

Sunny Soni

sunnysoni2021@gmail.com <https://sunnysoni.me> [linkedin.com/in/sunny-soni1089](https://www.linkedin.com/in/sunny-soni1089) github.com/Sunny-Soni00

Summary

B.Tech student in **Digital Transformation** with hands-on experience in **machine learning, generative AI, computer vision, and AI deployment**. Experience spans applied ML research, perception systems, edge deployment, and LLM-based applications. Interested in developing and evaluating AI systems for practical applications, with particular interest in intelligent systems, generative AI, and deployment-oriented machine learning.

Education

Atria University, Bengaluru
B.Tech in Digital Transformation

2023 – 2027
CGPA: 8.5/10

Experience

AI Intern

July 2026

Atria University – CoE eMobility

Personal Rapid Transit (PRT) Perception & Autonomy

- Independently developed an early-stage **PRT perception pipeline**, working end-to-end across real-world data collection, annotation, track segmentation, object detection, output fusion, and real-time perception testing.
- Collected and prepared track video data, annotated approximately **700–900 frames in CVAT**, trained a **U-Net** for track segmentation and **YOLOv8** for pedestrian and road-relevant object detection, and fused both outputs into a unified perception pipeline.
- Evaluated the perception system on **recorded track footage and live camera input**, observing reliable segmentation on relatively straight track sections while identifying limited training-data diversity as the primary constraint on broader robustness.
- Extended the work toward the autonomy stack by installing and testing **ROS 2 Humble and Autoware** on an **NVIDIA Jetson AGX Orin 64 GB**, and investigating integration requirements for LiDAR, radar, GNSS, and other sensors.
- Explored **LingBot-Map** for video-based 3D reconstruction, generating a short local reconstruction while investigating GPU-memory constraints and the practical feasibility of mapping approaches for the platform.

Research Experience

Brain Tumor Segmentation and Edge Deployment using Deep Learning

3 months

Manuscript submitted for publication

[\[GitHub\]](#)

- Developed a deep-learning pipeline for brain-tumor segmentation from multimodal MRI data using a **2.5D approach**, comparing U-Net, AttentionDRUNet, and MobileNetV2-UNet architectures.
- Investigated the trade-off between segmentation performance and computational efficiency using **Hybrid Focal-Dice loss**, local normalization, and architectural variations for resource-constrained deployment.
- Converted trained models through **ONNX and TensorRT FP16** and deployed them on an **NVIDIA Jetson Nano**, evaluating segmentation quality alongside latency, throughput, model size, and power consumption.
- Achieved **83.08% Dice, 82.79% recall, 53.56 ms latency, and 5.08 W power consumption** for the AttentionDRUNet deployment configuration on Jetson Nano.

Projects

Legal Chatbot | Llama-3, RAG, Agentic AI, Groq

3 months [\[GitHub\]](#)

- Developed an LLM-based legal assistant using **Llama-3 and retrieval-augmented generation (RAG)** to ground responses in relevant legal material rather than relying solely on model knowledge.
- Designed an **agentic workflow** combining the language model with retrieval and supporting tools for multi-step legal information processing.
- Built supporting capabilities including a real-time speech-to-text interface and an NLP document analyzer for legal clause extraction and highlighting.

Reinforcement Learning – Pong | DQN, Reinforcement Learning

Project

- Developed a Pong environment and reinforcement-learning agent, training and evaluating the policy through recorded gameplay, training curves, reward analysis, and performance metrics.
- Used a **DQN-based approach** with training analysis and supporting experimentation, documenting the methodology and results through a technical presentation and recorded demonstration.

Speech Enhancement | Signal Processing, Audio Analysis

Technical Study

- Conducted a comparative study of speech-enhancement approaches, analyzing the resulting audio and supporting signal-level visualizations.
- Presented the methodology, comparison, and findings through a self-developed technical presentation and supporting plots.

Technical Skills

Machine Learning & AI: PyTorch, TensorFlow, Scikit-learn, XGBoost, CNNs, RNNs/LSTMs, Seq2Seq, Reinforcement Learning

Generative AI: LLMs, Llama-3, RAG Pipelines, Agentic AI, Claude Code, Prompt Engineering, Hugging Face, Groq

Computer Vision: OpenCV, YOLO, U-Net, CVAT, Image Processing, Medical Image Segmentation

Programming: Python, SQL, Bash/Shell

Cloud, MLOps & Deployment: AWS (S3, EC2, SageMaker, CodeBuild, CodeCommit, CodePipeline, CodeDeploy), ONNX, TensorRT, NVIDIA Jetson Nano, NVIDIA Jetson AGX Orin, ROS 2 Humble, Autoware, Docker, Git/GitHub, Google Cloud Platform, Flask, NumPy, Pandas, Google Colab

Honors & Achievements

Aspire Leaders Program Finalist | Aspire Institute

Jun 2025 – Aug 2025

- Selected as one of **6,175 finalists from 36,050 applicants** globally (Top 17% of Cohort 3).
- Completed a leadership curriculum designed by **Harvard University faculty**, with a focus on AI and social impact.

Certifications

Natural Language Processing Specialization | DeepLearning.AI (Coursera)

[\[Certificate\]](#)

Getting Started with AI on Jetson Nano | NVIDIA Deep Learning Institute

Relevant Coursework

Machine Learning & AI: Advanced Machine Learning, Agentic AI, Reinforcement Learning, Supervised Learning, Unsupervised Learning

Computer Vision & Systems: Computer Vision, Digital Image Processing, Embedded AI, Software Design & Architecture