ABHEY TIWARI

Noida, Uttar Pradesh |+91 9810440326 | abheytiwarikvs@gmail.com | LinkedIn: @abheytiwari | GitHub: @AbheyTiwari

SUMMARY

AI/ML Engineer building production-ready LLM systems. Deploy RAG and local LLM (Ollama) pipelines on GCP; build real-time FastAPI backends with Docker, Nginx, and systemd.

SKILLS

- **AI/ML:** LLMs, RAG, embeddings, vector search, NLP, CNNs, ML
- **Frameworks:** FastAPI, Flask, Django, Streamlit, LangChain, LlamaIndex, Ollama
- **ML Tools:** Scikit-learn, TensorFlow, PyTorch
- **Languages:** Python, C++, JavaScrip
- t**Infra:** Docker, Git, Linux, REST, WebSockets, WebRTC

EXPERIENCE

Tech, Research & Innovation Intern — Draupadi Dream Trust
Jul 2025 – Aug 2025

- Researched Yamuna River (religious, cultural, environmental) and analyzed Delhi Jal Board datasets to produce stakeholder-ready reports.

Frontend Developer (Remote) — BeyondRiffs
Sept 2024 – Dec 2024

- Rebuilt major React components prior to launch; learned and applied React rapidly to stabilize production.
- Coordinated with design and backend to deliver a reliable release under tight timelines.

Website Developer — Mangalam Valley Resort
Jul 2023 – Nov 2023

- Delivered booking-enabled website with integrated payments and SSL.
- Managed hosting, domain, and deployment workflows.

PROJECTS

Maitri AI — real-time emotionally aware companion (100% local)
DeepFace, Ollama, FastAPI, WebSockets

- Privacy-first mental-health assistant with local emotion detection and contextual dialogue generation.
- Integrated real-time emotion recognition with local LLMs; zero cloud data transmission.

INDICA v1.0 — modular, voice-controlled personal cognitive assistant
Python, RAG, Gemini API, embeddings, asyncio, Playwright

- Built modular voice agent (STT → intent → dispatcher → skills).
- Implemented dual memory (short-term + semantic embeddings) and RAG workflows for research, planning, and code generation.

Cancer Detection AI — lightweight multimodal diagnostics
TensorFlow, PyTorch, XGBoost

- Hybrid CNN + ML approach for MRI and skin-lesion analysis optimized for low-resource devices.
- Deployable on standard laptops; supports interpretable predictions.