

MANU REJIMON

Kochi, Kerala (Open to Relocation) | manurejimonoffl@gmail.com | +91-XXXXXXXXXX
linkedin.com/in/manurejimon | github.com/Manu080405 | leetcode.com/Manu080405

PROFESSIONAL SUMMARY

Final-year Computer Science (Data Science) student specializing in backend engineering, Application Programming Interface (API) design, and Artificial Intelligence (AI)/Machine Learning (ML) systems. Built production-grade services handling 300+ document corpora with sub-4-second latency using Python, FastAPI, and FAISS. Co-founded a registered AI research startup, independently delivering end-to-end software systems from architecture to deployment.

TECHNICAL SKILLS

Languages: Python, SQL, C, JavaScript (JS)

Backend and APIs: FastAPI, Node.js, Express.js, REST Application Programming Interface (API) Design, JSON Web Token (JWT) Authentication, Role-Based Access Control (RBAC), Microservices

AI and Machine Learning (ML): Retrieval-Augmented Generation (RAG), FAISS, Sentence Transformers, BGE Embeddings, Large Language Models (LLMs), Scikit-learn, NumPy, Pandas

Databases: PostgreSQL, MySQL, FAISS Vector Indexing

Tools and Platforms: Git, GitHub, Postman, VS Code, React.js, Vite

Data and Analytics: Microsoft Excel (MS Excel), Power BI (Business Intelligence), Pivot Tables, DAX, Dashboard Design

EXPERIENCE

Software Engineer Intern (AI/ML Intern)

Feb 2025 – Jun 2026

Bharat Tech Foundation

Bangalore, Karnataka

- Led a 3-engineer Large Language Model (LLM) sub-team within an 8-member cross-functional group, owning delivery of all AI modules (reviewer matching, summarization, plagiarism detection, format validation) on schedule.
- Reduced manual conference reviewer assignment time by ~70% by architecting a semantic reviewer-matching system using LLaMA 3.1 8B (via Hugging Face API) and deploying it as a FastAPI microservice with sub-4-second response latency – delivered via clean Application Programming Interface (API) contracts to a 5-member backend team.
- Cut submission originality review effort by ~60% by engineering a plagiarism detection pipeline that generated dense embeddings from ArXiv paper metadata (title, authors, abstract) and performed semantic similarity search against a live corpus of 10,000+ papers.

Software Engineer Intern (AI/ML Intern)

Jan 2026 – Present

Avark Research Intelligence Technologies Pvt. Ltd.

Kochi, Kerala

- Architected MOUDEL – a production RAG backend ingesting 236 arXiv papers into 15,357 searchable chunks via section-aware PDF chunking (600-word sliding windows, 100-word overlap), BGE-small-en-v1.5 embeddings, and FAISS IVF indexing – isolating retrieval latency to under 4ms (FAISS-only, benchmarked) while keeping the full application layer under 270ms overhead.
- Achieved sub-2-second end-to-end query latency (median 1.4s, p95 1.49s, CV 10.8%) by engineering a single-pass async inference pipeline (FastAPI + LLaMA 3.1 8B on NVIDIA NIM) with similarity threshold filtering (0.25), per-source deduplication, and confidence scoring – sustaining zero failures under 10 concurrent users with stable jitter under 160ms.
- Hardened the backend to a 97% security pass rate (69/71 checks) by implementing input validation blocking SQL, NoSQL, shell, and path traversal injection, locking CORS to approved origins, disabling production API docs, and enforcing rate limiting on all inference endpoints – delivering a full-stack research assistant (React + Vite, FastAPI) within 4 weeks of project start.

PROJECTS

LexAI — FastAPI, FAISS, BGE Embeddings, Ollama, Python

March 2025

- Achieved sub-3-second average query latency on CPU-only infrastructure by building a legal research backend over 350+ Bharatiya Nyaya Sanhita (BNS) sections using FAISS Approximate Nearest Neighbor (ANN) search with BGE-large embeddings – eliminating GPU infrastructure costs entirely.
- Improved retrieval precision by ~30% on legacy legal queries by constructing a cross-reference dataset mapping 500+ Indian Penal Code (IPC) provisions to BNS equivalents, enabling automatic query reformulation for legacy legal terminology.
- Reduced hallucinated legal citations to zero across 100+ benchmarked real-world queries by designing a citation-aware response layer that anchored every generated answer to verified BNS source passages.

EDUCATION

SCMS School of Engineering and Technology

Expected May 2027

Bachelor of Technology (B.Tech) – Computer Science and Engineering (Data Science)

Karukutty, Kerala

- Relevant Coursework: Data Structures and Algorithms (DSA), Python Programming, Database Management Systems (DBMS), Operating Systems, Machine Learning (ML)

CERTIFICATIONS AND TRAINING

MERN Stack Development Internship Certificate – MERN (MongoDB, Express.js, React.js, Node.js) ICT Academy of Kerala, Jul 2024