

SUBAN SHRESTHA

+977 9848655422 • subanshrestha244@gmail.com • <https://www.linkedin.com/in/suban-shrestha-65b7731b3/> • github.com/suban244

SUMMARY

AI/ML Engineer specializing in LLM agents, RAG systems, and production AI backends. Deployed LLM-powered applications handling 100,000+ automated interactions across multilingual customer service and enterprise document workflows.

PROFESSIONAL EXPERIENCE

REG-1, London, UK: AI / ML Engineer

Feb 2026 - Present

- Built a **general purpose chat agent** using **pydantic-ai**, that serves as second interface to the platform, letting users create and complete forms and interact with UI elements, and answering questions across internal regulations, contracts, and policy documents.
- Developed **long-running background agents** that automate document and form-based business workflows, integrating with existing enterprise systems.
- Instrumented microservices with OpenTelemetry for distributed tracing, performance monitoring, and alerting.

reAlpha, Ohio, US: Applied AI Engineer

April 2024 – Feb 2026

- Developed **multi-agent orchestration layer** for **customizable** conversational AI agents that automated **100,000+** customer service interactions across multiple languages.
- Engineered conversational AI backend enabling **interactive chat interfaces** with observability via OpenTelemetry, integrated performance monitoring, and evaluations.
- Built **FastAPI/Django** services and **LLM-powered data pipelines** for real estate data integration at scale, enabling housing market analysis and enhanced search.
- Implemented end-to-end document parsing system with **RAG-powered search**, combining full-text, vector, and hybrid search methodologies using **OpenSearch and Qdrant**.
- Collaborated with **cross-functional teams across multiple countries** to define requirements and tune performance.

reAlpha, Ohio, US: Software Engineering Intern

Jan 2024 – Mar 2024

- Built **RAG** over financial documents using **LlamaIndex**, with advanced chunking strategies and specialized parsers for 10-K and 10-Q filings.

NAAMII, Lalitpur, Nepal: Research Intern

Nov 2023 – Mar 2024

- Researched **Transformer-based vision models** for medical image segmentation and built reproducible workflows for replicating published results.

PROJECTS

Watson: Self-hosted personal AI agent on a Raspberry Pi 🐉

FastAPI, Pydantic-AI, TaskIQ, SvelteKit, PostgreSQL, Docker

- Expense and budget tracking, analysis and reminders.
- Model sandboxing for generated SQL, Cloudflare Tunnels for access and GitHub Actions CI/CD

Extractive Nepali Question Answering System 📄

Fine-tuned muRIL-base-cased model for Nepali Question Answering

- Achieved **60.5%** exact match and **76.89** F1 score on XQuAD benchmark evaluation.
- Implemented and evaluated multiple **translation strategies** for low-resource language adaptation with comprehensive **dataset validation**.

TECHNICAL SKILLS

Programming & Web Development: Python, FastAPI, Django, REST API, GraphQL, Git/GitHub, Celery

AI/ML & Data Science: LLMs, OpenAI, Pydantic-AI, LlamaIndex, LangChain, PyTorch, Scikit-Learn, NumPy, Pandas

Databases & Search Technologies: PostgreSQL (PostGIS, pgvector), OpenSearch, Qdrant, Pinecone

Cloud & Infrastructure: AWS, Docker

EDUCATION

Bachelors in Computer Engineering

Nov 2019 - April 2024

Pulchowk Campus, Tribhuvan University, Nepal

Relevant coursework: Artificial Intelligence, Big Data, Enterprise Computing, Statistics

AWARDS

- Winner - Gritfeat AI Health Hackathon (Jan 2023)
- Winner - Ace Ignite 48-hour Hackathon (Nov 2022)
- Winner - IT MEET 2022, KU, Image Challenge (Aug 2022)
- Finalist - Hult Prize, Pulchowk (2022)

ACTIVITIES

- **Children in Technology** (Nov 2023): Led a team of 13 undergraduate students to 5 underserved rural communities in eastern Nepal, engaging with 13 schools and approximately 2,000 students on internet safety and cyber law.
- **Robocamp Instructor** (July 2023): Instructed electronics fundamentals through advanced microcontroller programming.
- **Software Fellowship** (Aug 2023) : Designed and taught software engineering fundamentals and machine learning algorithms to 100+ students
- **Human Resource Manager, PDSC Pulchowk Campus**: Organized and managed 5+ technical workshops and community outreach events.