Akansh Sharma

☎ (219) 368-2464 ✉ sharm995@pnw.edu github.com/akansh194 linkedin.com/in/sharma-akansh

Education

Purdue University Northwest, Hammond, IN Aug 2024 – Present
Master of Science in Computer Science
SRM Institute of Technology and Science, Delhi, India Jul 2018 – Aug 2022
Bachelor of Technology in Computer Science CGPA: 8.44/10

Technical Skills

Tools: Pytorch, TensorFlow, Scikit-learn, Keras, LangChain, Jupyter, MLflow, Datadog, Vercel, cPanel
Languages: Julia, Java, Python, JavaScript, React, flask, NodeJS, ExpressJS
Databases: SQLite, MySQL, PostgreSQL, MongoDB
Technologies: Machine learning, Large language Models, Docker, Kubernetes, VM Management, Virtual Networks, Jenkins, React, CI/CD pipelines, Token Management, Git

Certifications

Machine Learning Specialization – Stanford University
Deep Learning & NLP Specializations – DeepLearning.AI
Microsoft Azure AI Fundamentals (AI-900)
Microsoft Azure Developer Associate (AZ-204)

Work Experience

Insight, India Jul 2022 – Jul 2024
Software Engineer

- Deployed and managed machine learning models on cloud infrastructure using AWS SageMaker and Azure ML, enabling scalable AI workloads with 99.9% uptime
- Implemented AI-driven monitoring and anomaly detection systems using CloudWatch and custom ML algorithms, reducing incident response time by 35%
- Integrated serverless AI inference pipelines using AWS Lambda and API Gateway, processing 50K+ predictions daily with sub-second latency
- Architected a cloud backup and disaster recovery strategy that cut recovery time by 60% and boosted successful data restores by 50%, strengthening business continuity for production workloads
- Evaluated and piloted emerging cloud platforms and services, driving the adoption of a new provider that reduced monthly infrastructure spend by 30% while improving system performance by 25%.
- Diagnosed and resolved complex IaaS issues across compute, storage, and networking, delivering a 40% reduction in unplanned downtime and measurably improving overall system stability
- Optimized IIS application pool configurations to proactively recycle resources for high-traffic clients, preventing overload and enhancing responsiveness and reliability of production applications and websites

Academic & Personal Projects

End-to-End RAG Knowledge Base Platform (Azure, Python) Aug 2025 – Dec 2025

- Designed and implemented an end-to-end Retrieval-Augmented Generation (RAG) platform on Azure using Python, Azure OpenAI, Azure AI Search, and Azure Blob Storage for multi-document knowledge management and chat-based querying
- Developed a FastAPI backend that orchestrates query embedding, hybrid retrieval, and Azure OpenAI calls to generate grounded answers, enforcing strict ”answer from context” behavior and robust error handling for production use
- Implemented citation-aware prompting and response post-processing to attach chunk-level source IDs, powering UI features like inline citation markers and side-panel snippet highlighting directly mapped to original documents

Backdoor Attack Analysis on LLM Models Aug 2024 – Dec 2024

- Implemented and evaluated backdoor attacks on text classification models to study vulnerabilities in modern NLP pipelines
- Trained baseline Transformer-based classifiers using Python, PyTorch, and Hugging Face on labeled text datasets
- Injected poisoned samples with trigger phrases into the training data and quantified attack success rate and impact on clean accuracy
- Compared multiple defense strategies (data filtering, activation clustering, fine-tuning) and measured their effectiveness under different threat settings
- Developed reproducible experiments and visualizations to communicate findings on model robustness and AI safety risks
- Documented methodology and results in an IEEE-style report and presented key insights to faculty and peers
- GitHub Repository: Link

Cloud Migration and Optimization for E-Commerce Platform Jul 2022 – Jan 2023

- Containerized application services with Docker and orchestrated workloads using Amazon ECS, improving scalability, deployment speed, and system uptime
- Implemented Infrastructure as Code (IaC) using Terraform to automate provisioning and ensure reliable, version-controlled deployments
- Architected and deployed a scalable cloud infrastructure on AWS, leveraging EC2 for compute, RDS for relational database management, and S3 for fault-tolerant object storage