrastructure: AWS, Kubernetes, Terraform, GitHub Actions, Datadog

ML & AI: PyTorch, TensorFlow, Keras, scikit-learn, spaCy, Hugging Face, MLflow, OpenAI API

Data Science: pandas, NumPy, Jupyter, Tableau, PowerBI, Databricks

Web & APIs: FastAPI, Next.js, GraphQL, REST APIs

Practices: CI/CD, Async Programming, Pytest, Documentation & Testing