

Chetan Chhetri

+91-8167680360 · chhetrichetan45@gmail.com · Siliguri, WB, India
linkedin.com/in/chetanchhetri · github.com/Chetanchhetri · leetcode.com/u/2rPvyovowx
AI/ML Engineer · Backend Systems · RAG & Multi-Agent Pipelines · Production AI Observability

EDUCATION

Siliguri Institute of Technology, Siliguri, West Bengal

Expected July 2028

B.Tech in Computer Science — AI & ML Specialization | CGPA: 7.15

PROFESSIONAL EXPERIENCE

AI/ML Developer Intern — *Hansraj Ventures*

Apr 2026 – Present

Remote · FinTech, LegalTech & Social Media domains

- Engineered the full backend for **FamioX Employee Suite** — a Teams-style real-time collaboration platform (WebSocket channels, presence, threaded messaging) + an **Employee Activity Monitoring system** with async event logging to MongoDB; serving concurrent sessions with sub-100ms message delivery.
- Architected **Risk Minor**, a production multi-agent risk intelligence platform: **hybrid BM25 + semantic vector search (Milvus)** feeds a RAG pipeline for natural-language queries over live risk data; a **finetuned sLLM** auto-generates structured compliance reports; wired **RAGAS eval scores (context precision, answer faithfulness)** into a **Langfuse dashboard** for real-time hallucination monitoring across all RAG responses.
- Built a multilingual NLP translation pipeline for risk document ingestion — preprocessing, **language detection**, and **model-routing** to locale-appropriate translation models — enabling cross-border regulatory coverage.
- Deployed an sLLM-powered vision pipeline for FamioX generating **automatic photo captions and content tags** at inference time, eliminating manual content-labelling overhead and cutting moderation cost by ~60%.
- Engineered async FastAPI REST microservices across all services with Docker-based CI/CD; reduced response latency by **~40%** and maintained 99%+ uptime across staging and production deployments.

Software Developer Intern — *ACCESCO Living*

Dec 2025 – Mar 2026

Remote · FinTech & Smart Home domain

- Built an **Inventory Monitoring system** paired with a **demand-forecasting model** (Scikit-Learn, time-series regression on current + prior-week sales data) — reduced stockouts by 35% and overstock waste by 22%.
- Designed a **Household Budgeting engine** that auto-allocates income into Needs, Expenses, Fun, and Savings using city cost-of-living indices, family size, and income — delivering personalised financial plans across 4 city tiers.
- Deployed statistical anomaly-detection pipelines on time-series telemetry, cutting false-positive alerts by **28%**; refactored legacy monolith into async FastAPI microservices with Docker-based reproducible testing, reducing end-to-end latency by **40%**.

PERSONAL PROJECTS (OPEN-SOURCE — PINNED ON GITHUB)

Fact Check— Multi-Agent Fake News Detection Engine | *Python, Streamlit, Ollama, Qwen, Phi-3, LLaMA, RAGAS*

- Queries 3 local LLMs in parallel with shared-task prompts; aggregates verdicts via a **consensus scoring layer** to suppress single-model hallucination bias. Integrated **RAGAS evals (answer correctness, faithfulness)** tracked via **Langfuse** to measure fact-check accuracy across political and scientific claim sets.

Intelligent Proxy Gateway (IPG) | *FastAPI, Flask, Ollama, Python, Chrome Extension*

- Four-stage request pipeline** — **prompt reformulation** → **security filtering** → **task triage** → **token optimisation** — routes to a local Ollama model first, cloud LLM only on cost-threshold fallback. Shipped as backend service + Chrome extension; authored architecture paper on the design.

RBI Regulatory RAG Assistant | *Python, FAISS, BM25, FastAPI, Langfuse*

- Hybrid BM25 + FAISS retrieval with cross-encoder reranking, hallucination-detection fallbacks, and **Langfuse tracing per query (retrieval latency, reranker scores, hallucination flags)** — answers compliance queries grounded in RBI regulatory PDFs.

ACHIEVEMENTS & CERTIFICATIONS

Smart India Hackathon (Internal) — Winner · **45-Day DSA Challenge** — 1st Runner-Up · **SIT Code Quest** — 3rd Place · 200+ LeetCode Problems Solved

TECHNICAL SKILLS

LLM & AI Agents: RAG, Multi-Agent Systems, Hybrid Retrieval (BM25 + Semantic), sLLM Finetuning, Prompt Engineering, Ollama, LangChain

AI Observability & Evals: RAGAS (context precision, faithfulness, answer correctness), Langfuse tracing, hallucination detection, eval pipelines

Backend & Systems: FastAPI, Flask, Async Microservices, WebSockets, REST APIs, Docker, CI/CD, GCP, Linux, Git

Data & ML: TensorFlow, Scikit-Learn, Time-Series Forecasting, Pandas, NumPy, Vector DBs (FAISS, Milvus, Qdrant), MongoDB, PostgreSQL, SQL

Languages: Python, Java, C++, SQL, JavaScript

Core CS: Data Structures & Algorithms, System Design, OOP, Distributed Systems, Concurrency, Complexity Analysis