UDAY SHANKAR GATTU

(617) 971-7892 | udaygattu007@gmail.com | linkedin.com/in/udayshankargattu/ | github.com/UdayGattu | HuggingFace | Portfolio

SKILLS

Programming Skills: Python, Java, C++, JavaScript, SQL
AI/ML & Generative AI: PyTorch, TensorFlow, Transformers, RAG, LLMs (GPT-4, Llama 2), Stable Diffusion, OpenAI
AI Agent Frameworks: LangChain, Semantic Kernel, Prompt Chaining, RAG, RBAC, ABAC
Cloud & DevOps: AWS (EC2, S3, SageMaker), GCP (Vertex AI), Azure (Data Lake), Kubernetes, Docker, Terraform, CI/CD
Backend & APIs: FastAPI, Flask, Django, RESTful APIs, Microservices, Serverless (Lambda, Azure Functions)
Tools & Libraries: Hugging Face, OpenCV, Scikit-learn, Pandas, Git, Postman, React.js, VoiceFlow

WORK EXPERIENCE

Tata Consultancy Services **June 2022 – August 2023**

Machine Learning Engineer - Cloud Exponence Microsoft Azure

- Built and deployed predictive AI models on Azure using Python, automating cloud governance processes and reducing operational costs by 15%, while ensuring consistent 99% service availability
- Developed scalable ML pipelines and real-time dashboards using Flask and JavaScript, streamlining analytics and cutting manual monitoring time by 30% across enterprise workloads
- Integrated secure data access layers using Azure Data Lake and RBAC policies, reducing query latency by 30% and improving role-based data governance across distributed agent systems
- Automated agent deployment pipelines with Kubernetes, Docker, and CI/CD, reducing release cycles by 30% while enhancing resiliency, monitoring, and fault recovery across distributed AI applications

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

- Developed RESTful APIs to power internal agent services on Azure, optimizing query routing and improving response times by 25% for dynamic workflow automation
- Created infrastructure-as-code modules with Terraform and Azure Functions to automate secure multi-agent provisioning workflows, decreasing setup time by 40% and improving audit consistency

Xane.ai **June 2020 – September 2020**

Artificial Intelligence Engineer

- Deployed TensorFlow-based real-time vision models with adaptive threshold tuning, achieving 90% detection accuracy across dynamic environments; enabled real-time alerts for crowd safety through vision-to-audio pipelines

APPLIED PROJECTS

Image Alchemist: AI-Driven eCommerce Image Enhancement Link

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

- Enhanced product visuals by improving clarity, shadows, and layout using OpenCV and Stable Diffusion, ensuring eCommerce compliance and increasing image quality across catalogs
- Automated multi-style background generation and built a real-time image editing system, reducing manual editing time by 40% and enabling seamless user-driven enhancements

Innovative Text-to-Video System for Multi-Modal Content Creation Link

Tech Stack: Fast API, Lang Chain, Transformers, RAG, OpenAI API, Model Scope, TensorFlow, PyTorch, Runway AI

- Built a modular LLM-based agent system integrating RAG and Transformers to autonomously generate video content, improving pipeline scalability by 30% and enabling multi-step prompt chaining
- Integrated external LLM and vision models into SaaS agent framework with prompt orchestration logic, reducing video response latency by 20% and enabling real-time multimodal inference

Cloud-Native Application (Cloud Computing Google Cloud Platform) Link

Tech Stack: JavaScript, GCP, Postman, GitHub, Terraform, Packer, MySQL

- Automated GCP infrastructure with Terraform and Packer, cutting VM provisioning time by 50% and improving deployment consistency for AI workloads
- Secured cloud environments using VPC peering and encryption keys, while integrating CI/CD pipelines to reduce deployment errors by 40% and boost operational efficiency by 25%

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