

Pradeep Singh Ramola

Buffalo, NY 716-275-5455 pramola@buffalo.edu
linkedin.com/in/pradeep-singh-ramola github.com/pradeep-ramola leetcode.com/u/pradeepramola/

EXPERIENCE

NEC Software Solutions

Jun 2021 – Jul 2025

Software Engineer

- Resolved **30+** high-priority production incidents by reproducing issues, identifying code-level fixes, supporting RCA discussions, collaborating in daily stand-ups, and participating in cross-team KT sessions; reduced mean time to resolution (MTTR) by **20%** and post-release defects by **35%**.
- Optimized **5+** backend batch workflows processing **10,000+** daily operational records, cutting runtimes from **3 days** to **5 hours**; Delivered **20+** API endpoints and UI screens improving system reliability and reducing customer-reported issues by **25%**.
- Contributed **8+** core modules including Activity Requests, Activity Request Authorisation, Administration, Overtime Generation, TOIL Balance Calculation, and Payment Calculation; added new business rules to system function parameters reducing manual configuration effort by **30%**.
- Crafted and optimized web and mobile screens using Ext JS, .NET Framework, and PL/SQL; developed interface tools to export application data to CSV/Excel for **50,000+** officers across **20+** UK police forces, eliminating manual data extraction.
- Led end-to-end development of a Leave Management System serving **20,000+** employees using Angular, .NET Core, SQL Server, and Azure AD, replacing **100%** of manual workflows and reducing processing time from weeks to **2–3 days**.
- Integrated CI/CD pipelines cutting deployment effort by **40%**; led **15+** sprint demos and client calls, conducted **50+** code reviews, and mentored **3** junior engineers on backend architecture and testing practices.
- Led CARM upgrades from version 4.3 to 5.3 and 5.6, architecting **10+** modules, refactoring **40K+** lines of scheduling logic, and resolving **100+** critical bugs, reducing module load times by **30%** and improving cross-team maintainability.

TECHNICAL SKILLS

Languages: Python, Java, C#, JavaScript, C++, TypeScript, SQL

Frameworks & Libraries: .NET Core, Angular, React, Spring Boot, Ext JS, PyTorch, Hugging Face Transformers, OpenCV, Scikit-learn, Pandas, NumPy

AI / ML: DL, RL, Fine-Tuning (LoRA/PEFT), RAG, Agentic AI, NLP, Computer Vision, Transformer Architectures, LangGraph, LangChain, Ollama, Vector Databases, ChromaDB, Optuna, LangSmith, Ragas, Guardrails, Bayesian Optimization

Databases: Microsoft SQL Server, PL/SQL, PostgreSQL, Oracle 12c, MongoDB

Cloud & DevOps: Azure, AWS, Docker, Git, SVN, CI/CD, REST APIs, Microservices, Postman, Jira

Tools & IDEs: Visual Studio, VS Code, Sencha Cmd, Jupyter Notebook, PyCharm, SSMS, PL/SQL Developer

Practices: Agile, Scrum, TDD, Code Review, RCA, KT Sessions, System Design

PROJECTS

TinyGPT: Character-Level GPT | Python | PyTorch | Transformers

- Engineered a **145K-parameter** decoder-only GPT from scratch with causal multi-head self-attention, pre-norm blocks, GELU MLPs, residual connections, positional embeddings, and tied input/output embeddings.
- Trained on TinyShakespeare with AdamW, cosine LR scheduling, gradient clipping, and checkpointing; reduced validation loss from **3.185** to **2.725** and implemented temperature-based autoregressive generation.

AI-Native Delivery Control Tower | Python | FastAPI | Streamlit | Ollama | SQLite | Agentic AI

- Created a local-first agentic software delivery platform that manages engineering task intake, planning, patch generation, validation, review, explainable risk scoring, policy gates, canary simulation, rollback handling, and artifact storage.
- Implemented planner, coder, reviewer, security, release, and incident agents with mock and Ollama-backed modes; exposed REST APIs and dashboards for task/runs tracking, approvals, rollback actions, metrics, and policy visibility.

BERT Fine-Tuning: Hate Speech Classification | Python | PyTorch | Hugging Face | NLP

- Fine-tuned BERTweet on THOS Twitter dataset for 3-class classification; improved Test Macro F1 from **0.6486** to **0.6848** using domain adaptation, focal loss, and cosine LR scheduling.
- Applied LoRA to cut trainable parameters from **109M** to **297K** (0.27%), retaining **98.3%** of full fine-tuning performance; ran fairness analysis and attention-based explainability across target groups.

EDUCATION

University at Buffalo, SUNY

Aug 2025 – Dec 2026

M.S. Computer Science GPA: 3.77 / 4.0

Courses: Algorithms, Machine Learning, Deep Learning, Data Intensive Computing

KJ Somaiya Institute of Engineering and Information Technology

Jun 2017 – Jul 2021

B.E. Information Technology GPA: 4 / 4.0

PUBLICATIONS

Product Analyst Advisor: NLP and sentiment analysis for mobile product research. *IRJET*, Vol. 10, Issue 10, Oct 2023

HACKATHONS

- IBM Hackathon (Apr 2026):** Built Victor AI, an agentic AI assistant for international students using IBM watsonx Orchestrate, multi-agent workflows, email/calendar tools, and web scraping across 50+ universities.
- UB Hacking (Nov 2025):** Designed BrailleSense, a real-time speech/text/video to Braille translator app for accessibility.