Shivam Kumar

+91 9199932637 | shivam.sk2003@gmail.com | linkedin.com/in/shivam-kumar2003 | github.com/shivam2003-dev
Delhi, India

Education

BITS Pilani 2025 - 2027
M.Tech in Artificial Intelligence and Machine Learning *Ongoing*

Chandigarh College of Engineering and Technology (CCET), Chandigarh 2020 - 2024
B.E in Computer Science and Engineering *CGPA: 8.17/10*

Guru Gobind Singh Public School (GGPS), Bokaro Steel City 2018 - 2020
High School with Mathematics *96.6%*

Work Experience

Site Reliability Engineer II/ MLOps Engineer II June 2025 – Present
Kimbal Technologies, Delhi, India

- Trained and deployed ML models for electricity demand and supply prediction using time series forecasting, improving grid load balancing accuracy by 25%
- Developed and productionized AI-powered theft detection models using anomaly detection algorithms, processing 1M+ smart meter readings daily
- Built and deployed DTR (Distribution Transformer) mapping models using computer vision and geospatial analysis for utility infrastructure optimization
- Orchestrated ML model deployment pipelines on Kubernetes with automated scaling, versioning, and A/B testing capabilities
- Optimized PostgreSQL performance through configuration tuning, reducing CPU utilization by 35% and query response time by 40%
- Deployed VPC endpoints for AWS S3 and services, cutting data transfer costs by $5,000/month and eliminating NAT Gateway charges
- Resolved IP address exhaustion in production Kubernetes clusters by deploying Calico CNI plugin
- Configured local DNS caching in Kubernetes nodes, reducing CoreDNS CPU utilization from 700% to 1%
- Established multi-level alerting in Zabbix with email, Microsoft Teams, and PagerDuty integration
- Automated PostgreSQL backups using pgBackRest with Ansible, enabling point-in-time recovery in S3
- Built AWS Lambda functions triggered by EventBridge, reducing manual intervention by 60%
- Integrated AWS SNS with internal services for real-time alerts and CloudWatch metrics tracking

Site Reliability Engineer June 2024 – June 2025
Kimbal Technologies, Delhi, India

- Orchestrated Kubernetes, Docker, Helm, and ArgoCD for scalable deployments with GitOps-driven workflows
- Engineered Terraform modules to provision and manage cloud infrastructure as code, improving deployment efficiency
- Built and maintained CI/CD pipelines for 10+ microservices using GitHub Actions, enabling automated deployments
- Leveraged AWS services including EKS, EC2, S3, VPC, WAF, Inspector, and firewall configurations
- Automated monitoring and alerting using Zabbix integrated with internal systems for real-time visibility
- Administered HashiCorp Vault for secrets management and Nexus Repository for artifact distribution
- Created Ansible playbooks for configuration management of EMQX, OpenVPN, and PostgreSQL clusters
- Deployed Wazuh for security monitoring across 10+ environments and SonarQube for code quality analysis
- Established EMQX production cluster with 3 nodes, auto-discovery, and MQTT communication
- Automated 50+ recurring tasks including system maintenance, provisioning, and operational workflows
- Managed databases (MSSQL, PostgreSQL, Cassandra, Neo4j) focusing on performance tuning and backup strategies
- Led Disaster Recovery strategy and executed 30+ successful DC-DR and DR-DC drills for utility clients

Site Reliability Engineer Intern January 2024 – June 2024
Kimbal Technologies, Delhi, India

- Optimized CI/CD pipelines using GitHub Actions, reducing run times by 60–70%
- Gained hands-on experience with Terraform, Kubernetes, Docker, and cloud infrastructure

Technical Skills

Programming Languages: Python (Primary), Java, Bash, PowerShell, Go, C++, JavaScript, YAML
Machine Learning & AI: PyTorch, TensorFlow, scikit-learn, Transformers, Pandas, NumPy, Matplotlib, Seaborn
Large Language Models: GPT, LLaMA, Falcon, Mistral, Hugging Face Transformers, LangChain
LLM Techniques: Prompt Engineering, Fine-tuning, Embeddings, Vector Databases (FAISS, Pinecone), RAG
AI Applications: Natural Language Processing (NLP), Computer Vision, Recommendation Systems, Time Series Analysis
ML Algorithms: Regression, Classification, Clustering, Anomaly Detection, Neural Networks, Deep Learning
Data Processing: SQL, Data Cleaning, Preprocessing, Feature Engineering, ETL Pipelines, Large Dataset Handling
Cloud Platforms: AWS (EKS, EC2, S3, VPC, SNS, Lambda, EventBridge, CloudWatch, WAF, Inspector), Azure, GCP
Container & Orchestration: Kubernetes, Docker, Helm, ArgoCD, Production Cluster Management
Infrastructure as Code: Terraform, Ansible, CloudFormation, Configuration Management
CI/CD: GitHub Actions, CodePipeline, GitOps Workflows, Automated Deployments
Monitoring & Security: Zabbix, Wazuh, PagerDuty, Prometheus, Grafana, HashiCorp Vault, SonarQube
Databases: PostgreSQL, MSSQL, Cassandra, Neo4j, pgBackRest, Performance Tuning
Networking: VPC, DNS, Calico CNI, MQTT, OpenVPN, TCP/IP, Firewall, Hybrid Infrastructure
Message Brokers: EMQX, MQTT Protocol, 3-Node Cluster Configuration
Tools & Platforms: Git, GitHub, Nexus Repository, Linux (Ubuntu, RHEL), Windows Server
Disaster Recovery: DR Strategy, Backup Solutions, High Availability, 30+ DC-DR Drill Executions

Extracurricular Activities

AI/ML/GenAI Guild Leader [Date Range]
Kimbal Technologies, Delhi, India

- Lead internal AI/ML/GenAI guild, teaching and mentoring 20+ colleagues on machine learning fundamentals, LLMs, and generative AI
- Conducted weekly knowledge sessions on topics including prompt engineering, fine-tuning, vector databases, and RAG architectures
- Developed training materials and hands-on workshops on PyTorch, TensorFlow, and LLM frameworks (GPT, LLaMA, Mistral)
- Mentored team members on building AI-powered applications, from POC development to production deployment
- Facilitated collaborative learning environment, enabling colleagues to implement AI solutions in their respective projects

Vice-President July 2022 – 2024
CCET Student Council

- Managed technical and cultural clubs, coordinated campus-wide events with 500+ participants

Database Head January 2023 – 2024
Training and Placement Cell, CCET

- Maintained placement database for 200+ students and optimized reporting workflows

Projects

LLM-Powered Document Q&A Assistant | *Python, GPT-4, LangChain, FAISS, FastAPI*

- Built intelligent Q&A system using GPT-4 API with RAG architecture for document-based question answering
- Implemented vector embeddings with FAISS for semantic search and document retrieval from large corpora
- Designed prompt engineering strategies to improve answer accuracy and reduce hallucinations by 40%
- Automated document processing pipeline for PDF parsing, chunking, and embedding generation

Fine-tuned LLM for Domain-Specific Tasks | *Python, LLaMA, PyTorch, Transformers, Hugging Face*

- Fine-tuned LLaMA model on domain-specific dataset using parameter-efficient fine-tuning (LoRA) techniques
- Preprocessed and cleaned 10K+ text samples, implemented data augmentation strategies for training
- Achieved 15% improvement in task-specific accuracy compared to base model through hyperparameter tuning
- Deployed fine-tuned model using Hugging Face Transformers with optimized inference pipeline

AI-Powered Electricity Theft Detection | *Python, PyTorch, scikit-learn, Time Series, Anomaly Detection*

- Developed ML models using regression, classification, and clustering algorithms for smart meter data analysis

- Engineered features from time series data and implemented anomaly detection techniques for pattern recognition
- Achieved 92% accuracy in identifying energy theft and tampering patterns using ensemble methods
- Preprocessed and cleaned large-scale datasets (100K+ records) with SQL for efficient data handling

Document Summarization System | *Python, GPT-3.5, OpenAI API, NLP, Text Processing*

- Built automated document summarization tool using GPT-3.5 API for information retrieval and extraction
- Implemented chunking strategies for processing long documents and maintaining context coherence
- Designed workflow automation using LLM APIs to process 100+ documents daily with structured output

Mesh Network Analyzer | *Python, FastAPI, React, Firebase, tshark*

- Engineered platform to detect mesh networking apps with real-time packet capture capabilities
- Designed FastAPI backend and React frontend for network interaction detection and data visualization

DevOps HomeLab | *Kubernetes, Docker, Terraform, Ansible, ArgoCD*

- Constructed home lab infrastructure with orchestration, CI/CD automation, and monitoring (Prometheus/Grafana)
- Deployed GitOps workflows using ArgoCD for declarative infrastructure management

Performance Profiling Tool | *Python, TCP/IP*

- Created tool for analyzing system bottlenecks and network traffic optimization
- Reduced application latency by 45% through performance hotspot identification

Certifications and Achievements

Certified Cloud Security Practitioner – AWS (CCSP-AWS)
TryHackMe - Top 1% [0xA][WIZARD] - Elite cybersecurity ranking
Fundamentals of Machine Learning for Software Engineers - Educative
Key Achievements:

- Automated 50-100+ operational tasks, improving team productivity by 70%
- Reduced CI/CD pipeline runtime by 60-70% through optimization
- Successfully executed 30+ disaster recovery drills for utility clients
- Deployed security monitoring (Wazuh) across 10+ production environments
- Built and deployed 3+ LLM-powered applications with production-ready workflows