UDAY SHANKAR GATTU

(617) 971-7892 | udaygattu9949@gmail.com | linkedin.com/in/udayshankargattu/ | github.com/UdayGattu | https://udaygattu.github.io/ | HuggingFace

PROFESSIONAL SUMMARY

Founding AI Engineer with production experience in building agentic LLM pipelines and scalable FastAPI backends. Spearheaded the 0-to-1 development of adversarial security platforms, optimizing inference latency for Generative AI systems. Expert in deploying secure cloud infrastructure on Azure/GCP and translating complex R&D initiatives into production-grade deployments.

SKILLS

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

AI/ML: PyTorch, TensorFlow, LangChain, Transformers, RAG, OpenAI API, Llama 2, Stable Diffusion, Vector Databases (Pinecone)

Backend: FastAPI, Django, Flask, WebSocket, RESTful APIs, Microservices, Redis, Celery, GraphQL

Cloud & DevOps: AWS (EC2, Lambda), Azure (Data Lake), GCP (Vertex AI), Docker, Kubernetes, Terraform, CI/CD

WORK EXPERIENCE

Galexor | Boston, MA **June 2025 – Present**

Founding AI Engineer

- Spearheaded the end-to-end engineering of the Galexor MVP, delivering a fully functional adversarial prompt processing pipeline in a condensed 3-month timeline to secure critical pre-seed VC funding and launch initial pilot programs
- Architected a low-latency backend using FastAPI and WebSockets, implementing fan-out aggregation that reduced query response times by 40% while handling concurrent adversarial attacks during live enterprise security demonstrations
- Engineered adversarial testing frameworks to stress-test Large Language Models against prompt injection, finetuning open-source models on synthetic vulnerability datasets to establish the company's core safety compliance standards
- Designed the complete MLOps strategy for agentic workflows by integrating LangSmith for observability and Redis for state management, allowing the R&D team to rapidly iterate on prototypes and deploy production-ready agents with 99.9% uptime

Tata Consultancy Services | Hyderabad, INDIA **June 2022 – August 2023**

Machine Learning Engineer - Cloud Exponence Microsoft Azure

- Deployed predictive maintenance pipelines on Azure Machine Learning to automate cloud governance, processing 5TB of log data daily to identify resource bottlenecks and reduce operational infrastructure costs by 15% through preemptive resource scaling
- Built real-time monitoring dashboards using Flask and JavaScript to visualize distributed system health, enabling the operations team to detect anomalies faster and cutting the average mean time to resolution (MTTR) for critical server incidents by 30%
- Optimized data ingestion workflows by implementing distributed storage solutions with Azure Data Lake and SQL, refactoring legacy query logic to improve data retrieval speeds for high-volume enterprise workloads without compromising data integrity

Python Developer Intern - Cloud Exponence Microsoft Azure **June 2021 – June 2022**

- Developed RESTful APIs for cloud automation using Python and Flask to replace manual provisioning scripts, improving the consistency of resource deployment and significantly reducing the setup time required for new development environments
- Implemented infrastructure-as-code automation using Terraform and Azure Functions, creating reusable modules that standardized cloud resource allocation and eliminated configuration drift across multiple staging and production environments

KEY PROJECTS

BrandAI – AI-Powered Advertisement Critique Engine **Link**

Tech Stack: FastAPI, LangGraph, Vertex AI (Imagen 2, Veo 3.1), Gemini Vision, CLIP, OpenCV, Docker, Redis

- Built a comprehensive ad-generation engine orchestrating Gemini Vision and CLIP to autonomously critique and generate brand-aligned marketing assets, reducing the manual review cycle for creative teams by 50% while ensuring strict brand compliance
- Implemented a fault-tolerant agent workflow using LangGraph with parallel execution and retry logic, backed by Redis for state persistence, which streamlined the generation of multi-modal ad variations and cut end-to-end processing latency by 40%

Image Alchemist: AI-Driven eCommerce Image Enhancement **Link**

Tech Stack: FastAPI, Streamlit, YOLOv8, OpenCV, Stable Diffusion, GANs, Pillow, NumPy

- Developed a production-grade image enhancement pipeline utilizing Stable Diffusion and OpenCV to automatically correct lighting, remove backgrounds, and standardize product layouts, ensuring catalog consistency for high-volume eCommerce
- Engineered a real-time editing interface allowing users to perform semantic image manipulation via text prompts, integrating a custom GAN architecture that reduced the computational load for image upscaling tasks by roughly 35%

EDUCATION

Northeastern University, Boston, MA **May 2025**

Master of Science in Software Engineering Systems *GPA: 3.7*

- **Courses**: Advanced Techniques with LLMs, Generative AI, NLP, Cloud Computing, Responsible AI, Algorithms
- **Graduate Teaching Assistant**: Generative AI, Natural Language Processing, Prompt Engineering