Ripunjay Singh

singhripunjay09@gmail.com | +91-9934820613 | Github | LinkedIn

Professional Summary

Experienced Software Engineer specializing in custom LLM development, multi-agent AI systems, and production-scale infrastructure automation. Expert in training and deploying domain-specific language models using LoRA fine-tuning, implementing distributed processing pipelines with advanced caching mechanisms, and orchestrating containerized microservices. Proven track record of building scalable ML pipelines that handle 100+ concurrent requests with comprehensive observability frameworks and custom model serving infrastructure.

Skills Summary

- **Languages:** Python, C++, C, Java, JavaScript, SQL, Bash, Dart, Swift
- **LLM & AI Frameworks:** LangChain, LangGraph, LlamaIndex, vLLM, MLflow, LMCache, KServe, PyTorch, MLX-VLM, LoRA Fine-tuning
- **Cloud & Infrastructure:** AWS, Docker, Kubernetes, Terraform, Jenkins, Nginx, Supervisord, Redis, Celery, Flower
- **Monitoring & Observability:** Prometheus, Grafana, Custom Latency Diagnostics, Real-time Performance Analytics
- **Frameworks & APIs:** FastAPI, SpringBoot, Django, Flask, SwiftUI, Flutter, React, Next.js
- **Databases:** Neo4j, PostgreSQL, MySQL, Redis, MongoDB, Qdrant Vector DB, SQLite

Education

Bennett University — Greater Noida, Uttar Pradesh, India
Bachelor of Technology in Computer Science — Sept 2022 – July 2026
CGPA: 8.54 (till 6th semester)

St. Xavier's School — New Delhi, India
High School — April 2021 – July 2022
CGPA: 8.7

Experience

Visa2fly — New Delhi, India
AI Engineer / Backend Developer (AI/ML Focus) — *Dec 2024 – Present*

- **Graph-based Multi-Agent Conversational AI:** Developed and deployed a conversational AI system integrating multiple LLM-driven agents with a Neo4j-powered knowledge graph. Employed LangChain and LangGraph frameworks to orchestrate dynamic, context-rich interactions, significantly enhancing response accuracy and speed.
- **Custom LLM Development & Deployment:** Engineered domain-specific language models using LoRA fine-tuning techniques on travel and visa documentation datasets. Implemented custom model serving pipeline using vLLM and KServe, achieving 40% faster inference times compared to standard deployments with LMCache optimization.
- **Autonomous Document Validation System:** Architected production-scale LangGraph-powered multi-agent system that autonomously validates complex visa documents with 95% accuracy. Orchestrated 7 specialized agents (document processor, normalizer, validator, repair, reporter) handling 100+ concurrent validation requests daily.

- **Model Context Protocol (MCP) Server Implementation:** Engineered MCP server module to standardize context injection for multi-agent workflows. This solution enables dynamic external data retrieval and synchronization between AI agents, bolstering overall system reliability and performance.
- **Advanced ML Infrastructure Pipeline:** Designed and implemented end-to-end ML pipeline for custom models and microservices utilizing Docker for containerization, Kubernetes for orchestration and auto-scaling, Jenkins for continuous integration and deployment, and Terraform for infrastructure as code on AWS. Pipeline automated testing, container builds, deployment, and real-time monitoring with Prometheus and Grafana.
- **Knowledge Graph RAG System:** Architected Retrieval-Augmented Generation pipeline interfacing with Neo4j database storing 10,000+ validation rules and document relationships. Implemented graph-based context retrieval delivering context-aware responses for travel visa assistant, reducing error rates and ensuring regulatory compliance.
- **Production Deployment & Observability:** Orchestrated multi-service deployment using Nginx load balancing, Supervisord process management, and FastAPI microservices. Implemented comprehensive monitoring with Flower for Celery task monitoring, custom performance analytics, and automated error reporting generating 150+ detailed validation reports daily.

Visa2fly New Delhi, India
Backend Engineer – SpringBoot *Dec 2023 – Nov 2024*

- Spearheaded development of core microservices using SpringBoot, delivering refund management system processing 500+ transactions daily and internal analytics dashboard serving 50+ business users.
- Architected robust RESTful APIs with MongoDB and MySQL integration, implementing secure JWT authentication and role-based access control supporting 1,000+ active users.
- Automated CI/CD pipelines using Jenkins and Docker, reducing deployment time by 60% and establishing automated testing frameworks achieving 85% code coverage.
- Integrated early AI capabilities including conversational chatbots and automated document processing, laying foundation for advanced ML implementation.

Visa2fly Remote & On-site
SpringBoot Intern *June 2023 – Nov 2023*

- Contributed to development and optimization of SpringBoot microservices, implementing new features and resolving critical bugs in agile development environment.
- Collaborated with cross-functional teams to deliver customer-facing features, gaining foundational experience in enterprise software development and API design.

Software Development Projects

LangGraph Document Validation Agent (Advanced ML Pipeline)
(LangGraph, FastAPI, Neo4j, Redis, Docker, MLX-VLM, vLLM)

- Developed autonomous multi-agent system for intelligent document validation using state machine architecture with LangGraph orchestration.
- Implemented distributed processing with Celery workers, Redis backend, and comprehensive observability monitoring using custom metrics.
- Achieved 95% validation accuracy with real-time processing capabilities and extensive audit trail generation.

Custom LLM Training & Serving Infrastructure
(LoRA, vLLM, KServe, MLflow, LMCache, Kubernetes)

- Designed and implemented custom LLM fine-tuning pipeline using LoRA techniques for domain-specific applications.

- Built production-ready model serving infrastructure with vLLM and KServe, implementing intelligent caching with LMCache.
- Achieved 70% cost reduction in inference costs through optimized batching and resource management strategies.

AWS Multi-Region Infrastructure Automation
(Terraform, Jenkins, AWS, Kubernetes)

- Automated deployment of highly available AWS infrastructure across multiple regions with auto-scaling capabilities.
- Implemented Infrastructure as Code (IaC) with Terraform, managing load balancing, DNS routing, and secure network configurations.
- Established monitoring and alerting using Prometheus and Grafana for comprehensive system observability.

Image Enhancement & Computer Vision System
(Python, GANs, PyTorch, OpenCV)

- Developed advanced image enhancement system using Generative Adversarial Networks (GANs) for deblurring and glare removal.
- Implemented custom neural network architectures achieving 85% improvement in image clarity metrics.

Urban Workers Platform
(Django, Firebase, Google Maps API, PostgreSQL)

- Architected scalable platform connecting 500+ roadside workers with employment opportunities, implementing geolocation-based matching.
- Integrated real-time notification system and secure payment processing, achieving 90% user satisfaction rate.

Gitreverb Developer Social Platform
(Flask, SQLite, Firebase, Android-Java, iOS-SwiftUI)

- Built comprehensive social media platform for developers with portfolio showcasing and mentorship features.
- Implemented GitHub OAuth integration and real-time messaging system supporting 200+ active developers.

FRIDAY Virtual Assistant
(Python, Selenium, Automation)

- Developed AI-powered assistant to automate system tasks, manage schedules, and perform web searches via voice commands.
- Implemented natural language processing for command interpretation and task execution automation.

Certifications

- **AWS Certified Cloud Practitioner**
  Verify Credential

Research Publications

- **Bridging the Gap: Integrating DevOps Practices into Educational Curricula for Seamless Industry Transition**
- **A Shared Memory Framework for Coherent and Context-Aware Multimodal AI** – Focus on distributed AI systems and memory optimization

Honors and Achievements

- **MLH Hackathon Winner** (Asia-Pacific Region, Aug 2022) – Developed innovative AI-powered solution
- **Hacktoberfest Challenge Completion** (Nov 2022) – Contributed to open-source ML and DevOps projects
- **Advanced ML Implementation Recognition** – Developed production-scale multi-agent validation system handling 100+ daily processes

Leadership & Volunteer Experience

Microsoft Learn Student Ambassador (Sept 2023 – Present)

- Conducted 15+ technical workshops on cloud computing, AI/ML, and DevOps practices, training 200+ students.
- Mentored students in implementing real-world software engineering solutions and ML best practices.

Google Developer Student Clubs – Technical Core Team Member (Aug 2022 – Aug 2023)

- Coordinated Google Week events and Android Campus Fest, fostering tech community of 300+ students.
- Provided mentorship on mobile development, cloud integrations, and AI/ML project implementations.

Computer Society of India – Tech Head (Aug 2023 – Present)

- Organized 8+ technical workshops on AI, ML infrastructure, and cloud technologies, mentoring 150+ students.
- Led digital transformation initiatives, developing automation tools and web applications for organizational efficiency.

Additional Information

- **Production Systems Experience:** Deployed and maintained systems handling 10,000+ daily API requests with comprehensive monitoring and alerting.
- **Open Source Contributions:** Active contributor to LangChain, FastAPI, and ML infrastructure-related projects on GitHub.
- **Continuous Learning:** Pursuing advanced certifications in Kubernetes (CKA), AWS Solutions Architect, and ML Engineering.
- Open to relocation and hybrid work models, with experience in distributed team collaboration and agile methodologies.