SOURAV TRIPATHY

+91 9337191375 · Bhubaneswar, India · lipuntripathy74@gmail.com
www.linkedin.com/in/sourav-tripathy-astrophile/ · https://github.com/Sourav-Tripathy

EDUCATION

B.Tech, IGIT Sarang, Odisha 2020 – 2024
Electronics and Telecommunication Engineering
Relevant Coursework: Computer Networks, ML, Digital Electronics, Communication Engg., Soft Computing
Project: **Sink-hole Attack Detection using Data from VANET Simulation**

SKILLS

Programming: Python, C
Machine Learning: Supervised/Unsupervised Learning, NLP
Deep Learning: Neural Network (PyTorch, Tensorflow), Transformers
Reinforcement Learning: Policy Gradients, Value-based
Generative AI: LLMs, Retrieval-Augmented Generation (RAG), Context Optimization, Vector Search
Fine-tuning: PEFT (LoRA, QLoRA), SFT, COT
Backend: FastAPI, WebSocket, Async Programming, Docker
Databases: MongoDB, Neo4j, PostgreSQL
Tools & DevOps: Git, Linux, CI/CD

EXPERIENCE

Software Engineer (ML) Aug 2024 – Present
Cognitive View *Bhubaneswar, Odisha*

- Deployed an LLM-powered research assistant to query structured/unstructured data, reducing research friction by **90%** and cutting sales research turnaround from 1 day to under 2 hours.
- Built real-time multi-agent workflows (LangGraph + WebSockets) enabling parallel reasoning, delivering **3x faster LLM interaction** in multi-step tasks.
- Architected a hybrid retrieval system (MongoDB vector search + Neo4j graph queries) with semantic routing, reducing average RAG latency from 12–15s to **4–6s** across 1,000+ test queries.
- Developed a Graph-RAG pipeline with Cypher query generation, rewriting, and prompt tuning, achieving **95% factual accuracy** and cutting hallucinations by **70%** on 300+ benchmark queries.
- Designed an asynchronous FastAPI backend with persistent session memory and modular RAG orchestration layers, scaling to **500+ concurrent sessions** without performance drop.

PROJECTS

Agentic Book Writer (CrewAI). Multi-agent system using CrewAI with agents for ideation, drafting, fact-checking, and refining LLM-generated book content.

Embedding-Crosslinguality Study. Analyzed multilingual embeddings across 6 languages using LLaMA + multilingual-MiniLM. Compared vector alignment and cosine similarity across translated essays and words.

LoRA-based Fine-Tuning of Qwen-1.5. Fine-tuned Qwen-1.5B on a custom instruction set via PEFT LoRA. Demonstrated improved relevance on domain-specific prompts.

Simple Reinforcement Learning Agent. Q-learning agent for FrozenLake-v1. Tuned exploration-exploitation, visualized learning curves, and analyzed epsilon decay strategies.

EXTRA-CURRICULAR ACTIVITIES

- **Write regularly at siliconandsoul.substack.com on AI/ML and tech-philosophy.**
- Follow research in LLMs, retrieval systems, and alignment in personal study time.