# Vivek Kumar Gupta

Kushinagar, Uttar Pradesh, India

+91-9151429036 vivekgupta3749@gmail.com Portfolio Linkedin github.com/vivek34561

## Summary

AI Engineer specializing in building intelligent agents that interact with APIs, browsers, and databases. Experienced in developing LLM-based systems using LangChain, RAG, and FastAPI. Skilled in integrating local inference with Ollama and fine-tuning models using LoRA/QLoRA and quantization for performance optimization. Strong focus on end-to-end AI product engineering with Supabase, Docker, and MLflow in fast-paced, production-oriented environments. Certified in Oracle AI Foundations and Generative AI (LangChain, HuggingFace).

## Education

**Indian Institute of Information Technology (IIIT) Bhagalpur** **2023 – 2027**

*B.Tech in Mathematics and Computing – CGPA: 7.75 (absolute)* *Bhagalpur, Bihar, India*

## Experience

**Outlier – AI Response Evaluator (Freelance)** **Mar-July 2025**

*Evaluated AI model responses for accuracy, coherence, and prompt alignment in English, Hindi & code.* *Remote*

- Analyzed and rated multilingual LLM outputs, identifying hallucinations and logical errors to enhance response quality.
- Contributed to improving LLM response quality through structured feedback and prompt optimization techniques.

## Projects

**AI-Based Farmer Query Support and Advisory System** | Live Demo | Video Demo | GitHub **Sep 2025**

- Built an agentic advisory system combining RAG, FastAPI, and Streamlit to handle farmer queries in 8 languages.
- Integrated weather, crop, and price APIs with persistent conversational memory for contextual responses.
- Implemented crop recommendation using a scikit-learn pipeline over 7 agronomic inputs (N, P, K, temperature, humidity, pH, rainfall) with cached model loading.
- Designed modular FastAPI endpoints enabling API-based agent interaction and database logging.

**CardioCare AI** | Live Demo | Video Demo | GitHub **April 2025**

- Built a low-latency (250ms) FastAPI + Streamlit system for heart disease prediction and multilingual recommendations.
- Integrated Groq LLM for multilingual diet plans, lifestyle guidance, and doctor notes (English, Hindi, Tamil, Bengali), boosting user satisfaction.
- Experimented with Groq and quantized local models to optimize inference speed and memory footprint.
- Added structured logs and automated health monitoring with MLflow integration.

**AI-Powered Interview Preparation Tool** | Live Demo | GitHub **May 2025**

- Built an interactive agent using LangChain and FAISS that performs resume screening and real-time AI interviews.
- Integrated prompt memory, retrieval, and context caching to maintain conversational continuity.
- Used local inference with Ollama and fine-tuned lightweight models (QLoRA) for task-specific QA optimization.
- Integrated real-time voice features with TTS via GPT-4o-mini-tts and STT via Whisper-1, enabling partially hands-free mock interviews with 10 questions, automated feedback, and integrated job search via the Adzuna API.

## Technical Skills

**Languages & Databases:** Python, C++, MySQL
**Frameworks & Libraries:** FastAPI, Pandas, NumPy, Scikit-learn, TensorFlow
**AI & ML:** LoRA/QLoRA fine-tuning, Quantization, Transfer Learning, Supervised/Unsupervised Learning, Deep Learning, Transfer Learning, Model Optimization
**NLP & Generative AI:** Ollama, LangChain, LangGraph, RAG, HuggingFace, Pinecone, FAISS, ChromaDB
**MLOps & DevOps:** MLflow, DVC, Docker
**Tools:** Git/GitHub, Jupyter, VS Code, BeautifulSoup