NISHIKA YADAV

+91-8840139142 · nishikayadav26@gmail.com · linkedin.com/in/nishika-yadav · github.com/nishika26 · https://nishfish.vercel.app/

PROFESSIONAL SUMMARY

Python-based Backend & AI Engineer with 1.5+ years building production LLM platforms, multi-provider orchestration, RAG pipelines, vector stores, and fine-tuning large language models. Comfortable owning full backend systems from schema design to deployment and deeply excited by data in all its forms and eager for research opportunities.

EDUCATION

Madhav Institute of Technology & Sciences, Gwalior | B.Tech Electrical Engineering | 2020 - 2024

WORK EXPERIENCE

Tech4Dev (Project) | *Backend & AI Engineer — Kaapi,* Dec 2024 – Present

- Architected a **Python-based** multi-provider LLM gateway **(OpenAI, AWS Bedrock, Gemini)** in **FastAPI + SQLModel**, hardened the provider-agnostic platform with rate limiting and SSRF-protected webhook callbacks.
- Built a production document upload to optional document transformation to vector store pipeline with asynchronous support via **Celery + Redis + gevent**. Experimented with multiple file search setups including **AWS Bedrock + OpenSearch + S3** and **Google Gemini file search** to evaluate best-fit RAG setups for NGO clients.
- Contributed to a guardrails microservice (a separate FastAPI service) with autodiscovery and service-to-service auth, enabling safety validators to be added without modifying core platform code.
- Contributed to a production LLMOps frontend using **Next.js 16, React 19, and TypeScript**, delivering features including guardrails configuration UI, document management with upload progress, knowledge base creation, and speech-to-text evaluation.
- Engineered complete pipeline for fine-tuning LLM Models and evaluating those models with **OpenAI** — stratified data splits, background job tracking, and JSONL preprocessing.

Dalgo (via Tech4Dev) | *Data Engineer, Sept 2024 - Dec 2024*

- Built **DBT transformation pipelines** converting raw NGO field data into analytics-ready models, directly powering dashboards used by the Jal Jeevan Mission programme.
- Delivered **Apache Superset dashboards** tracking key mission metrics, enabling data-driven decisions for teams previously relying on manual reporting.

Calfus Inc. | *AI Engineering Intern* Feb 2024 – Aug 2024

- Delivered a production chatbot **(Ollama + LlamaIndex + LangChain)** that let customers query their cash flow database in natural language; integrated directly with **PostgreSQL**, replacing a 3-step manual reporting workflow.
- Fine-tuned Llama 2 and CodeLlama for Text-to-SQL using **SFTTrainer**, **ORPOTrainer**, and **DPOTrainer** — explored **QLoRA 4-bit quantisation (GGUF/GPTQ)** to run 7B parameter models on constrained hardware.

PACTA | *Research & Data Analyst Intern* Dec 2023 – Feb 2024

- Produced daily data analysis reports using **Python** and **Tableau**, surfacing trends in **Tamil Nadu's disability data** for the research team's decision-making.

Indian Institute of Information Technology | *ML Research Intern,* May 2023 – Jun 2023

- Trained **UX-Net**, a modified U-Net architecture combining convolutional and recurrent processing for low-latency speech separation, and applied it to speech separation and acoustic echo cancellation tasks in **PyTorch**

TECHNICAL SKILLS

AI / LLM: LLM fine-tuning (SFT, ORPO, DPO, QLoRA), LangChain, LlamaIndex, LangGraph, Agentic AI, MCP Servers, Langfuse, Transformers, RLHF, Guardrails AI, OpenAI, Anthropic, AWS Bedrock, Gemini, Ollama, HuggingFace
Cloud & Infra: AWS (ECS, EC2, CloudWatch), Docker, GitHub Actions CI/CD, Vercel
Full-Stack: FastAPI, SQLModel, SQLAlchemy, Pydantic, Alembic, Celery, Redis, SlowAPI, PostgreSQL, RESTful APIs, Webhooks, Async Python, Pytest, Next.js 16, React 19, and TypeScript
ML / Data: PyTorch, TensorFlow, Keras, Scikit-learn, Pandas, NumPy, CNNs, RNNs, OpenCV, DBT, Apache Superset, Tableau
Languages & Tools: Python, C++, SQL, Typescript, Git, DBeaver, Jupyter, Neo4j, Cursor

KEY PROJECTS

Llama 2 Fine-Tuning (QLoRA) github.com/nishika26/finetuning_llama2

- Fine-tuned Llama 2-7B on a free Colab T4 GPU using 4-bit QLoRA precision, reducing VRAM footprint by ~75% with no meaningful accuracy loss. Published multiple fine-tuned model variants on Hugging Face as well as on Ollama

BLOGS

Kaapi Guardrails: A Tattle-Tech4Dev Collaboration for AI Safety projecttech4dev.org
AI Platform Building Steadily: Khopoli Sprint Reflections projecttech4dev.org

CERTIFICATIONS & COMMUNITY

- Tech Volunteer - Save Mumbai Mangrove (April 2026 - Present)
- PEDP Data Science for Social Impact - Ashoka University (2025-2026)
- Summer Analytics 2023 — Consulting & Analytics Club, IIT Guwahati
- Volunteer — People+AI (March 2024 – Aug 2024)
- Neo4j & LLM Fundamentals — Neo4j (2024)