Jay Narendrabhai Joshi

(408)-207-5280 | jayjoshius810199@gmail.com | https://www.linkedin.com/in/jay-joshi-b08232200/ | San Jose, CA, USA

Education

San Jose State University — GPA: 3.56/4.0 San Jose, CA, USA
Masters of Science in Applied Data Intelligence Jan 2024 – Dec 2025
Coursework: Mathematics, Database Systems, Machine Learning, Big Data, Deep Learning, Generative Models, Distributed Systems

Birla Vishvakarma Mahavidyalaya Engineering College — GPA: 3.4/4.0 Anand, India
Bachelor in Information Technology Aug 2019 – May 2023
Coursework: Data Structures & Algorithms, Computer Networks, Operating Systems, Database Management Systems, Python, AI

Technical Skills

Languages: C++, Python, Java, C, JavaScript
Cloud & Databases: AWS (Amazon Web Services), GCP (Google Cloud Platform), Azure, PostgreSQL, MySQL, No-SQL Databases, MongoDB, Vector DB
Tools & Technologies: RESTful APIs, Git, GitLab, Kubernetes, Docker, Linux, Node.js, React, Cursor AI, PyTorch, vLLM, TensorFlow, NumPy, Pandas, JSON, HTML, Airflow, Inference Engines
AI/ML: Artificial Intelligence, Computer Vision, Natural Language Processing, Transformers, CNN, BERT, GPT, ELMo, Spacy, Graph Neural Network, Deep Learning, Engineering, Fine-Tuning, Quantization, QLoRA, Model Context Protocol, Deep Q-Learning, Distributed Training, Vector DB, RAG, LLM
Miscellaneous: System Design, Agile Development, SCRUM, Object-oriented Programming, Data Structures and Algorithms, Probability, Statistical modeling, Hugging Face

Experience

Software Engineer Intern | Launch Pad Jan 2023 – Nov 2023
[AWS, Python, PostgreSQL, Bash, Git, Docker, Airflow]

- Reduced data retrieval time by 30% by pipeline processing speed by setup of automated data pipeline between company's in-house application and Bloomberg Terminal executing **Airflow** for job scheduling, **AWS EKS** for container orchestration, and **Python**.
- Improved log monitoring efficiency by **40%** observed by operational oversight metrics by creating dashboard application with **Machine Learning**
- Reduced manual intervention with improved system reliability across applications by **automated job scheduling** using **Airflow**

Software Engineer Intern| Tatvasoft Jun 2022 – Jul 2022
[Vue.js, .Net, Node.js, HTML, CSS, JavaScript, MongoDB, AWS, WebSocket]

- Contributed to design and development of a book-selling e-commerce website utilizing **Vue.js** for front-end, **.NET** and **Node.js** for back-end, **MongoDB** for database management, and **REST APIs** for seamless data integration.
- Reached **80%** increased successful transactions calculated by transaction completion rate by streamlining the payment gateway and reducing checkout steps.
- Improvement in page load time by **20%** and customer satisfaction by utilizing **WebSocket** to optimize server requests.

Projects

Multi-Agent Collaboration System for Software Writing [*Crew AI, RAG, Google ADK, GPT, Docker, GCP, Python, PyTorch, vLLM*]

- Accelerated project completion by **70%** metered by development time reduction by AI-based Multi-Agent Collaboration systems with **AI Agents and LLMs**
- Increased agent accuracy by 85% by solution quality metrics by leveraging **Agentic frameworks** with **RAG-based fine-tuning** and custom **prompt engineering**.
- Accomplished software generation in seconds by **time reduction** compared to **human development** by deploying **multi-agent AI system**
- Implemented live **inference deployment** measured by **production readiness** by deploying fine-tuned LLMs using **Hugging Face**, **vLLM** and **RunPod**.

Stock Market Prediction [*Python, PyTorch, React.js, Node.js, Fast API, WebSocket, GCP*]

- Archived **97%** stock prediction accuracy seen by model performance on **2000+** stocks by developing **LSTM/XGBoost** models using **PyTorch**.
- Deployed real-time analytics platform with **millisecond** latency with **WebSocket** response times by designing frontend utilizing **React.js** with **JavaScript** and **FAST API** Python backend for technical analysis, forecasting, and news integration on **GCP**.

Study Buddy- Berkeley Cal Hacks (Hackathon) [*React.js, GPT, Hume AI, LangChain, Ollama, RAG, Google Gemma, ChromeDB*]

- *Accomplished over **90%** socratic response relevance measured by tutor interaction metrics by integrating **Hume EVI** and fine-tuning **Gemma 2B** with **4-bit quantization using QLoRA**.*
- *Attained **70%** reduction in out-of-context errors measured by error metrics by implementing **RAG pipeline** with custom **Model Context Protocol (MCP)** for vectorized course material injection*