UDAY SHANKAR GATTU

(617) 971-7892 | udaygattu9949@gmail.com | linkedin.com/in/udayshankargattu/ | github.com/UdayGattu | Google Scholar

EDUCATION

Northeastern University, Boston, MA **April 2025**
Master of Science in Software Engineering Systems *GPA: 3.6*

- **Courses**: Advanced Techniques with LLMs, Generative AI, NLP, Cloud Computing, Responsible AI, Algorithms.
- **Graduate Teaching Assistant**: Generative AI, Natural Language Processing for Fall2024, Prompt Engineering for spring 2025.

Harvard University, Boston, MA **October 2022**
CS50 Introduction to Computer Science.

Goka Raju Ranga Raju Institute of Engineering and Technology, Hyderabad, India **June 2022**
Bachelor of Technology in Mechanical Engineering.

SKILLS

Programming Skills: Python (PyTorch, TensorFlow, Scikit-learn), MySQL, Fast API, Django, C++, Java.

AI Concepts: Large Language Models (LLMs), Fine-tuning, Transformers, BERT, GPT, Generative Adversarial Networks (GANs), Reinforcement Learning, Encodings, Retrieval-Augmented Generation (RAG), LLM Agents, Prompt Engineering, LangChain.

Machine Learning Frameworks: CNN, LSTM, Hugging Face, Stable Diffusion, Deep Learning, Data Science, OpenCV, Flask.

Cloud Computing Skills: AWS, GCP, Vertex AI, Azure, Amazon Sage Maker, Git, Azure DevOps.

Certification: AWS Academy Cloud Practitioner, Microsoft Azure Fundamentals & Developer, EDX CS50.

WORK EXPERIENCE

Northeastern University | Boston, MA **January 2025 – Present**
Graduate Teaching Assistant – *Prompt Engineering for Generative AI*

- Deployed predictive models on Azure using Spark for real-time predictions across large datasets, reducing response times by 30% and improving operational efficiency by 20%.
- Designed and implemented Databricks-integrated scalable ML pipelines for fault-tolerant storage and queries, reducing latency by 30% and storage costs by 15%.
- Built and deployed SaaS-oriented machine learning applications using Python and Flask, achieving a 25% improvement in deployment efficiency and reducing latency by 15% in production systems.
- Collaborated with cross-functional teams to implement robust MLOps workflows, increasing deployment speed by 30% and ensuring 99.9% system uptime for enterprise-grade solutions.

Graduate Teaching Assistant – *Generative AI, Natural Language Processing* **September 2024 – December 2024**

- Deployed predictive models on Azure using Spark for real-time predictions across large datasets, reducing response times by 30% and improving operational efficiency by 20%.
  Designed and implemented Databricks-integrated scalable ML pipelines for fault-tolerant storage and queries, reducing latency by 30% and storage costs by 15%.

Tata Consultancy Services | India **June 2022 – August 2023**
Machine Learning Engineer - *Cloud Exponence Microsoft Azure*

- Deployed predictive models on Azure using Spark for real-time predictions across large datasets, reducing response times by 30% and improving operational efficiency by 20%.
- Designed and implemented Databricks-integrated scalable ML pipelines for fault-tolerant storage and queries, reducing latency by 30% and storage costs by 15%.
- Built and deployed SaaS-oriented machine learning applications using Python and Flask, achieving a 25% improvement in deployment efficiency and reducing latency by 15% in production systems.
- Collaborated with cross-functional teams to implement robust MLOps workflows, increasing deployment speed by 30% and ensuring 99.9% system uptime for enterprise-grade solutions.

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

- Developed backend services with Python & Flask, implementing RESTful APIs and databases, boosting data retrieval by 25% and scalability.
- Optimized SQL/NoSQL databases, reducing query latency by 30% and improving system reliability for high-load applications.
- Earned Azure Fundamentals certification, enhancing cloud deployment, authentication, and monitoring, improving performance by 20%.

Edu Skills | India **June 2020 – September 2020**
AI-ML Engineer Intern

- Trained in advanced machine learning technologies through AWS Academy, mastering Amazon SageMaker to develop and optimize machine learning models, led to a 30% improvement in their accuracy and performance.
- Gained extensive hands-on experience with AWS cloud technologies, focusing on the scalable deployment of AI solutions that improved operational efficiency by 25% across various projects.

- Spearheaded the deployment of machine learning and deep learning models in the cloud, achieving a 20% increase in project efficiency and a 15% reduction in system downtime.

The Sparks Foundation | India **June 2020 – September 2020**

Data Science and Business Analytics intern

- Demonstrated proficiency in the application of Supervised Machine Learning techniques, notably Linear Regression, for predictive modeling.
- Employed Unsupervised Machine Learning methods, specifically Iris Clustering, for exploratory analysis. Utilized Decision Tree algorithms to inform decision-making processes.
- Conducted Stock Market Prediction through a comprehensive approach involving both numerical and textual analysis.
- Performed in in-depth Business Analytics for strategic insights

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

Artificial Intelligence Engineer

- Developed TensorFlow models for mask detection in retail, achieving 85% accuracy and 87% faster detection speeds.
- Implemented preprocessing techniques like resizing and normalization, increasing recognition accuracy by 25%.
- Deployed AI solutions on edge devices, enhancing real-time performance and usability in customer-facing systems.

APPLIED PROJECTS

Innovative Text-to-Video System for Multi-Modal Content Creation | Northeastern University | *Link* **January 2025**

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

- Built a text-to-video system using RAG and fine-tuned models, increasing video generation accuracy by 30%.
- Evaluated DALL-E and Model Scope for multi-modal applications, optimizing system capabilities for creative content.
- Developed APIs for seamless SaaS integration, reducing content delivery time by 20%.

Interactive AI-based Tutor for Physics Education | Northeastern University | *Link* **November 2024**

Tech Stack: *Fast API, Lang Chain, Transformers, RAG, OpenAI API, PyTorch, MySQL*

- Developed an AI-driven Physics Tutor Bot designed to enhance learning experiences for students by providing interactive and adaptive problem-solving support. This project focuses on bridging theoretical concepts with real-world applications to foster a deeper understanding of physics.
- Evaluated DALL-E and Model Scope for multi-modal applications, optimizing system capabilities for creative content.
- Developed APIs for seamless SaaS integration, reducing content delivery time by 20%.

AutoMate: AI-Powered Car Assistant | Northeastern University | *Link* **January 2025**

Tech Stack: *Voiceflow, AI Chatbot, NLP, FastAPI, JSON, Dialog Management*

- Developed an AI-driven virtual assistant for car companies, automating appointment bookings, test drive scheduling, and agent interactions, reducing manual scheduling efforts by 60%.
- Implemented intelligent error handling and adaptive fallback workflows, reducing user drop-off rates by 35% and improving conversation success rates by 50%.
- Optimized dialogue workflows and structured user interactions, cutting response time by 40% and enhancing customer satisfaction scores by 30%.

Cloud-Native Application (Cloud Computing Google Cloud Platform) | Northeastern University | *Link* **May 2024**

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

- Automated GCP infrastructure using Terraform, reducing VM setup times by 50% and improving deployment efficiency.
- Secured cloud resources with VPC peering and encryption keys, cutting deployment errors by 40%.
- Integrated CI/CD pipelines, enabling seamless updates and enhancing operational efficiency by 25%.

Dermatological Image Generation Using Latent Diffusion Models | Northeastern University **May 2024**

Tech Stack: *Stable Diffusion, Hugging Face, Clip, GCP, Transformers, Unet, PyTorch*

- Developed text-to-image models using CLIP and Stable Diffusion, enhancing healthcare datasets by 20%.
- Built scalable pipelines on GCP, transforming text inputs into photorealistic images for diagnostic use.
- Added predictive titling for streamlined diagnostics, improving accuracy in AI-generated medical reports.

Travel Agent Chatbot | Northeastern University **May 2024**

Tech Stack: *Python, LangChain, Fast API, MySQL, Transformers, NLP SQL, PyTorch*

- Designed and implemented an advanced chatbot for a travel agency utilizing Fast API, Large Language Models (LLMs), and ChatGPT, focusing on real-time travel inquiries and personalized package recommendations.
- Developed both front-end and back-end systems, incorporating AI technologies to streamline user interactions and improve service delivery.
- Engineered a custom Natural Language Processing (NLP) Query-SQL engine to enable dynamic database interactions, enhancing the flexibility and responsiveness of the chatbot, supporting market research and increasing customer engagement.
- Achieved a 30% increase in customer engagement by integrating personalized travel options and real-time assistance, resulting in a more interactive user experience.
- Boosted booking conversions by 25% using targeted AI-driven recommendations and efficient handling of customer inquiries.

RESEARCH EXPERIENCE

- *Johnson Cook Material Model*, **Materials Today: Proceedings**, 2022. **Citations: 118*** *Link*