DHIVYA SREEDHAR

(412) 954-7892 ◇ Pittsburgh, PA
dhivyasreedhar@gmail.com ◇ 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 Machine Learning Intern – LLM Reasoning & Evaluation May 2025 – Present
Scale AI *Remote, USA*

- Developed and evaluated multi-turn agentic systems using LangGraph and LangChain, integrating RAG pipelines with vector databases (Qdrant, FAISS) and LLM tools for contextual task execution
- Fine-tuned HuggingFace LLMs for code generation and QA using LoRA/DPO strategies; deployed via vLLM and accelerated using DeepSpeed on AWS

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

- Developed a multi-modal robotic sorting system using NIR, XRF, and RGB sensor fusion on NVIDIA Jetson AGX Orin, applied transfer learning and Vision Transformer (ViT) fine-tuning, achieving 93.5% classification accuracy across 6 material categories
- Performed ETL, optimized, deployed cloud-based inference pipelines on AWS SageMaker, integrating with EMR, Lambda, CloudWatch, and S3 Data Lake for event-driven automation of a multimodal AI system (image, audio, text). Applied TensorRT, ONNX, quantization-aware training, and model pruning to achieve ¡50ms latency, low-power inference, and high throughput. Visualized model performance using confusion matrices, precision, recall, and latency distribution plots

Applied Scientist January 2025 — May 2025
Bank of New York *Pittsburgh, USA*

- Developed human-in-the-loop RLHF workflows, LoRA with automated prompt engineering, reinforcement learning, A/B testing, and red-teaming, reducing hallucinations by 43% and improving responsible model deployment.
- Applied knowledge distillation and efficient transfer learning for multimodal financial data (image, tabular, text) fusion; implemented scalable MapReduce pipelines on Spark over AWS S3, enabling deployment across 9 AI tasks and driving $4.2M annual savings

Applied ML Scientist August 2022 — August 2024
Zoho Corporation - Part of the Manage Engine - Log360 Cloud OD Team *Chennai, India*

- Designed and deployed real-time anomaly detection pipelines for Log360 Cloud SIEM, applying z-score, EWMA, Isolation Forests, and autoencoders to detect threats including privilege escalations, lateral movement, and rare event anomalies
- Built and productionized scalable machine learning pipelines for ingesting and analyzing cloud security logs, leveraging Splunk-compatible HTTP Event Collectors, Airflow, Docker, and AWS Lambda
- Developed RAG-style analytics over large-scale observability and cybersecurity log data (via S3, pgvector, Qdrant), integrating with automated vulnerability triage workflows to reduce MTTD (Mean Time to Detect)

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 & PATENTS

1. **Acoustic-based resin identification using contrastive learning; Applied for patent**
2. **Typing Reinvented: Towards Hands-Free Input via sEMG , NeurIPS**
3. **Neural Networks for Music Instrument Recognition**, *A*dvances in Speech and Music Technology: **Springer International**

PROJECTS

emg2qwerty — *PyTorch, TensorFlow, NumPy, SciPy, Signal Processing* Led the development of a neuromusculoskeletal interface translating surface EMG signals into text input for AR/VR and spatial computing platforms. Achieved <5% CER and <30ms latency using a hybrid Conformer–Transformer with spectral feature extraction, self-attention, and CTC loss. Built a real-time beam search decoder with Flan-T5 and GPT-4 Turbo for autocorrection. Introduced EMG-specific augmentations (SpecAugment, RandomBandRotation, Temporal Jitter) and causal modeling for low-latency inference. Applied large-scale, high-dimensional time series analysis for robust cross-user generalization.

Agentic Interview Simulator — *LangChain, LangGraph, LLMs, NLP, Generative AI, RAG, Agent Workflows* Developed a Generative Agentic AI system simulating adaptive job interviews using Retrieval-Augmented Generation with Llama 3. Engineered large-scale embedding and indexing pipelines (FAISS, Hypothetical Document Embeddings) for structured information retrieval. Integrated multi-modal context from resumes and job descriptions to generate personalized, context-aware interview questions. Implemented voice-interactive workflows, reinforcement learning for adaptive questioning, and an LLM-as-a-judge module for response evaluation.

MyTorch — *Python, NumPy, PyTorch* Built a custom deep learning library from scratch with an Autograd engine for forward/backpropagation, loss functions, optimizers, linear/convolutional/recurrent layers, batch normalization, and pooling. Implemented MLPs, CNNs, LSTMs, RNNs, GANs, GNNs, and GRUs, showcasing deep learning and probabilistic modeling capabilities.

Retrieval Augmented Generation — *Python, PyTorch, FAISS, Hugging Face* Implemented an end-to-end RAG system for large-scale Q/A from scratch, including knowledge corpus curation, synthetic data generation, model fine-tuning, and statistical modeling for query optimization. Integrated state-of-the-art embedding/indexing methods, Hypothetical Document Embeddings (HyDE), document summarization, and model quantization for production-ready performance.

Mini Llama — *Python, PyTorch, Q-LoRA, Hugging Face* Implemented Llama 3.1 in PyTorch with Q-LoRA parameter-efficient fine-tuning, pretraining and finetuning from scratch on large-scale, high-dimensional datasets. Applied multi-modal extensions, reinforcement learning strategies, and optimization for scalability and downstream NLP tasks in production environments.

Movie Recommendation System — Built and deployed a content-based movie recommendation system serving 1M+ simulated users. Containerized with Docker and automated retraining pipelines via Jenkins to ensure scalability and 99% uptime. Utilized MLflow for model versioning and integrated Prometheus and Grafana for real-time monitoring, metrics tracking, and performance optimization.

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, CI/CD, MapReduce, Tableau
ML Libraries & tools: Tensorflow, PyTorch, OpenCV, Numpy, Pandas, XGBoost, HuggingFace Transformers, OpenAI APIs, Scikit-Learn, Keras, Jax, PySpark, VLLM, LlamaFactory, TensorRT, MLOps, MLflow, Kubeflow