DHIVYA SREEDHAR

(412) 954-7892 ⋄ Pittsburgh, PA
dsreedha@andrew.cmu.edu ⋄ linkedin.com/in/dhivya-sreedhar-03b541168/ ⋄ https://dhivyasreedhar.github.io

EDUCATION

Carnegie Mellon University December 2025
Master of Science (MS) - Information Systems (Machine Learning & Natural Language Processing), GPA: 4.0/4.0 Pittsburgh, PA
Relevant Coursework: Advanced Natural Language Processing, Deep Learning (PhD), Machine Learning in Production, Generative AI

Anna University May 2022
Bachelor of Engineering - Computer Science Engineering, GPA: 8.7/10 Chennai, India
Relevant Coursework: Data Structures & Algorithms, Distributed Systems, Artificial Intelligence, Linear Algebra, Statistics, Probability

WORK EXPERIENCE

Applied AI Research Intern – LLM Reasoning May 2025 – Present
Scale AI *Remote, USA*

- Led large-scale evaluation of frontier LLMs (GPT-4, Claude, LLaMA) on generative reasoning tasks—including multi-hop QA, chain-of-thought inference, and long-context summarization—to benchmark emergent capabilities
- Fine-tuned instruction-following models using SFT, LoRA, and DPO on tasks such as code generation, math problem solving, and open-domain QA; improved coherence and factuality by 18.7% on internal evaluation suites

Computer Vision and Robotics Intern May 2025 – Present
Reclamation Factory (CMU Robotics Startup) *Pittsburgh, USA*

- Built and deployed real-time plastic classification system combining NIR, XRF, and RGB modalities on NVIDIA Jetson AGX Orin, achieving 93.5% accuracy across 6 material categories
- Applied transfer learning and Vision Transformer (ViT) fine-tuning, with domain adaptation and self-supervised pretraining to reduce lighting-sensitive error rates by 27%
- Optimized inference pipelines with TensorRT, ONNX, and CUDA, enabling sub-50ms low-latency predictions for embedded robotic sorting under real-world constraints

Applied Scientist January 2025 — May 2025 (Duration : 5 months)
Bank of New York *Pittsburgh, USA*

- Pioneered human-in-the-loop (HITL) reinforcement learning from human feedback (RLHF) for optimizing conversational flows, implementing automated prompt engineering, A/B testing strategies and red-teaming capabilities that reduced hallucination rates by 43% while ensuring responsible AI model training, optimization & deployment
- Deployed knowledge distillation and parameter-efficient transfer learning techniques for scalable, personalized, multi-channel deployment across 9 core AI capabilities: Content Generation, Anomaly Detection, Decision Reapplication, Code Modernization, Virtual Assistance, Data Migration, Scenario Creation, Prediction/Classification, and Unstructured Data Q&A resulting in $4.2M in annual cost savings

Applied ML Scientist August 2022 — August 2024 (Duration : 2 years)
Zoho Corporation - Part of the Manage Engine - Log360 Cloud OD Team *Chennai, India*

- Designed and deployed real-time anomaly detection systems for cloud log analytics using a combination of z-score analysis, EWMA, Isolation Forests, and autoencoders, improving detection of privilege escalations and rare activity patterns
- Built HTTP Event Collector module for low-latency ML data ingestion, optimizing log stream processing across US, UK data centers

RESEARCH EXPERIENCE

Graduate Research Assistant (Collaboration with Prof Bhiksha Raj Ramakrishnan) January 2025 — Present
Machine Learning for Signal Processing Group, Language Technologies Institute, CMU *Pittsburgh, USA*

- Conducting research on Multimodal Chain-of-Thought (CoT) frameworks for integrating vision-language reasoning in large language models, improving interpretability and structured inference in multi-hop QA tasks
- Engineered scalable CoT prompting and alignment strategies, boosting ScienceQA task accuracy by 16% and reducing reasoning errors by 23% through joint vision-text embeddings and modular decoding

PUBLICATIONS

1. **Typing Reinvented: Towards Hands-Free Input via sEMG , NeurIPS**
2. **Comparison of Deep Learning Techniques for Face Forgery, Journal of Social Network Analysis and Mining ,**
3. **Neural Networks for Music Instrument Recognition**, *A*dvances in Speech and Music Technology: **Springer International**

PROJECTS

emg2qwerty | *PyTorch, TensorFlow, and signal processing libraries (NumPy, SciPy)*
Led the development of EMG2QWERTY, a hands-free neuromusculoskeletal interface translating surface EMG signals into text input for AR/VR and spatial computing platforms such as Apple Vision Pro and Meta Quest. Achieved < 5% character error rate (CER) and < 30 ms latency using a hybrid Conformer-Transformer architecture with spectral feature extraction, self-attention, and CTC loss. Built a real-time beam search decoder with Flan-T5 and GPT-4 Turbo for word- and sentence-level autocorrection. Introduced EMG-specific augmentations (SpecAugment, RandomBandRotation, Temporal Jitter) and causal modeling for generalization and low-latency inference. Delivered seamless mid-air typing with animated hand overlays, advancing tactile-free input for immersive spatial interfaces.

AI Interview Agent | *LangChain, LangGraph, LLMs, NLP, Generative AI, Agentic AI Architecture, Agent Workflows*
Developed a Generative Agentic AI interview agent using RAG, LangChain, and Llama 3 to simulate job interviews. The system vectorizes resumes, cross-checks job descriptions, and generates personalized interview questions with GPT-based models. It conducts voice-interactive conversations, dynamically adapting with context-aware follow-ups. An LLM-as-a-judge module evaluates responses, leveraging NLP and structured information retrieval for AI-driven assessments

LLM-Powered Dataset Auditor | *LangGraph, LangChain, CLIP, FAISS, GPT-4*
Designed a Retrieval-Augmented Generation (RAG) pipeline for visual label validation in image classification datasets. Integrated CLIP embeddings with FAISS to retrieve visually similar images, enabling contextual label judgment by LLM validators (GPT-4/Claude) within a LangGraph agentic workflow. The system audits predictions, flags inconsistencies, and logs decisions to structured JSON/CSV formats

MyTorch | *Python, NumPy, PyTorch*
Built a custom deep learning library from scratch using Python and Numpy. Built the Autograd engine (back and forward propagation), implemented Loss functions, Optimizers, Linear layer, Convolutional layer, Recurrent layer, Batchnorm2D, Meanpool2D, Maxpool 2D, sequence packing and other pytorch utilities. Implemented MLPs, CNNs, LSTMs, RNNs, GANs, GNNs & GRUs using the custom library

SKILLS

Programming/Scripting Languages: Java, Python, C++, C, C#, MySQL, PHP, Javascript, HTML5 / CSS3
Frameworks & tools: Struts, Flask, Django, CUDA, GNU, AWS, GCP, NodeJS, ReactJS, AngularJS, Containerization (Docker), Kafka, Kubernetes, NVIDIA GPUs, SQL (Snowflake, BigQuery),REST APIs, AJAX, OpenGL, Apache Beam, Spark MapReduce, Tableau
Machine Learning : Computer Vision, NLP, Large Language Models, Image & Video Processing, Multimodal Learning
Libraries: Tensorflow, PyTorch, OpenCV, Numpy, Pandas, Matplotlib, Scikit-Learn, Keras, Jax, PySpark, VLLM, LlamaFactory, TensorRT