# DHEERAAJ PINJALA

Boston, MA | pinjala.d@northeastern.edu | +1 (857) 426-1304
linkedin.com/in/dheeraaj-pinjala | github.com/dheeraajpinjala | dheeraajpinjala.vercel.app

## SUMMARY

AI/ML Engineer and CS Master's student with nearly **3 years of experience** across the **software development lifecycle** and building and scaling **end-to-end AI products**. Focused on delivering **RAG systems** and **agentic workflows** by building robust **development and deployment pipelines,** with hands-on experience **fine-tuning** models and **evaluating and optimizing model performance** in production.

## EDUCATION

**Northeastern University, Khoury College of Computer Sciences, Boston, MA** Sep 2025 – Dec 2027
Master's in Computer Science GPA: 3.8/4.0
*Coursework: Data Management and Processing, Algorithms, Programming Design Paradigm*

**Sri Sivasubramaniya Nadar College of Engineering (SSNCE), Chennai, India** Aug 2019 – May 2023
Bachelor's in Information Technology GPA: 8.8/10
*Coursework: Artificial Intelligence, Machine Learning, Big Data Engineering, Database Systems, Data Structures, Probability & Statistics*

## TECHNICAL SKILLS

**Languages:** Python, R, SQL, JavaScript, TypeScript, Linux Bash Scripting, Java, C++
**Frameworks & Libraries:** PyTorch, TensorFlow, Scikit-Learn, FastAPI, React, Next.js, MongoDB, PostgreSQL, Power BI, Tableau, Git, Excel, Airflow
**ML & AI:** Large Language Models (LLMs), Retrieval-Augmented Generation, Deep Learning, Reinforcement Learning, Graph Neural Networks, Natural Language Processing (NLP) (BERT), Transformers, MCP, Supervised Machine Learning, Agentic AI, Data Science, FAISS
**Cloud & Infrastructure:** Azure (AKS, ML, OpenAI), Docker, Kubernetes, Terraform, Prometheus, Grafana
**Certifications:** Microsoft Certified: Azure AI Fundamentals, The Linux Foundation: Certified Kubernetes Administrator, *Data Analytics with Python*

## WORK EXPERIENCE

**Software Engineer (AI/ML Engineer) | Hewlett Packard Enterprise (HPE), Bengaluru, India** Sep 2023 – Aug 2025

- Accelerated incident resolution by **25% for 100+ engineers** via **LangChain** Parent-Document indexing and Map-Reduce
- Reduced hallucinations from **9% to 4%** through **fine-tuning** and few-shot prompting across **15K monthly troubleshooting queries**
- Increased **Top-3 RAG** search accuracy from **81% to 93%** via **Pinecone Cross-Encoder re-ranking** for 20K+ documentation queries
- Decreased agentic logic failures from **12% to 5%** via **LangGraph self-correction loops**, automating error-retry strategies
- Leveraged **Prometheus** & **LangSmith** to monitor latency and token usage, optimizing system performance & reducing monthly **API costs**
- Automated **Azure AKS provisioning** via **Terraform** and **Airflow**, securing OpenAI integrations through private network endpoints
- Facilitated **cross-functional collaboration** to translate AI capabilities into **business requirements**, scaling adoption across 4 global teams

**Software Engineer Intern | Hewlett Packard Enterprise (HPE), Bengaluru, India** Mar 2023 – Aug 2023

- Designed **React** dashboards reducing incident investigation time by **60%** for operations teams **monitoring 10K+ daily transactions**
- Built **Bash** and **Python scripts** within **Agile** workflows to **orchestrate container** builds and **Kubernetes** deployments
- Orchestrated **asynchronous REST API** endpoints with **Redis** caching, **reducing API latency** through strategic **PostgreSQL** indexing
- Collaborated with **product** and **DevOps teams** to define technical specs, benchmark system performance & **streamline CI/CD pipelines**

**AI/ML Core Member | SSNCE Coding Club, Chennai, India** Jun 2022 – Feb 2023

- **Upskilled 50+** students in ML fundamentals, focusing on **data pipelines** and **TensorFlow** architecture through hands-on workshops
- Boosted **student engagement by 40%** by coordinating **AI hackathons** and guiding participants through problem-solving
- Guided **10+ teams** by conducting **code reviews** and implementing Git-flow to optimize model architecture for latency and accuracy

## PROJECTS

**Multi-Agent AI Researcher System | Python, LangGraph, RAG, FastAPI, React**

- Pioneered a **four-stage sequential LangGraph pipeline** for multi-pass writing, automating outline generation, drafting, and critique
- Created a hybrid retrieval system using **Sentence Transformers** and **Cosine Similarity** to filter high-impact research papers
- Implemented **stateful quality gates** to enforce publication standards, triggering **autonomous revision cycles** based on critical feedback

**Custom Claude GitOps MCP Server | Python, MCP, Git, Claude, FastAPI**

- Enabled **autonomous GitOps workflows** for Claude by building **a custom MCP server** using a JSON-RPC communication implementation
- Developed a modular discovery interface using FastAPI to manage and **execute independent, script-based automation task sequences**
- **Automated Git CLI workflows** and **GitHub API integrations** for end-to-end repository provisioning and code deployment

**Indian LegalGPT | Python, ChromaDB, Groq, Mistral-7B, React, FastAPI**

- Engineered a high-precision RAG pipeline using **InLegalBERT** and **ChromaDB**, achieving **90%+ accuracy** on complex statutory queries
- Enabled real-time, **bilingual document** synthesis by integrating **Groq-accelerated Llama-3** inference and **OpenAI Whisper** voice-to-text
- Refined query relevance by implementing a **weighted retrieval**, prioritizing **Penal Code context** to resolve legal ambiguities

**Credit Score Analysis using Machine Learning | Python, Google Colab, TensorFlow, Seaborn**

- Achieved **98.6%** accuracy in multi-class credit risk prediction using a **10-model ensemble** on imbalanced financial datasets
- Resolved class imbalance by synthesizing **5,000+** minority samples via **CTGAN, SMOTE**, and **bootstrapping** to improve recall
- Optimized input matrices via **Seaborn** density mapping and targeted imputation of missing financial metrics