

ROSHAN CHHETRI

AI Engineer | GenAI, RAG & Backend Systems

Sikkim, India · Open to relocate · Requires visa sponsorship · +91 9749922509 · roshanchhetri931@gmail.com
github.com/Roshan-Chhetri28 · linkedin.com/in/roshan-chhetri28

SUMMARY

AI Engineer building production GenAI systems: LLM applications, RAG pipelines and agentic workflows with LangGraph, FastAPI and PostgreSQL/pgvector. Built and deployed an autonomous content-generation system on a 13,000-user platform, generating 4,000+ learning items in production over the past year and cutting educator content-preparation time by 60%. Builds end-to-end AI systems spanning LLM applications, retrieval, backend services and production deployment.

SKILLS

AI / GenAI:	LLMs, RAG pipelines, Agentic AI, OpenAI API, PyTorch, Deep Learning
Agent frameworks:	LangGraph, LangChain, LangSmith
Backend:	Python, FastAPI, Node.js, Express.js, REST APIs, SQL
Data & Retrieval:	PostgreSQL, pgvector, Pinecone, MySQL, Redis
Infrastructure:	Docker, AWS, GitHub Actions, CI/CD
Frontend:	JavaScript, React.js, Tailwind CSS

EXPERIENCE

Spaced Revision | AI Engineer Sikkim, India · 2025 - Present
Spaced-repetition study platform for students, ~13,000 users: flashcards, notes and active recall.

- Built and deployed a **LangGraph** agentic content-generation system that uses multi-step **RAG** to autonomously produce MCQs, flashcards and study notes from source material; generated **4,000+** items in production over the past year and cut educator content-preparation time by **60%**.
- Designed and shipped a real-time chat service using **Socket.IO** and **Express**, supporting 1:1 and group channels with deterministic room coordination across MySQL, MongoDB and Redis; in production for **9+ months**.

Bytesberry Technologies | Software Development Intern 2024 - 2025

- Collaborated on the development of a headless CMS using **React** and **Node.js/Express**, implementing role-based access control and contributing across the application lifecycle.
- Worked in an Agile team across the full release cycle, contributing to code reviews and cross-functional technical planning.

Waglogy Tech LLP | Data Analyst & AI Engineer (Freelance, concurrent) 2022 - 2025

- Built an **LLM-powered automation pipeline** that extracted and classified client data, routed customer-support requests, and incorporated Google search-trend data to identify emerging Sikkim tourism topics for SEO-focused content and recurring reports.
- Owned the extraction and routing layer, removing the manual data-entry step from the delivery process entirely.

PROJECTS

Threadly | Semantic note-taking application github.com/Roshan-Chhetri28/Threadly

- Built a semantic search pipeline over 1,000+ notes using OpenAI text-embedding-3-small with cosine-similarity nearest-neighbour retrieval in **pgvector**, returning the expected note in the top results for **80%** of a hand-built query set.
- Grouped related notes using embedding similarity, automatically surfacing semantically related content without manual tagging.

Neural Network from Scratch in C github.com/Roshan-Chhetri28/Neural-Network-in-C

- Implemented forward/backward propagation, SGD, Sigmoid and ReLU, Xavier/He initialisation and binary cross-entropy with **no ML libraries**.
- Traced a non-converging network to a data-loading bug: the loader was reading the CSV index column as the label; correcting the loader restored convergence.

EDUCATION & ACHIEVEMENTS

B.Tech, Computer Engineering. Sikkim Institute of Science and Technology (Sikkim University), 2025
NPTEL Machine Learning (ML). Top 1% Performer · 81%