Sree Dhyuti Nimmagadda

AI Researcher | Machine Learning Engineer | Data Scientist
sreedhyutin@gmail.com | linkedin.com/in/dhyutin | github.com/dhyutin

EDUCATION

Northwestern University | Evanston, IL — Sept 2024 - Dec 2025
Master's, Artificial Intelligence. CGPA: 3.98/4.00.
Focus: NLP, Deep Learning, GenAI, Machine Learning, Graph Neural Networks, Data Science, High Performance Computing.

Indian Institute of Information Technology, Kancheepuram | Chennai, India — Jul 2019 - May 2024
M.Tech + B.Tech, Computer Science Engineering. CGPA: 8.86/10.00.
Focus: Big Data Analytics, Digital Image Processing, Computer Graphics, Pattern Recognition, Computer Vision.

SKILLS

- **Languages/Platforms:** Python, C, C++, R, Matlab, SQL, Verilog, NASM, AWS, Microsoft Azure, Google Cloud
- **Frameworks**: TensorFlow, PyTorch, Keras, Scikit-learn, XGBoost, LightGBM, nltk, LangChain, OpenCV, PySpark.
- **Statistical Analysis:** Hypothesis Testing, Regression, Time Series Analysis, Predictive Modeling, A/B Testing.
- **Tools/Optimization:** CUDA, OpenMP, MPI, NVIDIA DeepStream, Docker, Git, CI/CD, MongoDB, Kubernetes.

EXPERIENCE

AI Researcher (Clariti | Remote, USA) — Jan 2026 - present
- Improved an LLM-based PDF plan-tagging system for construction floorplans using prompt engineering and Vertex AI foundation models on GCP, increasing tagging recall from 68.6% to 82.6%.
- Reduced inference cost and latency by optimizing the serving pipeline and consolidating model calls from 2 to 1 per prediction, improving throughput and deployment efficiency on cloud infrastructure.
- Designed a hybrid rule-based + LLM extraction pipeline for automated intake summarization and scope-of-work generation, improving structured information retrieval from noisy construction documents.
- Built a multi-stage floorplan segmentation pipeline orchestrating specialized YOLO models to detect and classify structural components (rooms, doors, walls, fixtures) from PDF blueprints, achieving > 90% precision and converting unstructured layouts into structured spatial representations for downstream reasoning and model fine-tuning.

AI Research Intern (Writer | San Francisco, California, USA) — Jun 2025 - Sept 2025
- Investigated attention mechanism variations & interpretability techniques to enhance long-context handling and transparency in model behavior.
- Designed & implemented an evaluation framework with 18-different datasets over 7 unique tasks for long-context modeling, identifying and addressing logical and coding discrepancies in Writer's evaluation framework.
- Integrated LLM-as-a-judge evaluation and few-shot prompting capabilities into the framework, and conducted manual comparisons with human checks to validate reliability and robustness.

AI Researcher (The Abazeed Lab | Chicago, Illinois, USA) — Mar 2025 - Jun 2025
- Built and deployed a DynUNet-based 3D segmentation model for 117 organs-at-risk (OARs) with 94% cross-validation Dice accuracy, now integrated into the radiation oncology workflow at Abazeed Lab.
- Designed a scalable CT preprocessing pipeline using cube-based volume chunking and conducted experiments with transformer-based models (Swin UNETR, ViT) for comparative performance analysis.

Machine Learning Research Intern (BioSystems & Controls Lab | Chennai, India) — May 2023 - Oct 2023
- Developed a regression model for predicting apple-sugar levels using Near InfraRed Spectrum, attaining a 0.4 R^2-score.
- Engineered a semi-supervised autoencoder regression model for real-time *Lactococcus lactis* bacteria fermentation monitoring, improving validation R^2-score from 0.61 to 0.89, and prediction R^2-score from 0.49 to 0.82.
- Initiated a multivariate T^2 analysis task, uncovering data inconsistencies and refining experimental data curation setup.
- Collaborated with cross-functional biotechnology teammates to integrate regressor to a hardware architecture.

Machine Learning Intern (Tiny Banyan Technologies Pvt Ltd | Chennai, India) — Aug 2022 - Dec 2022
- Deployed YOLOv5 model for real-time detecting potholes and cracks, revamping anomaly detection accuracy to 99%.
- Trained a team of interns to manage GCP for model training, hyperparameter tuning, testing & validation of ML models.

Theoretical Research Intern (IIITDM Kancheepuram | Chennai, India) — May 2021 - Jul 2021
- Created novel non-deterministic polynomial complete algorithms for Steiner tree problems in Split, Interval, and Chordal graphs, resulting in approximately 2x faster runtime compared to existing solutions.

PROJECTS

AutoRL
- Built a multi-agent reinforcement learning fine-tuning framework that converts natural language tasks into automated RL experiments, dynamically generating training plans across PPO, SAC, A2C, and GRPO pipelines.

- Designed an LLM-driven control stack (a RAG-Orchestrator, Doom Loop Sentinel, Env Doctor, Reward Designer, Evaluator) using the OpenAI Agents SDK to autonomously plan, monitor, debug, and optimize large-scale RL training runs.
- Developed a self-healing RL infrastructure with anomaly detection (NaNs, reward plateaus, entropy collapse) and automated hyperparameter recovery, improving training robustness through iterative agent-guided restarts and reward shaping.
- Integrated Weights & Biases Weave + Redis retrieval + Hugging Face for end-to-end observability, real-time distributed coordination, and model deployment, enabling live experiment tracking and reproducible policy benchmarking.

Qbitrade

- Built an end-to-end quantum-enhanced trading research pipeline integrating 14 alpha models (LightGBM, XGBoost, LSTM, VQC) with market microstructure signals and news embeddings for explainable financial prediction.
- Developed a hybrid quantum-classical meta-labeling framework using Qiskit (QSVM + ZZFeatureMap) that outperformed classical SVM baselines on 5/14 models, achieving up to +0.2136 F1 improvement on uncertainty-aware trade selection.
- Trained a Stable-Baselines3 PPO portfolio optimizer with a custom differential Sharpe Ratio reward function, improving portfolio Sharpe from 0.859 to 1.038 (+21%) over baseline allocation strategies.
- Designed Q-Atlas, a quantum market memory retrieval system leveraging quantum covariance kernels and Google DeepMind Gemini embeddings to identify analogous historical market regimes and generate interpretable, LLM-grounded trade narratives.

Multi-Task Embodied Learning for Bimanual Robot Manipulation

- Built a full-stack embodied AI system for bimanual robot learning, integrating teleoperation, calibration, data collection, and deployment via LeRobot + custom web tooling.
- Collected and curated 78 teleoperated demonstrations (~51K frames) for multi-task robotic manipulation, improving policy quality through trajectory-level filtering and dataset aggregation.
- Trained ACT and SmolVLA policies for specialist vs. generalist control, benchmarking task transfer, low-data scaling, and prompt-conditioned execution. Trained a single ACT policy to generalize across two distinct manipulation tasks - utensil stacking and liquid pouring studying multi-task behavior cloning in low-data real-world settings.
- Analyzed execution failures and designed RL-based recovery objectives for robust closed-loop adaptation in real-world manipulation. Won 2nd place at the makermod's Physical AI hackathon.

ACHIEVEMENTS

- **Weave Hacks 4, Weights&Biases, 2026:** One of the 8 finalists out of 70 teams for building AutoRL
- **Quantum AI Hackathon 2026, UC Berkeley:** Secured 1st place in the Finance Track for building Qbitrade
- **Physical AI Hackathon 2026, San Francisco**: Secured 2nd place - built an ACT-based robotic manipulation system
- **IEEE TENCON 2024, Singapore:** Presented my research paper on evaluation metrics for LLMs.
- **OpenCV AI Competition 2022:** Top 45 global finalists out of 1500 teams.