Varun Phanindra Shrivathsa

+1 312-478-2342 | varunpshrivathsa@gmail.com | Chicago, IL | (Open to Relocation) | linkedin.com/varps
www.varun-p.com | github.com/varunpshrivathsa | medium.com/@varunpshrivathsa

Summary

MS in Computer Science student at UIC with hands-on experience building machine learning and GenAI systems from data ingestion to production deployment at scale in real-world environments. Experienced in end-to-end ML workflows through internships, applied research and projects. Seeking Machine Learning Engineer or AI Software Engineer roles.

Skills

Programming Languages: Python, C/C++, SQL
Machine Learning: PyTorch, Scikit-learn, SparkML, Hugging Face, OpenCV
Generative AI: LLMs (GPT, LLaMA), LangGraph, CLIP, RAGs, LoRA
Frameworks & Systems: FastAPI, gRPC, Ray, Optuna, React, Redis, CUDA, Linux
Data Engineering & Databases: PostgreSQL, MongoDB, Pinecone, FAISS
MLOps & Deployment: MLflow, Docker, Kubernetes, AWS, GCP, GitHub
Monitoring & Testing: Prometheus, Grafana

Education

University of Illinois Chicago (UIC) Chicago, USA
Master of Science in Computer Science *Aug 2024 – May 2026*
- Thesis: Autonomous Robotic Framework for Navigation, Exploration & Mapping

Dayananda Sagar University Bangalore, India
Bachelor of Technology in Computer Science and Engineering *Jun 2020 – Jun 2024*

Experience

Machine Learning Engineer Intern Jan 2025 – Dec 2025
G19 Studio - Academic–Industry Collaboration (UIC) *Chicago, USA (Remote)*
- Worked with G19 Studio through UIC's Advanced NLP course and continued as an extended internship.
- Developed 'TwinVerse' a human digital twin platform for stress mitigation using real-time wearable sensor data.
- Built and evaluated time-series forecasting models for physiological signals improving simulation convergence by 25%.

Research Intern Jun 2023 – Jun 2024
Indian Institute of Science (IISc) - NIAS *Bangalore, India*
- Built GIS image-processing pipelines for environmental impact analysis of large-scale infrastructure projects.
- Trained and evaluated multi-spectral CNN models achieving 92% segmentation accuracy for land-use classification.
- Secured KSCST research funding to extend geospatial validation and automate GIS processing pipelines.

Projects

VideoTune: Multimodal Retrieval & Recommendation
Stack: Python, PyTorch, CLIP, Wav2Vec2, FAISS, AWS, Docker, Prometheus
- Built a multimodal retrieval and recommendation system to rank videos using visual, audio, and text embeddings.
- Designed unified representation and retrieval pipelines with structured optimization to support real-time inference.
- Achieved a 12% accuracy improvement over CLIP only baseline models and deployed as a scalable, monitored service.

SmallGPT: Fine-Tuned Transformer Language Model
Stack: Python, PyTorch, CUDA, LORA, KV Cache
- Built an end-to-end transformer language model training and inference pipeline in PyTorch for scalable deployment.
- Trained a 100M-parameter model on TinyStories dataset with stable convergence across training runs.
- Optimized inference efficiency using LORA fine-tuning and caching to reduce latency in production settings

GenCost: Multi-Agent LLM Cost Optimization
Stack: FastAPI, LLM orchestration, PostgreSQL, Redis, AWS, CI/CD
- Built an adaptive LLM routing platform to optimize inference cost under latency and quality constraints.
- Designed Multi-agent decision pipelines to dynamically select models based on real-time performance signals.
- Reduced inference costs by 40% and deployed a scalable, monitored backend with 99.9% uptime.

Certifications & Publications

Databricks Machine Learning Engineer Professional *Databricks — January 2026*
Environmental Impact Analysis using Satellite Image Processing: Bangalore STRR *2024 IEEE — Published*
Deepfake Detection using LSTM and XResNet *IJRASET — Published*