BUVANI PAI

+1 (872) 239-1418 | buvanipai@gmail.com | Portfolio | LinkedIn | Github

EDUCATION

Northwestern University, Evanston, Illinois — Sep 2023 – Dec 2024
Master of Science in Computer Science (Focus: Artificial Intelligence & Machine Learning) — CGPA: 3.71/4.0

Rajiv Gandhi Institute of Technology, Mumbai University, India — Aug 2019 – May 2023
Bachelor of Engineering in Computer Engineering — CGPA: 3.64/4.0

SKILLS

- **GenAI & LLM:** LangChain, FastAPI, ChromaDB, sentence-transformers, RAG Pipelines, SpaCy, NLTK
- **AI/ML Frameworks:** PyTorch (DDP), TensorFlow, Scikit-learn, NetworkX, Google OR-Tools
- **Programming & Dev:** Python, SQL, JavaScript, Java, Ruby, React.js, Node.js, Django
- **Cloud & Infrastructure:** Docker, Google Cloud Run, AWS, GCP, CI/CD, Git, PostgreSQL
- **Achievements:** 2nd Place, Google Cloud Hackathon (Agentic AI Product Launch); McKinsey Forward Program (Dec 2025)

WORK EXPERIENCE

AI/ML Engineer | Bhuvi IT, Schaumburg, IL — Jan 2026 – Present

- Designing a real-time voice AI pipeline using FastAPI, WebSockets and LLM APIs for speech transcription and intent classification, achieving **sub-800ms** end-to-end latency across concurrent calls.
- Architecting a **RAG** system with **ChromaDB** and sentence-transformers to retrieve domain-grounded context at inference time, eliminating hallucinations on policy-sensitive queries.
- Containerizing and deploying a serverless application on Google Cloud Run using Docker multi-stage builds, with CI/CD pipelines for automated deployment and auto-scaling to handle concurrent load.

AI/ML Engineer | Happy World Foundation, Inc., Evanston, IL — Aug 2025 – Dec 2025

- Developed data aggregation pipeline with Python, PostgreSQL, and **REST APIs** to process requests from web forms, Instagram API, and SMTP email parsing, ensuring same-week response delivery across distributed volunteer networks.
- Created automated pairing algorithm using Python, NetworkX graph clustering, and Google OR-Tools to match volunteers and teachers across 40+ countries, reducing manual scheduling effort by **70%**.

Machine Learning Engineer | Argonne National Laboratory, Lemont, IL — Jan 2025 – May 2025

- Deployed distributed 8-GPU training pipeline using PyTorch DDP and NCCL with gradient accumulation to process 2TB of astronomical data, reducing training runtime from **one week to 20 hours**.
- Optimized data ingestion through memory-mapped arrays and **PyTorch Distributed Samplers** for 300k+ high-resolution images and spectra, maximizing GPU throughput and eliminating compute starvation.
- Built data quality pipeline using Pandas and Matplotlib histogram analysis to detect and prune 13% of skewed observations, eliminating gradient explosions and **stabilizing model convergence**.
- Implemented multi-modal preprocessing workflows using NumPy normalization and feature scaling across image and spectral data, achieving 2x faster convergence compared to unnormalized baselines.
- Fine-tuned conditional diffusion models using PyTorch with **cosine annealing** and **warmup scheduling** to generate synthetic galaxy images, reducing visual artifacts and improving structural coherence in generated images.

Machine Learning Engineer | Mascot Dynamics, Mumbai, MH — Sep 2021 – Mar 2022

- Engineered real-time predictive maintenance using Isolation Forest and XGBoost to achieve an **0.85** F1-score with SMOTE.
- Developed FFT-based feature extraction from vibration signals, improving sensitivity by **20%** to prevent equipment downtime.
- Analyzed thermal-pressure correlations to identify failure patterns, reducing false alarms by **35%** for key industrial assets.

PROJECTS

Agentic CRM Entity Resolution Pipeline — Nov 2025 – Dec 2025

- Engineered graph-based deduplication system using **LangChain**, LLMs, and NetworkX BFS clustering to resolve ambiguous contact identities, achieving **0.96** F1-score with few-shot prompting and hard negative constraints.
- Implemented hybrid blocking strategy to reduce entity matching complexity from $O(N^2)$ to near-linear, decreasing database redundancy by **28%** while preventing transitive over-merging.
- Designed batch inference pipeline with conflict detection maintaining **95%** confidence across semantic matching, ensuring audit trail integrity for record linking.

Multi-Agent Text-to-SQL System — Apr 2025 – May 2025

- Built three-agent system using **LangChain** and LLMs to translate natural language into executable SQL across 11 schemas, improving query validity by **30%** and efficiency by **6%**.
- Established metadata-driven validation logic in Python to detect and correct syntax errors, increasing column-match accuracy by **12%** and execution success rate by **5%**.
- Constructed iterative testing framework to evaluate query patterns across diverse schemas, improving query execution reliability through systematic prompt engineering and parameter optimization.