Sourav Yadav

Delhi, India | sk.souravrao25@gmail.com | 97284 96993 | linkedin.com/in/souravydv22
github.com/sourav-ydv | Portfolio

Summary

Final-year CSE student at IIIT Senapati, Manipur (2027) who builds and tests AI systems: tool-calling agents, a fine-tuned small model with output checks, and a fault-injection tool. WABI 2026 co-author.

Education

Indian Institute of Information Technology (IIIT) Senapati, Manipur — Aug 2023 – 2027 (Expected)

- B.Tech in Computer Science and Engineering | CPI: 8.85 (6th Sem)

Experience

Research Intern, National Institute of Technology Kurukshetra — Nov 2025 – Apr 2026

Built a Python pipeline to clean and validate an 18,104-genome dataset; trained a coverage-aware **PyTorch** model against SVM, Random Forest, XGBoost and two published baselines (**F1-macro = 0.79**, **AUROC = 0.84**).
Fixed an evaluation leak a reviewer found (embeddings trained on held-out genes) and regenerated all results; co-authored the paper, accepted at **WABI 2026**.
Tools: Python, PyTorch, Scikit-learn, NumPy, Git

Publications

- Co-authored "*Biologically Grounded Explainable Antimicrobial Resistance Prediction Using Knowledge Graphs*", accepted at **WABI 2026**.
- Submitted: "*NeSy-AMR: Neuro-Symbolic Coverage Routing for Explainable AMR Prediction*", under review at **IEEE TCBB**.

Projects

Havoc: Fault-Injection Testing for AI Agents (Python, asyncio, MCP) — github-repo | Report

Built a proxy that injects 6 fault types into an agent's tool calls and grades runs by checking the final database state with 9 safety checks; delta debugging shrinks failures; 383 tests with CI.
Across 411 runs, found false refund confirmations in 25–48% of runs; a one-line prompt change cut them to 0–10% ($p = 0.0004$). A second model did not reproduce it; I reported that and my own checker mistakes.
Tools: Python, asyncio, MCP, SQLite, pytest, GitHub Actions

receipt-router: Fine-Tuned Small Model with Output Checks (Python) — github-repo | Model

Fine-tuned Qwen2.5-1.5B (QLoRA, GGUF, laptop CPU) to extract receipt fields; arithmetic and grounding checks accept its answer or escalate to a 120B model (95.3% of totals right on 150 held-out receipts vs 88.7% for the small model alone).
Found a silent failure: a one-currency model read 43.3% of Malaysian totals off by 100x and the checks accepted 59 wrong answers; training on both formats fixed it. Measured the OCR drop (65.3%). 121 tests.
Tools: Python, Pydantic, FastAPI, pytest, QLoRA, llama.cpp, Tesseract

Trip AI Agent (LangGraph, FastAPI, PostgreSQL) — github-repo | Demo-link

Built and deployed a ReAct agent (LangGraph, Groq) with 7 tools over live flight, hotel, weather and web-search APIs, PostgreSQL state, and an `interrupt()`-based approval step before any (simulated) booking.
Found it inventing fares and airport codes from its traces and added a rule for each.
Tools: Python, FastAPI, LangGraph, PostgreSQL, React

Technical Skills

LLMs & Agents: LangGraph, LangChain, Tool-Calling (ReAct) Agents, MCP, Human-in-the-Loop, RAG, Structured Output, Prompt Engineering, Groq / Gemini / Cohere APIs
Model Training & Evaluation: PyTorch, QLoRA Fine-Tuning (Unsloth), GGUF / llama.cpp, Hugging Face Hub, Scikit-learn, XGBoost, Held-out Evaluation (exact match, F1, AUROC), Output Checks
Backend & Testing: Python (FastAPI, Pydantic, asyncio, Pandas, NumPy), SQL, PostgreSQL, pytest, GitHub Actions, Fault Injection
Core CS & Tools: Data Structures & Algorithms, OOP, C++, C, JavaScript, Git, GitHub, Docker, Tesseract

Certifications & Languages

NPTEL Elite + Gold: Business Intelligence & Analytics (IIT Madras, 95%, Top 5%); Information Retrieval (IISER Kolkata, 92%, Top 2%); Competitive Programming (IIT Gandhinagar, 90%). **Elite + Silver:** Blockchain (IIT Kharagpur, 79%).
Korean Language: TOPIK Level 1 (Oct 2025)