

Ruman Dangol

Senior AI Engineer

rumancha12@gmail.com | +977-9808873409 | [linkedin.com/in/ruman-dangol](https://www.linkedin.com/in/ruman-dangol) | Kathmandu, Nepal

PROFILE

Senior AI Engineer with 7+ years of software engineering experience, now focused on production LLM systems, multi-agent architectures, and GPU-accelerated ML pipelines on AWS. Currently designing and building ECLAT at Maitri Holdings — a healthcare data platform combining large-scale OCR, clinical terminology extraction (LOINC / RxNorm), and orchestrated GPU inference. Strong full-stack and cloud-native foundation across Python, Java and React, with deep hands-on experience in AWS, Kubernetes, Docker, CI/CD, and observability. Proven track record leading cross-functional Agile teams and translating architecture decisions into measurable cost and latency improvements.

CORE EXPERTISE

AI / ML: LLM integration & prompt engineering (chain-of-thought, zero/few-shot, system-message tuning), multi-agent and Agent-to-Agent (A2A) architectures, agent orchestration & memory-aware state, RAG patterns, OCR pipelines, clinical terminology normalisation (LOINC, RxNorm), Gemini 2.5

Cloud & MLOps: AWS (SageMaker Async Inference, Batch, Step Functions, Lambda, EC2, S3, RDS, SNS, SAM), GPU workload orchestration, scale-to-zero vs. warm-fleet cost/latency design, GCP, Terraform, Kubernetes, Docker, serverless

Backend: Python, Java, FastAPI, Django, Flask, Spring Boot, Node/Nest.js, REST, WebSockets, microservices, PostgreSQL, MongoDB, SQL/NoSQL

Frontend: React.js, Next.js, Redux, JavaScript/TypeScript

Observability & QA: Prometheus, Grafana, Loki, Tempo, Pyroscope, ELK Stack, distributed tracing, k6 load testing, GitHub Actions CI/CD

PROFESSIONAL EXPERIENCE

Senior AI Engineer 2026 – Present

Maitri Holdings

- Lead engineer on ECLAT, a GPU-accelerated healthcare data pipeline covering document OCR, medical terminology extraction, and batch plus online model inference on AWS.
- Designed the GPU orchestration architecture, evaluating SageMaker Async Inference, AWS Batch, and Step Functions against throughput, cold-start, and cost profiles for each workload class.
- Established a tiered execution strategy — scale-to-zero for bulk backfill and OCR, warm worker fleets for latency-critical online paths — materially reducing idle GPU spend without regressing user-facing response times.
- Built extraction and normalisation services mapping unstructured clinical text to standard LOINC and RxNorm codes, improving downstream data quality and interoperability.
- Authored the platform architecture proposal and drove technical review with stakeholders, including pre-warming strategies for scheduled nightly backfill windows.
- Own end-to-end delivery: infrastructure-as-code, CI/CD, observability instrumentation, and production reliability for the inference stack.

Senior Software Engineer July 2025 – 2026

TaxMD

- Led development and integration of cookie-based authentication across TaxMD's multi-agent system, coordinating with the user-auth service and taxmd-common library to cover every endpoint.
- Integrated AI-driven capabilities into existing tax analysis tooling, ensuring authenticated access across all AI functionality.
- Designed and implemented scalable backend services supporting tax analysis workflows and AI features.
- Strengthened platform security through advanced validation processes protecting sensitive tax data, and led rigorous end-to-end testing of authentication across AI endpoints.

Senior Software Engineer 2021 – 2025

Fusemachines

- Led a 10-person cross-functional team as Full-Stack Engineer and Technical Lead across multiple AI-driven client projects.
- Spearheaded integration of Google's Agent-to-Agent (A2A) architecture into a multi-tenant GenAI SaaS platform using Next.js and FastAPI, orchestrating dynamic multi-agent interactions and contextual decision-making.
- Embedded Gemini 2.5 as the core LLM, enabling real-time reasoning, agent coordination, and autonomous task execution through structured prompting and memory-aware agent state.
- Engineered secure REST APIs and bi-directional WebSocket channels for real-time agent-user communication, contributing to a 40%+ increase in user engagement.
- Benchmarked single-agent, multi-agent, and A2A paradigms through structured experimentation, validating A2A as the most scalable and context-aware option for evolving use cases.
- Applied advanced prompting techniques — chain-of-thought, zero/few-shot, system-message tuning — to raise agent accuracy and task completion rates.
- Championed observability with Prometheus and Grafana, improving system reliability and cutting incident response time; built CI/CD pipelines with GitHub Actions and Docker.
- Managed technical alignment, prioritisation, and delivery across three concurrent AI projects, including sprint planning and mentorship.

Software Engineer 2020 – 2021

Fusemachines

- Developed and deployed a Rasa-based chatbot system, using Kubernetes Secrets for secure credential management.
- Implemented Velero for backup and disaster recovery and Scaffold for infrastructure provisioning within Codehub.
- Contributed to Codehub, a platform enabling students to learn AI by running Python programs.

Software Engineer Associate 2019 – 2020

Fusemachines

- Led customisation of JupyterHub and Jupyter Notebook environments for AI education using Tornado, Backbone.js, jQuery, and JavaScript.
- Revamped Codehub for 1,000+ active users, delivering a 40% increase in engagement and 30% improvement in satisfaction metrics.
- Deployed the platform on GCP with Kubernetes and Docker.

Software Engineer Associate Trainee 2018 – 2019

Fusemachines

- Developed and maintained fuse.ai, an AI learning platform for students.
- Applied machine learning techniques in the intern training programme, including hand recognition and Optical Character Recognition (OCR) projects.

SELECTED PROJECTS

Health Tracking Application (Dionysus)

Team Lead

- Directed web and mobile development on AWS with Python backend, SNS notifications, and SAM CLI for serverless deployment; served 500+ daily active users.
- Drove Agile delivery to a 30% productivity gain, 20% reduction in delays, and 50% reduction in user-reported issues via a structured feedback loop.

QA Automation & Observability

Food Coupon and HR Management Systems

- Built end-to-end observability across web and mobile using Prometheus, Tempo, and the ELK stack, improving incident response time by 20%.
- Validated scalability with k6 load testing, surfacing bottlenecks under high-traffic conditions.

School Management Platform (Fuse AI)

Backend Developer

- Built online examination, quiz, and grading features plus APIs for real-time student progress tracking and detailed reporting across multiple institutions.

EDUCATION

Master's in Computer Application, Indira Gandhi National Open University — 2024 – Present

Bachelor's in Software Engineering, NCIT, Pokhara University — 2014 – 2019

+2 in Science, Unique College (NEB) — 2012 – 2014

CERTIFICATIONS & TRAINING

- AWS Solution Architect training — secure, reliable, cost-optimised cloud architecture per the AWS Well-Architected Framework.
- Full Stack Data Science Course — Become a Data Scientist, Udemy.
- Automating Real-World Tasks with Python, Coursera.
- Introduction to Image Generation with Google Cloud, Udacity.
- Kubernetes for the Absolute Beginners — Hands-on, Udemy.
- AWS Essentials · Learn Amazon Web Services: The Complete Introduction, Udemy.
- Building RESTful Python Web Services with Tornado, Udemy.