Sumit Yadav

Portfolio: sumityadav.com.np
Github: github.com/rockerritesh
Email: rockerritesh4@gmail.com
Mobile: +977-9819856148
LinkedIn: linkedin.com/in/rockerritesh

Summary

AI Engineer/Researcher specializing in **natural language processing (NLP)** and **AI optimization/Safety**, with 5+ years of experience developing production-grade AI systems. Currently architecting:

- **Multi-agent RAG systems** with guardrails for secure information retrieval.
- **Context-aware chatbots** with post-conversation analysis capabilities.
- **LLM evaluation frameworks** for accuracy and reliability testing.
- **MCP Server** for easy and fast way to integrate Agents.
- **Tutor** Taught courses like Fundamental of Machine learning, compitive programming in C and Pyhton.

Proven track record in **AI/Machine Learning engineering** across many NLP projects including maithili text classification(low-resources) (0.87 accuracy) and multilingual document analysis systems. Authored 3 peer-reviewed publications and one open review paper on **machine learning optimization/security** and **low-resource language processing**

Core Competencies:

- **AI Prompt Design • LLM Fine-tuning • Security Document Analysis**
- **Technical Documentation • Cross-functional Collaboration • GRC Data Annotation**

Education

- **Pulchowk Engineering College** Kathmandu, Nepal
  Bachelor of Computer Engineering
  Courses: SDNs, FinTech, Operating Systems, Data Structures, Big Data, Artificial Intelligenc, Networking, Databases

Skills Summary

- **Languages**: Python, C, C++, Bash
- **Online Courses**: Deep Learning and GAN Specialization, Generative AI LLM, Image Understanding TensorFlow GCP
- **Tools/Module**: CI//CD, GIT,Pytorch,LangChain, LlamaIndex,Django, streamlit, MySQL, GraphQL
- **Soft Skills**: Leadership, Event Management, Writing, Public Speaking, Time Management
- **Hobbies**: Walking, Meditation, Deep Think, Meta-Thinking.

Experience

- **Astha.ai** USA
  AI Engineer (Remote) *May 2025 - Now*
  - **Zero Trust Agentic System**: On Project related to RAG, Agent its identity and agents control flow.
  - **MCP Security**: In-forcing policy on MCP tools, deterministic tool call, Assigning Crypto-Identities on Agentic Components.
  - **Vibe-Flow**: Architectural Design of vibe-flow, A flow that will be enterprise ready flow creation platform, that is designed to solve real-world problems.
  - **MCP-Proxy: Advanced SSE-to-SSE Proxy with Policy Enforcement**: This project implements a sophisticated SSE-to-SSE proxying system that enables secure bridging between upstream and downstream Server-Sent Events (SSE) servers. The implementation features a comprehensive policy engine supporting both v1.0 (simple allow/deny lists) and v2.0 (advanced role-based access control) policy formats, providing granular tool access control with default deny policies for maximum security.
  - **MCP-Proxy: Production-Ready Security Architecture**: The system incorporates enterprise-grade security features including role-based access control (RBAC) for admin/user/guest roles, configurable rate limiting to prevent abuse, conditional access controls with path restrictions and command whitelists, and comprehensive audit logging with full policy decision trails. The architecture implements fail-secure error handling and zero-configuration routing with automatic path generation from upstream URLs.
  - **MCP-Proxy: Intelligent MCP Tool Filtering and Monitoring**: The platform provides real-time MCP tool call interception and filtering capabilities, automatically evaluating tool access requests against JSON-defined policies with support for wildcard patterns, conditional rules, and context-aware decision making. It includes built-in health monitoring, CORS support for web integration, and seamless integration with existing MCP infrastructure while maintaining backward compatibility.
  - **MCP-Scanner: Comprehensive MCP Security Platform**: This project implements a sophisticated security analysis platform for Model Context Protocol (MCP) servers, featuring both vulnerability scanning and real-time defense capabilities. Built on the SAFE-MCP framework, it provides coverage for 78+ attack techniques across 14 security tactics including Initial Access, Execution, Persistence, Privilege Escalation, and Defense Evasion.

- **MCP-Scanner: AI-Powered Vulnerability Detection**: The system leverages Anthropic's Claude AI models for intelligent security analysis, automatically discovering MCP servers, enumerating available tools, and performing comprehensive vulnerability assessments. It includes advanced detection capabilities for context confusion, recursive injection analysis, and state manipulation detection with preliminary risk classification (LOW/MEDIUM/HIGH/CRITICAL).
- **MCP-Scanner: Dual Defense Architecture**: The platform implements two specialized defender systems: (1) *SAFE-T1102 Prompt Injection Defender* for multi-vector prompt injection detection with real-time analysis, and (2) *SAFE-M-1 Tool Poisoning Defender* based on Google Research's CaMeL system, providing control/data flow separation with provable security guarantees achieving 77% task completion rate.
- **MCP-Scanner: Production-Ready REST API**: Features a comprehensive FastAPI-based backend with 15+ endpoints supporting multiple scanning modes (single server, JSON configuration upload, file-based analysis), real-time defender analysis, and database persistence with SQLite storage. The API includes structured response models, comprehensive error handling, and Docker containerization support.
- **MCP-Scanner: Advanced Scanner Implementation**: The `adv` module provides sophisticated tool execution capabilities with dynamic sequence running, timeout management, safe tool execution patterns, and comprehensive logging. It includes an automated scanner runner with retry mechanisms, Anthropic-powered success validation, and structured JSON output parsing for integration with the broader security analysis pipeline.

- **Amnil Technology Pvt. Ltd** Lalitpur
  AI Engineer (Full-time) *May 2024 - May 2025*
  - **Generative AI and Machine Learning Engineering**: On Project related to RAG, Agent based, recursive query, Chatbot, SQL Agent, and scheduling optimization. Made the system like Guardrails, LLM evaluation and Report generation.
  - **LLM hosting, inference optimization, and API integration.**: I hosted different **embedding model** and the **completion model** eg.LLaMA 3.3 3B model on server using the vLLM inference engine, ensuring efficient performance and easy API integration.
  - **AI-Software Engineering**: Developed real-world AI applications and machine learning models for production deployment. Built efficient data pipelines, analyzed large datasets, and collaborated with cross-functional teams to identify and implement AI solutions. Demonstrated expertise in modern ML frameworks and maintained professional excellence throughout tenure. Tech: Python, TensorFlow, PyTorch, Scikit-learn, Hugging Face, Data Engineering, ML Pipelines

- **Ed-Acadia** Lalitpur
  Chief Data Oficer (Full-time) *May 2022 - 2023*
  - **AI/ML Projects**: Supervising the project and research related to Data Science. Works of different DocumentsAI system for low resources language.
  - **Question-Answer Sytem**: Large scale semantic embedding, simillarities calculation between the question and answer pair.
- **PDSC(Plan Design Solve Create)** Lalitpur
  Software Coordinator (Full-time) *May 2022 - 2023*
  - **Project Management**: Supervising the project and research related to Data Science.
- **DeepLearning.AI** Virtual
  GAN Mentor (Part-time) *Aug 2021 - Present*
  - **Course - GAN Specialization**: Helping the student in understanding the key concept behind Unsupervised learning (GAN).
- **Robotics Association of Nepal** Lalitpur
  AI and Robotics Member (Part-time) *2021 - Present*
  - **Making Robotics based system**: Done research and project related to Computer Vision based on raspberrypi microcontroller.

Publications

- **Can maiBERT Speak for Maithili?**: Natural Language Understanding (NLU) for low-resource languages remains a major challenge in NLP due to the scarcity of high-quality data and language-specific models. Maithili, despite being spoken by millions, lacks adequate computational resources, limiting its inclusion in digital and AI-driven applications. To address this gap, we introducemaiBERT, a BERT-based language model pre-trained specifically for Maithili using the Masked Language Modeling (MLM) technique. Our model is trained on a newly constructed Maithili corpus and evaluated through a news classification task. In our experiments, maiBERT achieved an accuracy of 87.02 percent, outperforming existing regional models like NepBERTa and HindiBERT, with a 0.13 percent overall accuracy gain and 5-7percent improvement across various classes. We have open-sourced maiBERT on Hugging Face enabling further fine-tuning for downstream tasks such as sentiment analysis and Named Entity Recognition (NER).
- **SafeConstellations: Steering LLM Safety to Reduce Over-RefusalsThrough Task-Specific Trajectory**: SafeConstellations addresses the problem of LLM over-refusal, where safety mechanisms cause models to reject benign instructions that superficially resemble harmful content (e.g., refusing to analyze sentiment of "How to kill a process"). The method discovers that LLMs follow distinct "constellation" patterns in embedding space for different tasks, enabling targeted inference-time steering that guides representations toward non-refusal pathways without compromising safety on genuinely harmful content. Their approach achieves up to 73 per Cent reduction in over-refusals across multiple models (Claude, GPT-4o, LLaMA, Qwen) while maintaining utility and preserving legitimate safety mechanisms.

- **Revolutionizing Currency Security: A Yolov8-Based Approach for Automated Detection of Counterfeit Nepali Banknotes**: Implemented YOLOv8 to achieve a true positive recall of 0.82 (front face) and 0.9863 (back face) in detecting counterfeit Nepali banknotes, demonstrating significant advancements in counterfeit currency detection.
- **Machine Learning Analysis of Tirhuta Lipi**: Achieved 0.97 accuracy in Tirhuta Lipi character recognition using MobileNet embedding and logistic regression, with applications in translation and OCR for low-resource languages.
- **SUPPORT VECTORS ARE A BETTER WAY OF TEXT CLASSIFICATION FOR IMBALANCED DATA**: Present a robust SVC method for text classification (100+ classes) using term-frequency vectorization, achieving superior test data results over neural networks.

Projects

- **SAFE-MCP Security Framework Contributor**: Authored three critical attack techniques for the Security Analysis Framework for Evaluation of Model Context Protocol (SAFE-MCP). Contributed SAFE-T1601 (MCP Server Enumeration), SAFE-T1703 (Tool-Chaining Pivot), and SAFE-T1110 (Multimodal Prompt Injection via Images/Audio), providing comprehensive documentation, Sigma detection rules, test cases, and mitigation strategies for emerging AI security threats. Tech: Cybersecurity Research, MITRE ATT&CK Framework, Sigma Rules, Python, YAML, Security Analysis
- **Agents.ai**: An intelligent multi-agent system that automatically selects the best agent and tool sequence to handle user queries. The system uses semantic similarity to match user requests with specialized agents, each equipped with specific tools for different tasks. Tech: Agent, MCP, Claude API keys, Python, Streamlit
- **Vibe-Coder**: Made an Agent that will do Streamlit and FastAPI. Tech: Agent, MCP, Claude API keys, Python, Streamlit
- **Retrieval Augmentation Generation System (RAG) and Intelligent Document Processing(IDP)**: Developed a retrieval-augmented reality system for enhanced information access and interaction. Tech: OpenAI, Gemini, Claude API keys, Python.
- **Nepali Chat with Doc**: Implemented a chatbot for Nepali language using Devanagari and Preeti fonts. Features include Guardrails system, post-conversation analysis, and agent-based systems like SQL Agent, Excel Agent, and Reflexive Agents. Preeti to Unicode Conversion. Tech: OpenAI, Gemini, Claude API keys.
- **Bachelor's Major Project: Evaluating Auto-Encoder Transformer Language Model for Maithili Text Classification**: Established a benchmark in this language. First to create a corpus in Devanagari Maithili language, trained LLM for Maithili, and performed downstream task classification. Tech: LLM, Transformer(bert), Pytorch, Streamlit & Big Data. (April '2024)
- **IRB (Image Recognition Based) Robotics Arm (Image Processing, Signal Processing, Actuator Control)**: Research-oriented, open-source project under UN's SDG3 - Good Health & Well-Being. Tech: Python, Arduino Programming, Arduino Toolkit, TensorFlow (May '2020).
- **Nepali Language Projects**: Developed multiple applications, including a Devanagari letter classifier using VGG16 (accuracy 0.94), a Nepali sentiment analysis model, and a simple OCR for Nepali text. Tech: Keras, Transformer, Pytorch, TF-IDF, NLTK. (Past 2 Years)
- **Unsupervised Model**: Explored the behavior of latent spaces using VAE, GAN, C-GAN, AC-GAN, and DC-GAN. Tech: Python, Numpy, TensorFlow. (Sep, 2021)
- **NEPSE Simple**: Presented Nepal stock market data in a minimal environment constraint. Tech: GitHub Workflow, Automation in Scraping, WebSockets, JavaScript, RSS, XML. (Since 2020)
- **Advanced Document and AI Systems**: Designed and implemented a variety of tools, including:
  - Chat systems for Nepali and multilingual documents with Preeti-to-Unicode conversion and guardrails for improved user interaction.
  - AI-powered memo creation and advanced Excel file manipulation tools.
  - Contract document analysis using recursive and advanced reasoning GPT systems.
  - Translation systems for Nepali documents using OCR and text conversion.
  - Chat and interaction systems for image and audio data with TTS and Whisper integration.
- **Verification and Financial Prediction Systems**: Developed:
  - A face and signature verification app using VGG-based advanced face detection and liveness detection algorithms.
  - A loan eligibility prediction system utilizing knowledge-based reasoning techniques.

Honors and Awards

- Winner of GritFeat AI Hackathon 2023, Locus - Feb, 2023,(SWIFT' is a wearable devices with hardware and AI models that detect falls in elderly people with 0.7986 accuracy, resulting in immediate emergency alerts to contacts.)
- First RunnerUP of Dataverse, Locus - Jan, 2023,Dataverse Solution (NLP pased problem to classify abstract.)
- Winner of Best AI Project of Deltathon, DELTA 3.0 - Jan, 2022,Nepali Harvest (Designed a portal to help farmers that can predicting diseases, identifying optimal harvest times, and aiding with crop health assessment.)
- Winner of Image Challenge, IT-Meet UP KU - Sep, 2022 (Have to train AI model to classify image of Ballot paper.)
- Winner of Capture The Flag, LogPoint - Feb, 2021 (Tasked of finding information and exploiting a binary file.)
- Runner's Up at DATARUSH by DOCSUMO - Feb, 2021 (NLP based model for classifying Abstract into Classes.)

Social Experience

- **Team of NPL Coders** — Global
  Conducted National level Data Science Coding Compition on Kaggle and Hacker Ranks. — *Sep 2023 - Present*
- **Joint Secretary at NTBNS Student Clubs, IOE, Pulchowk Campus** — Lalitpur, Nepal
  Conducted technical training & Organized nepal largest sarswati puja Program. — *Jan 2020 - Present*
- **Tutor of Children In Technology- WorldLink** — Nepal
  Aware the student about Risk and Safety of Internet. — *Nov 2023*