BIBEK RAWAL

Jumla, Nepal • +977 9864851923 • bibek.rawal.588@gmail.com
bibekrawal.com.np/ • www.linkedin.com/in/bibekrawal • github.com/BibekRawal5

AI Engineer working across the LLM stack: fine-tuning and domain adaption, agentic and multi-step RAG, MCP, multi-agent orchestration, and the evaluation, safety and observability layers that keep such systems honest in production. Research interest in NPL, low resource languages, evaluation and safety. Ships project end-to-end, from retrieval to development, B.Sc. CSIT with Distinction.

EXPERIENCE

Frost Digital Venture (Frost & Sullivan) | Kathmandu, Nepal
Associate AI Engineer | Oct 2025 – Present

- **Architectured and Built Flow.** An AI-agent native low-code workflow automation platform, replacing about 40% of backend services with **MCP chats**. Added multiple workflow communication channels through platforms likes Teams, Slack, WhatsApp and Telegram.
- **Merge Request (PR) Automation.** Developed a custom agentic workflow to automatically verify internal merge request changes and identify conflicts and security issues before human review. Generating inline comments with email notification to assigned developer with actionable requests for resolution prior to final merge approval. Deployed on GitLab runner for constant monitoring.
- **OCR and vision-language models.** Optimized a document OCR and segmentation platform to improve the extraction of formatted tables and figures, achieving about 30% better detection of tables and figures and reducing incorrect separation of single figures caused by gaps. Evaluated and tested DETR- and YOLO-based object detection models for document layout analysis.
- **Multi-agent orchestration.** Designed a multi-agent system called DocPilot with retrieval grounding and confidence-based escalation, reducing human intervention 80% in customer assistance.
- **RAG and fine-tuning.** Built multi-step retrieval pipelines with query decomposition and re-ranking; fine-tuned open-weight LLMs with LoRA, QLoRA and PEFT.
- **Delivery.** Shipped FastAPI backends and containerized deployments using docker and Kubernetes, owning the AI side end-to-end on assigned client projects across financial services and retail, optimal agentic coding with Codex, ClaudeCode, antigravity and other tools.

GaoTek Inc. | Remote, US
AI Tech Intern | Jun 2025 – Aug 2025

- **Architectured and Built CCRAG.** An action oriented RAG customer care chatbot with multiple tools to perform live actions like updating KYC.
- **Retrieval Testing.** Used Precision@7, Recall@7, MRR and nDCG where on each test achieved over 80% result.
- **SEO tool.** Developed an AI-native SEO evaluation tool, that measures the SEO suitability of a website automatically without any source code. Then also prroivdes measure and steps to improve the SEO score.
- **Web Scraping:** Used Playwright and selenium to extract websites content for SEO evaluation and Knowledge base creation.
- **Delivery.** Completed the chatbot and SEO application from initial POC to final dockerized deployed product in less than 2 months.

Sprouts Technology (CloudFactory)| Remote, Nepal
Data Specialist | Jan 2025 – Oct 2025

- Performed quality **data annotation and 3D modeling** across different computer vision and geospatial projects.
- Contributed to Nearmap's 3D roof detection project, supporting the development and refinement of large-scale **geospatial datasets.**
- Surpassed exepected **project-defined quality and productivity SLAs**.
- Active engagement in project discussions, review sessions, and management meetings, aligning on requirements, quality standards and timely delivery.
- Collaborated with project teams to identify annotation challenges, improve workflows, detect bug in internal tools and maintain dataset quality and consistency.

TECHNICAL SKILLS

Vision and Document AI: Vision-language models, OCR for Nepali and low-resource scripts, image preprocessing, document understanding

Machine Learning: Supervised and unsupervised learning, deep learning, NLP, transformers, exploratory data analysis, feature engineering

LLMs and Agents: Fine-tuning (LoRA, QLoRA, PEFT), domain adaptation, multi-step and agentic RAG, LangChain, LangGraph, multi-agent orchestration, tool use, MCP, human-in-the-loop escalation, retrieval and re-ranking

Evaluation and Safety: Ragas, MLflow, LLM-as-judge, custom code-level evaluation pipelines, groundedness, hallucination, safety and refusal scoring, Grafana observability

Libraries and Frameworks: PyTorch, TensorFlow, scikit-learn, FastAPI, Pydantic, NumPy, Pandas, Matplotlib, Seaborn, Streamlit,

Data and Storage: Vector databases, Neo4j, graph machine learning, SQL, Postgres, RustF and S3 bucket

Programming and Tools: Python, C, C++, SQL, Git, GitHub, Docker, WebSockets, linux

PROJECTS

NLTTA: Nepali Language Transliteration and Translation Automation

- Specially trained mBART model to support Romanized Nepali, Devanagari Nepali, and English translation and transliteration.

- Curated a dataset from Wikipedia scraping, open-source datasets, and translation tools for fine-tuning.
- Compared NLTTA with other state of the art models, beating some bigger models like LLAMA.
- NLTTA reliability and accuracy tested with 39.84 and 0.9613 BLEU and BERT score respectively.
- Built a custom hybrid retrieval system (embedding similarity + custom BM25) using NLTTA output to get relevant news articles for users.
- Achieved remarkable and **TU highest 98/100** on the project.

CCRAG: Customer Care RAG Chat

- Multi-stage Customer Support agentic RAG over GaoTek content, with a shared backend API serving web based clients.
- Created pipeline of retreival and generation to handle customer and employee queries using internal documents.
- RAG-backed conversational search with optional image input, and a constraint-based trip planner.
- Action based chat as well including actions like KYC update.
- Designed a contribution and expert-review pipeline with versioned content states, keeping generated answers traceable to human-verified sources.
- Tested retrieval quality with Precision@7 and Recall@7, Generation tested using Groundness, Relevance, and BLEU scores.

SRGAN: Single Image Super Resolution using GAN

- Generating High-Resolution image out of a Low Resolution input(based on SRGAN paper).
- Re-implementation of model proposed by SRGAN paper
- Trained on about 1000 images of cars and still portrait poses.
- Trained with limited compute and still achieved remarkable results in 20 epochs.

EDUCATION

Tribhuvan University, Bhaktapur Multiple Campus | Bhaktapur, Nepal

B.Sc. CSIT — 83.64%, Distinction | 2022 – 2026

- Ranked top 2 in the cohort. Final-year project: Nepali Language Transliteration and Translation Automation(NLTTA), got the highest score of 98/100.
- Relevant coursework: Calculus, Linear Algebra, Statistics I & II, Numerical Methods, Object-Oriented Programming, Data Structures and Algorithms, Design and Analysis of Algorithms, Operating Systems, Computer Networking, Database Management Systems, Data Warehousing and Data Mining, Compiler Design, Core AI/ML Concepts, Software Engineering, System Analysis and Design.

WORKSHOP AND TRAINING

- **Academic-Industry Dialogue, Research Management Cell, Bhaktapur Multiple Campus**
  - **Represented Bhaktapur Multiple Campus CSIT Department.** Shared perspectives and recommendations on academic and research development with the **Dean and Rector of TU** and **industry specialists.**
  - **Research and Academic Development.** Presented perspectives on strengthening research activities, improving academic initiatives, and supporting the continued development of the Campus and University.
  - **Academic-Industry Engagement.** Contributed to discussions on connecting academic research with industry needs and identifying opportunities for stronger collaboration between academic institutions and industry.
- **AI in Healthcare, NAAMI and B.P. Eye Foundation**
  - **Medical Research and Critical Analysis.** Analyzed medical research papers to extract key findings, research methodologies, limitations, and identified gaps and weakness in existing AI/ML approaches**.**
  - **AI/ML Research and Healthcare Challenges**. Researched challenges in medical data acquisition and annotation, ethical considerations, and methodological gaps in reliability and applicability of AI/ML research in healthcare.
  - **Research Findings Presentation**. Extracted and presented findings from medical research literature, highlighting opportunities for improving AI-driven healthcare research and applications.

SCHOLARSHIPS AND COURSE CERTIFICATIONS

- **Scholarships:**
  - Jehendar Scholarship | Tribhuvan University Academic Excellence Scholarship (Ranked #1 On Campus)
  - Merit Scholarship for Technical Bachelor Studies | Directorate of Education Development, Karnali Province (Best IT undergrad)
- **Course Certifications:**
  - TensorFlow Developer Specialization, DeepLearning.AI (Coursera)
  - Structuring Machine Learning Projects, DeepLearning.AI (Coursera)
  - Natural Language Processing in TensorFlow, DeepLearning.AI (Coursera)