Rajan Devkota

Kathmandu, Nepal | r.devkota.98@gmail.com | +977 9803006302
linkedin.com/in/devkotarajan | github.com/rajandevkota98

Senior AI Engineer with 3+ years building and operating production AI systems, currently leading a team of five and owning the stack end to end: document intelligence and RAG pipelines, multi-agent orchestration, vLLM serving on Kubernetes, and GPU cost engineering. Depth in high-throughput inference (vLLM, TGI, Triton) and parameter-efficient fine-tuning (QLoRA), delivered across AWS, Azure, GCP, and DigitalOcean.

EDUCATION

Bachelor of Engineering, Computer Engineering, IOE Pashchimanchal Campus, Tribhuvan University *Jan 2018 – June 2023*

- Graduated First Division

EXPERIENCE

Senior AI Engineer, Wildenova, Texas, USA (Full-time, Remote) *Nov 2025 – Present*

- Own DevOps, MLOps, and AI engineering end to end while leading a team of five across 4 services, building from Kubernetes manifests through to shipped product
- Architected a two-stage hybrid document pipeline that decouples layout from transcription: a deterministic layout detector (PP-DocLayoutV3 via the GLM-OCR SDK) segments and crops semantic regions (paragraphs, tables, formulas, charts), then a compact vision-language decoder (Qwen 3.5 4B) served on vLLM transcribes each high-resolution crop in context
- Built long-horizon document workflows on LangChain DeepAgents, using planning, sub-agent delegation, and a shared virtual filesystem to keep multi-step extraction jobs coherent across 500 documents.
- Re-engineered the serving path onto A40 class GPUs at roughly one-sixth the hourly rate of H100, holding throughput with continuous batching and Redis-backed queuing across a Rust and Python service boundary
- Built serverless, demand-driven inference on DigitalOcean Kubernetes (DOKS), fronted by an Nginx ingress behind a DigitalOcean load balancer, with gRPC services, KEDA-based autoscaling across GPU droplets, and CI/CD pipelines driving build and rollout
- Engineered a Graphiti-style temporal memory layer on PostgreSQL and Qdrant over end-to-end encrypted storage, cutting the in-memory index 32× with binary quantization (float32 to 1 bit per dimension) while preserving recall through rescoring, and scaled the vector store horizontally with sharding and replication

AI Engineer (Part-time, 30 hrs/week, Remote), Herix Co, Tokyo, Japan *Jan 2025 – Feb 2026*

- Designed a cost-effective due diligence application reaching 95% accuracy on over 500 annual reports through optimized LLM orchestration and efficient data processing with Azure Functions
- Developed deep research agents for due diligence using LangGraph, enhancing research and grounding of analytical insights
- Built enterprise-grade hybrid RAG system combining vector databases with knowledge graphs, improving precision and enabling multi-hop querying capabilities
- Optimized Document Intelligence model deployment with Triton Inference Server and deployed applications on Azure spot instances and Azure Functions, cutting inference cost 50%
- Delivered across a multi-cloud footprint, running managed model workflows on Vertex AI and event-driven document processing on Azure Functions

AI Engineer, Frost Digital Ventures, Kathmandu, Nepal (Full-time, Onsite) *Sep 2023 – Nov 2025*

- Designed and developed end-to-end chatbot system leveraging advanced retrieval techniques and vector databases, serving over 5,500 case studies and documents across 120+ users a day
- Built a Nepali-language speech-to-text application by fine-tuning Wav2Vec2 on a custom audio dataset, reaching production-ready transcription accuracy
- Architected server infrastructure and optimized resource allocation for efficient deployment of open-source Large Language Models, utilizing TEI for embeddings and vLLM for model serving, 25
- Developed and managed multi-agent systems using LangGraph, integrating agentic frameworks from LlamaIndex and LangChain for complex task orchestration
- Fine-tuned Llama 3 (8B) using QLoRA techniques for domain-specific adaptations and performed instruction tuning of GPT-4o-mini with GPT-4o generated data for cost-efficient inference

Data Scientist Intern, iNeuron.ai, India (Internship, Remote) *July 2023 – Sep 2023*

- Designed and developed end-to-end computer vision project leveraging transfer learning techniques, deployed with FastAPI for production use
- Containerized and deployed applications on AWS using Docker and GitHub Actions for automated CI/CD workflows
- Implemented MLOps practices including model versioning, automated testing, and monitoring for seamless machine learning model lifecycle management

PROJECTS

Sensor Fault Detection System

- *Technologies:* Python, Scikit-learn, AWS EC2/ECR, Apache Kafka, GitHub Actions, MongoDB, Docker
- Designed decoupled microservice architecture for sensor fault detection with automated CI/CD pipelines for training and inference workflows
- Implemented Apache Kafka for real-time data streaming and MongoDB for document storage and retrieval
- Applied MLOps Level 2 architecture patterns on AWS cloud infrastructure for scalable machine learning deployment

Covid Detection Using X-Ray Analysis

- *Technologies:* Python, TensorFlow, AWS (EC2, ECR, S3), GitHub Actions, Apache Airflow, Docker, FastAPI, VGG16
- Developed medical image classification system using deep learning and computer vision techniques to assist radiologists in Covid detection
- Implemented S3-based data storage with APIs supporting both batch and single image inference capabilities
- Integrated Apache Airflow for automated training pipelines and model monitoring workflows

TECHNOLOGIES

AI/ML Frameworks: PyTorch, Transformers, LangChain, LangGraph, DeepAgents, LlamaIndex, vLLM, TGI, Ollama
Model Serving & Optimization: vLLM, Triton, TensorRT, KEDA autoscaling, continuous batching, QLoRA, binary quantization
Vector Databases: Qdrant, Weaviate, Pinecone, FAISS, ChromaDB
Data Processing: PDF/DOCX parsing, OCR, Document layout detection, Excel processing, Web scraping
Computer Vision & NLP: Transfer Learning, Vision-Language Models, Stable Diffusion, OCR Optimization, Speech Recognition
Backend Development: Python, Rust, FastAPI, gRPC, PostgreSQL, MySQL, MongoDB, Redis, Celery, Alembic
Infrastructure & CI/CD: Docker, Kubernetes, Nginx, GitHub Actions, Apache Airflow, Terraform
Observability: Prometheus, Grafana, OpenTelemetry, Sentry, Arize Phoenix, LangSmith
Cloud: AWS (EC2, SageMaker, Lambda, S3), Azure (VM, Blob, Functions), GCP (Vertex AI), DigitalOcean (DOKS, Droplets)