RISHAV UPADHAYA

+977-9865993983 | rishavupadhaya266@gmail.com | linkedin.com/in/rishav-upadhaya | github.com/Rishav-Upadhaya

PROFESSIONAL SUMMARY

AI Engineer with hands-on experience building Generative AI, RAG, and multi-agent systems, including a live production deployment. Designed and implemented LLM-powered backends, vector search pipelines, and LLM observability workflows. Skilled in Python, LangChain/LangGraph, FastAPI, and ML fundamentals including NLP, classification, and fine-tuning. Experienced collaborating in cross-functional Agile teams to translate business requirements into scalable AI solutions. Graduating December 2026.

EDUCATION

Tribhuvan University Kathmandu, Nepal

Bachelor of Science in Computer Science and Information Technology, Expected December 2026 *Nov 2022 – Dec 2026*

- Relevant Coursework: Data Structures and Algorithms, Artificial Intelligence, Machine Learning, Database Systems, Natural Language Processing, Operating Systems, Software Engineering, Computer Networks

TECHNICAL SKILLS

AI & ML: LLMs (GPT-4, Gemma), RAG Systems, Multi-Agent Orchestration, NLP, Sentiment Analysis, Text Classification, Model Fine-tuning (QLoRA/PEFT), Model Evaluation
ML Frameworks: Hugging Face Transformers, scikit-learn, LangChain, LangGraph, LangSmith (LLM observability & tracing)
Primary Stack: Python, FastAPI, PostgreSQL/pgvector, Docker, Git, Linux
Cloud & AI Services: AWS Textract, Azure Document Intelligence, Google Vision API
Data & Evaluation: Pandas, NumPy, OpenCV, Pydantic, SQL, Feature Engineering, Cross-Validation, Precision/Recall/F1/ROC-AUC, CER/WER
Databases: PostgreSQL, Pinecone, pgvector (HNSW indexing)
Practices: Agile/Scrum, Code Review, Technical Documentation

PROFESSIONAL EXPERIENCE

AI Engineer (Part-Time) May 2025 – Present

AsterGaze Technologies *Kathmandu, Nepal*

- Designed and built multi-agent AI system for a CRM platform in study abroad consultation, orchestrating agent workflows with LangGraph and FastAPI achieving 3–5 second end-to-end response times
- Architected RAG pipeline with hybrid search combining semantic (neural) and metadata retrieval, optimizing vector search from approximately 10 seconds to under 4 seconds using pgvector HNSW indexing
- Reduced LLM inference costs by up to 55% through prompt engineering, context compression, and token optimization; tracked LLM trace observability, token usage, and latency via LangSmith
- Designed scalable PostgreSQL schema with SQLAlchemy ORM, implementing N+1 query optimization, lazy loading, and OpenAI text-embedding models for semantic search

Backend Developer Intern Nov 2025 – Feb 2026

Proshore.eu *Kathmandu, Nepal*

- Built multi-engine OCR pipeline integrating AWS Textract, Azure Document Intelligence, and Google Vision API for automated document data extraction in regulated workflows, reducing manual processing from 8 human-hours to under 3 seconds per batch while maintaining 96% data fidelity
- Engineered computer vision preprocessing pipeline using OpenCV for skew correction, grayscale conversion, and orientation detection, improving OCR accuracy by 35% measured via Character Error Rate (CER) and Word Error Rate (WER)
- Developed RESTful APIs with Pydantic schema validation, comprehensive error handling, and OpenAPI documentation for document processing workflows
- Collaborated in Agile sprints with cross-functional teams of 10+ engineers (PM, QA, Frontend, Backend), translating business requirements into scalable technical solutions

Student Partner Apr 2025 – Oct 2025

Leapfrog Technology, Inc. – Leapfrog Student Partnership Program *Kathmandu, Nepal*

- Selected from 500+ applicants for competitive software engineering apprenticeship; led development of Reviso.ai, an LLM-powered exam platform with automated question generation and NLP-based evaluation capabilities
- Built modular FastAPI backend integrating GPT-4, LangChain, and Pinecone vector database with CI/CD pipelines, supporting concurrent evaluations with real-time analytics
- Applied Agile methodologies, testing practices, and SOLID principles; participated in peer code reviews to maintain code quality and system maintainability

FEATURED PROJECTS

SWLP: Sliding Window Layer Pipeline | ***Python, PyTorch, Hugging Face Transformers, MLX*** 2026

- Engineered a memory-efficient FP16 inference system enabling a 26 GB transformer model to run on a 16 GB machine at 1.7 GB peak RAM, no quantization, no quality loss, by streaming layers one at a time from NVMe SSD to RAM to compute, then evicting to zero-byte meta tensors; achieved SOTA throughput on the MLX backend
- Implemented concurrent I/O and compute via background prefetch threading (layer N+1 loads while layer N runs), making SWLP 2× faster than AirLLM; added speculative decoding via n-gram matching for up to 4 accepted tokens per disk sweep, with linear batch throughput scaling at a flat 0.28s sweep cost

JobRAG: Hybrid Search Pipeline for Job Intelligence | ***LangGraph, FastAPI, pgvector, Gemini API, Python*** 2026

- Built a production-grade RAG pipeline over 1,000+ job listings using a LangGraph state graph (intent → retrieve → synthesise → judge) with PostgreSQL + pgvector HNSW for vector search, a Jina cross-encoder reranker for precision, and a pluggable embedder abstraction supporting Gemini, Cohere, and Hugging Face providers
- Designed a split schema (`jobs` for metadata filtering, `job_chunks` for embeddings) with direct psycopg2 SQL for predictable performance; unified all LLM calls through a single Gemini factory for centralized rate limiting, prompt versioning, and cost control across classifier, synthesizer, and judge nodes

DiabetesInsight: Multi-Class ML Pipeline | ***scikit-learn, Pandas, NumPy, Python*** 2026

- Built an end-to-end ML pipeline for 3-class diabetes prediction (Non-diabetic / Pre-diabetic / Diabetic) combining supervised classification across multiple models, unsupervised K-Means clustering with PCA visualization, and full EDA covering class distributions, feature correlations, and data validation
- Implemented comparative model evaluation with confusion matrices, precision, recall, F1-score, and ROC-AUC tracking; exported trained model artifacts, scaler, figures, and CSV comparison tables for full reproducibility

Reviso.ai: Exam Generation & Evaluation Platform | ***GPT-4o, LangChain, Pinecone, FastAPI*** 2025

- Co-built a full-stack exam lifecycle platform supporting configurable LLM-generated question sets (descriptive and analytical, per-difficulty distribution), automated answer evaluation with accuracy/clarity/depth scoring, AI-generated flashcards and quizzes, and an async batch evaluation pipeline for concurrent assessment loads
- Integrated Pinecone vector search for subject-filtered document retrieval across 4 domains, advanced proctoring with cheating detection (eye movement, multi-person, audio anomaly, spoofing), and per-student and admin-level analytics dashboards; deployed with CI/CD pipelines as part of Leapfrog Student Partnership Program

LEADERSHIP AND ACHIEVEMENTS

Hackathons: Locus (Winner) – built AI-powered product integrating agentic system; also participated in SecurityPal, Sandbox, and CodeYatra

Campus Director & Mentor, Hult Prize at Samriddhi College (2024–2025) – led 15+ students in social entrepreneurship competitions

Executive Member, Samriddhi IT Club (2023–2024) – organized 10+ technical workshops with 200+ participants

Certifications: LangChain Chat with Your Data (DeepLearning.AI), Data Analysis with Python (IBM), Python for Beginners (Cisco), IT Essentials (Cisco)