PARTH VAIBHAV PANSE

Tempe, AZ | 480-953-2424|ppanse1@asu.edu|parthpanse|

SUMMARY

Software engineer with a Master's in Computer Science and hands-on experience building full-stack web applications, scalable AI systems, and machine learning models. I've contributed to both startup teams and academic research, developing solutions that improve performance, automation, and user experience. Skilled in Python, React, AWS, and ML frameworks like PyTorch and TensorFlow, I enjoy solving real-world problems with clean, efficient code. Whether it's deploying AI tools or creating smart, user-friendly applications, I bring creativity, curiosity, and a strong focus on collaboration to every project.

TECHNICAL SKILLS

- **Languages:** Python, C, C++, Java, Go, HTML, CSS, JavaScript
- **Frameworks:** Django, ReactJS, Angular, NodeJS, ExpressJS, Selenium, Flask, .NET
- **Machine Learning:** Scikit-Learn, PyTorch, Keras, TensorFlow, Pandas, Numpy
- **DevOps & Cloud:** Git, Jenkins, Docker, Kubernetes, CI/CD, GCP, AWS, API Gateways
- **Databases:** PostgreSQL, MySQL, MongoDB (SQL/NoSQL)

WORK EXPERIENCE

Software Engineer| *Sedai Inc., Remote, USA* **June 2025 – Present**

- Currently architecting and implementing a scalable, on-premise Agentic Evaluation Platform using DeepEval to enable automated, repeatable testing of LLM agents across critical workflows.
- Designing automated evaluation workflows and integrating custom metrics to enable safe deployment of multi-step enterprise AI agents.

Software Engineer Intern| *Oytie Pvt. Ltd, Pune, India* **February 2022 – May 2022**

- Led development of a full-stack CRM web application using **Django, React, and PostgreSQL**, contributing to a 10% increase in user retention by streamlining client interactions and improving UX flow.
- Optimized **RESTful APIs using Django REST Framework** by profiling and refactoring bottlenecks, reducing average response times by 40% and enabling the system to scale to 10,000+ daily user requests.
- Implemented containerized development and deployment workflows using **Docker and GitHub Actions for CI/CD**, reducing deployment time by 30% and eliminating 90% of manual errors.
- Designed and executed automated unit and **integration tests using Pytest**, boosting software reliability by 35% and maintaining 99.9% uptime post-deployment.
- Drove **Agile practices** by facilitating sprint planning and backlog refinement sessions on Jira, resulting in a 25% increase in team productivity and a 20% acceleration in release cycles.
- Collaborated in peer code reviews and pair programming, improving code quality and reducing post-deployment bugs by 25%.

PROJECT EXPERIENCE

Scalable Face Recognition System on AWS **January 2025 – May 2025**

- Prototyped a real-time simulation system on **AWS, leveraging EC2 and SQS** to build a fault-tolerant and scalable data processing workflow
- Owned the performance optimization of the system, troubleshooting bottlenecks to cut latency from 3s to 1.8s and ensure reliability for high-throughput analysis.
- Automated deployment and monitoring using **CloudWatch** and infrastructure-as-code, enabling rapid iteration and robust system observability.

Automated Brochure Generator with GPT-4 and Web Scraping **January 2025 – March 2025**

- Built an end-to-end web application to automate brochure creation using GPT-4 for content generation and **Python-based web scraping**, cutting manual design time by 80%.
- Architected the solution using a microservices-based stack (**React frontend, Flask backend**), showcasing cross-functional experience in **GenAI, front-end design, and API integration**.
- Set up automated checks and tests to make sure the brochure content was accurate, reaching 95% accuracy and helping businesses quickly create brochures without mistakes or the need for a design team.

Predictive Pricing Assistant for E-commerce **March 2025 – May 2025**

- Collected and cleaned a custom dataset of product descriptions and prices, making sure the data was accurate and easy for the model to learn from.
- Fine-tuned **OpenAI's GPT-4o-mini using LoRA**, so the model could predict prices based on short product descriptions.
- Tracked training progress and tested model accuracy using Weights & Biases, helping to compare results and spot areas for improvement.
- Set up a clear, repeatable process with version control and experiment logging to make sure others can run and understand the project results.

RAG-Powered Medical Documentation Chat System **March 2025 – May 2025**

- Built an interactive chat system that helps users retrieve precise answers from complex medical records, leveraging **Retrieval-Augmented Generation (RAG)** and a user-friendly web interface.
- Developed and fine-tuned core components—including document **chunking, OpenAI-based embeddings, and a ChromaDB vector retriever**—experimenting with chunk sizes and retrieval settings to maximize answer accuracy.
- Validated the system's performance by testing on real clinical questions, achieving 100% retrieval precision, and ensuring that every answer could be traced back to the correct source document.

Image Denoising Using Convolutional Neural Networks **August 2023 – December 2023**

- Demonstrated a strong background in computer vision by developing a **CNN-based (U-Net) model** in PyTorch to process and enhance spatial data.
- Optimized the deep learning model through iterative training cycles, achieving stable convergence at **0.0154 MAE** loss on the CIFAR-10 dataset.
- Highlighted practical use cases such as improving medical image clarity and enhancing object recognition in autonomous systems, showcasing the model's value beyond academic evaluation.

EDUCATION

Master of Science, Computer Science **August 2023 – May 2025**
Arizona State University **Tempe, AZ**
Relevant Coursework: Cloud Computing, Statistical Machine Learning, Foundation of Algorithms, Statistical Learning Theory, Software Verification/Validation/Testing

Bachelor of Engineering, Computer Engineering **August 2019 – June 2023**
University of Pune **Pune, India**
Relevant Coursework: Data Structures and Algorithms, High Performance Computing, Theory of Computation, Machine Learning, Natural Language Processing