Sai Nithin Krishna Souram

Chicago, IL, USA | +1 331-666-3092 | sai.saianu.nithin@gmail.com | LinkedIn | GitHub

Professional Summary

Senior AI Engineer with 5+ years of experience designing and deploying generative AI and machine learning solutions. Skilled in Python-based back-end development, API design with FastAPI/Flask, cloud deployments on AWS SageMaker, and integrating HuggingFace transformers and OpenAI APIs. Proven track record of automating ML pipelines to boost accuracy by 25% and reduce processing time by 30%, enabling efficient, scalable AI applications.

Technologies

- **Deep Learning & GenAI**: Vision Transformers, CLIP, DINO, GPT-2/3, RAG, HuggingFace Transformers, OpenAI APIs, ML Model Deployment, Generative AI Technologies, Large Language Models, Interest In AI
- **Programming & Web Development**: Python, C, Java, R, Bash, JavaScript, HTML, CSS, Full-stack Development
- **Frameworks & Libraries**: TensorFlow, PyTorch, Keras, OpenCV, Flask, Django
- **MLOps, Deployment & Cloud**: Docker, Kubernetes, FastAPI, AWS SageMaker, MLflow, Jenkins CI/CD, Cloud Platforms (AWS), Git
- **APIs & Database Management**: RESTful API Design, Database Management

Experience

Tata Consultancy Services (TCS) **Jul 2019 - Aug 2023**
Artificial Intelligence Engineer *Hyderabad, India*

- Designed and deployed machine learning pipelines that automated internal reporting processes, reducing manual effort by 40% and saving over 150 team hours per quarter.
- Led the development of comprehensive AI applications by managing phases like data preprocessing, model training, deployment, and monitoring, effectively applying MLOps best practices on AWS to streamline operations, enhance application performance, and improve predictive accuracy.
- Architected and optimized ML models to enhance performance, reliability, and scalability across distributed systems and database management platforms, resulting in a 25% increase in model accuracy and a 30% reduction in processing times.
- Conducted research on computer vision and transformer-based architectures, integrating large language models and generative AI technologies into production applications to improve functionality and enable advanced features.
- Mentored junior engineers on deep learning frameworks, coding standards, and ML system design patterns, resulting in faster project completion times and improved quality of the team's codebase through a culture of collaboration and continuous learning.
- Collaborated cross-functionally with product owners and cloud architects on full-stack development projects, aligning AI-driven front-end and back-end components with strategic objectives.
- Automated deployment workflows and implemented CI/CD pipelines using Docker, Kubernetes, and Git for robust model serving and version control.
- Provided technical leadership in design reviews, establishing RESTful API design, ML governance, model versioning, and monitoring frameworks.

Education

DePaul University **Aug 2023 - Jun 2025**
MS, Artificial Intelligence *chicago*

- **GPA:** 3.5 / 4.0
- **Achievements:** Graduate Assistant
- **Coursework:** Deep Learning, Reinforcement Learning, NLP, Computer Vision, AI Ethics

G. Pullaiah College of Engineering & Technology **2015 - 2019**
B.Tech, Electrical & Electronics Engineering *india*

Projects

Autonomous Driving using CARLA Simulator **Jan 2025 - Jul 2025**
https://github.com/Nithin9Krishna/AI2_Final_Project
Depaul University *chicago*

- Trained reinforcement learning models (PPO, A2C) for autonomous vehicles in simulated environments
- Architected ML pipelines focusing on lane navigation and real-time decision-making for self-driving agents
- Utilized CARLA simulator to test, evaluate, and refine AI models for real-world applicability

Multi-modal Generative AI Project **Jan 2023 - May 2023**
Depaul university *chicago*

- Developed a multi-modal transformer integrating text and image features for contextual response generation
- Fine-tuned vision-language models (CLIP + GPT-2) using PyTorch and Hugging Face on custom QA datasets

- Designed prompt tuning and evaluation pipelines aligning outputs with visual cues and human feedback

Research & Publications

- Cosmos World Foundation Model Platform for Physical AI.*Analyzed NVIDIA's Cosmos WFM for Physical AI, focusing on modular architecture, video curation, and tokenization. Recommended stronger application-level validation, benchmark comparisons, and accessibility improvements for lower-resource research environments*
- Explainable AI for Autonomous Driving.*Reviewed a comprehensive survey of XAI methods for AVs, including visual, RL-based, and decision tree explanations. Provided feedback on framework integration, user-oriented design, and regulatory compliance to improve transparency and adoption.*
- Application of Robotics.*Authored a detailed study on robotics applications in delivery, healthcare, agriculture, construction, and surgery. Analyzed limitations, environmental challenges, and optimization strategies for improved performance in varied conditions.*
- Research – Exploring the Design, Sensors, and Challenges of Modern Robotic Systems.*Examined humanoid, industrial, and autonomous mobile robot designs, highlighting sensor roles in perception, HRI, and adaptive behavior. Identified limitations in mobility, efficiency, and operational environments, proposing design enhancements.*
- Research – Application of Robotics (Defense & Industrial Focus).*Explored robotics applications across defense, healthcare, manufacturing, and exploration. Emphasized AI integration in UAVs, bomb disposal robots, AGVs, and swarm robotics, with insights on ethics and future military technology trends.*
- Research – Exploring Minimal-Sensing Robotics: Passive Dynamic Walkers, Weasel Ball Robots, and Soft Robot Manipulators.-*Investigated low-sensor robotic designs relying on mechanical properties and natural dynamics. Presented mathematical models and application scenarios for energy-efficient, cost-effective robotics in controlled environments.*