UDAY SHANKAR GATTU

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

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, C++, Java, C#.

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.

Web Frameworks: JavaScript, Flask, Fast Api, Django, Html, CSS, Tailwind CSS, React JS.

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*

- Guide 40+ students in prompt engineering and generative AI, improving coursework completion by 25%.
- Create hands-on assignments, real-world case studies, and videos to enhance student engagement by 30%.
- Maintain attendance records, streamline troubleshooting, reducing coursework-related queries by 35% and ensuring participation.

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

- Mentor 60+ students in applied AI and NLP, providing structured guidance on diffusion models, transformers, RAG, and language model fine-tuning.
- Design and refine course materials, ensuring alignment with industry standards and enhancing clarity in Gen AI and NLP concepts.
- Provide detailed feedback on assignments and hands-on projects, helping students bridge theoretical AI knowledge with real-world applications.

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

- Deployed predictive models on Azure Cloud Exponence using Microsoft Azure and Python, automating cloud governance workflows and reducing operational costs by 15% while ensuring 99% uptime.
- Developed scalable ML pipelines and interactive dashboards using Flask, JavaScript, and Chat JS, improving real-time analytics and reducing manual monitoring efforts by 30%.
- Implemented distributed storage and query systems with Azure Data Lake, SQL, and NoSQL databases, reducing query latency by 30% and enhancing system reliability for enterprise workloads.
- Automated MLOps deployment workflows using Docker, Kubernetes, and CI/CD pipelines, accelerating model release cycles by 30% and improving fault tolerance for large-scale AI applications.

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

- Designed and developed backend RESTful APIs for cloud automation using Python, Flask, increasing data retrieval efficiency by 25% and improving API response times.
- Optimized SQL and NoSQL database performance using MySQL and MongoDB, reducing query execution time by 30% and improving scalability for high-traffic applications.
- Built infrastructure automation scripts using Azure Functions, Terraform, and Bash scripting, reducing cloud provisioning time by 40% and enhancing operational efficiency.

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

- Developed and optimized ML models using Amazon SageMaker, improving accuracy and performance by 30%
- Deployed AI solutions on AWS Cloud, leveraging EC2, S3, and Lambda, increasing operational efficiency by 25%.
- Implemented deep learning deployments, enhancing project scalability by 20% and reducing system downtime by 15%.

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

Data Science and Business Analytics intern

- Developed predictive models using Supervised Learning, improving forecasting accuracy for business insights.
- Applied Unsupervised Learning techniques, including clustering and decision trees, enhancing data-driven decision-making.
- Built a stock market analysis pipeline, integrating numerical and textual data, improving trend forecasting efficiency.

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

Artificial Intelligence Engineer

- Developed deep learning models for face mask detection using TensorFlow and Computer Vision, achieving 95% accuracy.
- Optimized data collection and preprocessing, improving detection speed by 20% through precise parameter tuning.
- Deployed real-time AI models, ensuring 90% efficiency in varied lighting and crowd conditions for public health safety.

APPLIED PROJECTS

Image Alchemist: AI-Driven eCommerce Image Enhancement | Northeastern University | *Link* **January 2025**

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

- Enhanced product images by improving clarity, shadows, and positioning, ensuring compliance with eCommerce standards.
- Automated background generation with four styles, reducing manual editing time by 40% and improving visual consistency.
- Developed an interactive image processing system, enabling real-time enhancements and seamless user control.

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, enabling interactive problem-solving and improving student engagement by 40%.
- Implemented real-world application modules, bridging theoretical physics concepts with practical examples, increasing comprehension rates by 35%.
- Designed dynamic Q&A and adaptive hints, reducing query resolution time by 30% and enhancing the learning experience.

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

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

- Built an AI-driven virtual assistant for car companies, automating scheduling and reducing manual booking efforts by 60%.
- Integrated error-handling and adaptive fallback workflows, decreasing user drop-off rates by 35% and improving conversation success by 50%.
- Optimized dialogue workflows and structured interactions, cutting response time by 40% and boosting customer satisfaction 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 | *link* **May 2024**

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

- Designed an AI-powered travel chatbot, automating real-time inquiries and increasing user engagement by 30%.
- Developed a custom NLP Query-SQL engine, enabling dynamic database interactions and improving response accuracy.
- Implemented AI-driven travel recommendations, boosting booking conversions by 25% and optimizing customer experience.

RESEARCH EXPERIENCE

- *Johnson Cook Material Model*, **Materials Today: Proceedings**, 2022. **Citations: 119*** *Link*