

PRADEEPRAJ PRABHU RAJ

Chicago, Illinois, USA

+1 312 998-3116 | pradeepraj18062002@gmail.com | [LinkedIn](#) | [GitHub](#) | [Website](#)

Education

University of Illinois at Chicago , Chicago, Illinois	Expected May 2026
Master of Science, Computer Science	3.88/4.0
Sri Venkateswara College of Engineering , Chennai, India	Aug 2020 - Jul 2024
Bachelor of Engineering, Computer Science	3.36/4.0

Experience

Sportthon Inc. AI/ML Developer Intern	Jun 2025 – Present
	Chicago, Illinois
<ul style="list-style-type: none">Accelerated performance decisions by 40% by developing n8n AI agents that processed natural language queries to deliver real-time athlete readiness recommendations based on wellness and weather data.Increased user engagement by 30% by developing and deploying LLM-based applications using open-source models integrated into production systems.Enhanced contextual understanding of fitness queries by 15% by fine-tuning LLMs on domain-specific sports and wellness datasets.Built production API endpoints integrating OpenAI and Hugging Face models to enable AI-powered features and streamline data processing.Accelerated model experimentation cycles by 50% by evaluating performance, latency, and deployment trade-offs of state-of-the-art LLM architectures.	
KaaShiv InfoTech Software Engineer Intern	Feb 2022 - Mar 2022
	Chennai, India
<ul style="list-style-type: none">Developed a nutrition tracker app by building 12+ RESTful API endpoints with Node.js and Express.js, reducing manual data handling time by 40%.Implemented JWT-based authentication that secured access for 100+ test users, increasing system reliability and protecting sensitive health data.Optimized database operations by integrating MongoDB (NoSQL) and SQL databases, which improved query performance by 30% and supported real-time tracking of 500+ food entries.	

Skills

- Programming Languages:** Python, JavaScript, Java, C, C++, SQL
- AI & Machine Learning:** PyTorch, Transformers, CNN, RNN, Generative Models, TensorFlow, Scikit-learn, OpenAI APIs, Whisper
- LLM Finetuning & Optimization:** Axolotl, LoRA, QLoRA, BitsAndBytes, LLMCompressor, PEFT
- AI Agentic Frameworks:** LangChain, LangGraph, CrewAI, AutoGen, Letta, Botpress, Microsoft Semantic Kernel
- Cloud & Deployment:** AWS (S3, EC2), Apache Airflow, Google Colab, Jupyter Notebook
- Databases & Vector Stores:** MySQL, PostgreSQL, MongoDB, Astra DB, ChromaDB, Pinecone
- API Development & Integration:** Node.js, Postman, Webhooks, Express.js
- Data Processing & Visualization:** Pandas, Matplotlib, Seaborn, Power BI, MS Excel
- Web Development:** HTML, CSS, React, Angular, Jasmine Framework
- Developer Tools:** Git, GitHub, Visual Studio Code, Selenium, Beautiful Soup, PySpark

Projects

AI Code Generator with Multi-Agent Architecture GitHub	Oct 2025 - Nov 2025
<ul style="list-style-type: none">Reduced web application development time by 70% as measured by time-to-deployment by building an autonomous AI coding system using LangGraph that converts natural language prompts into complete, production-ready web applications.Designed a multi-agent workflow with specialized agents (Planner, Architect, Coder) using LangChain and Pydantic schemas, enabling structured code generation with proper file organization.Orchestrated iterative code generation with state management and conditional edges in LangGraph, allowing the system to loop through multiple implementation tasks autonomously.	
Full Stack PDF RAG Application GitHub Live App	Mar 2025 - May 2025
<ul style="list-style-type: none">Developed a full-stack RAG system using Next.js, OpenAI, and Astra DB to enable semantic question answering over uploaded PDFs, resulting in a 3× faster information retrieval experience compared to manual search.Implemented PDF parsing and chunking pipeline with LangChain and text embedding models, reducing preprocessing time by 50%.Deployed app on Vercel with serverless APIs, ensuring reliable uptime and seamless cross-device chat.	
Fine-Tuning LLaMA 3.2 for IPL Match Strategy Prediction GitHub	Jan 2025 - Feb 2025
<ul style="list-style-type: none">Fine-tuned LLaMA 3.2 using QLoRA and PEFT on IPL match data, achieving a 90% improvement in venue-specific toss and score predictions compared to the base model.Built an efficient training pipeline using Hugging Face Transformers, bitsandbytes, and LoRA adapters for 4-bit quantized fine-tuning on limited GPU resources.	
Bias-Free Music Recommendation System GitHub	Aug 2024 - Nov 2024
<ul style="list-style-type: none">Developed a music recommendation system using CNN, Word2Vec, and Node2Vec, reducing popularity bias by 50% and ensuring fair exposure for lesser-known artists.Reduced cold-start problem impact by utilizing hybrid models that integrated content-based and metadata-based embeddings, achieving accurate recommendations for new users and items, and reducing its impact by 40%.	