

# Vaibhav Gupta

Alwar, Rajasthan | vaibhavgupta1148@gmail.com | +91-7424809555 | linkedin.com/in/vaibhav-gupta-b929ba237

## SUMMARY

AI/ML Engineer with 3+ years of experience building production-grade Agentic AI systems and Generative AI pipelines. Expert in multi-agent frameworks (LangGraph, LangChain), LLMs (Llama 3, GPT-4, Qwen), and RAG systems. Proven track record of delivering scalable AI solutions with measurable business impact.

## EXPERIENCE

### Machine Learning Engineer | Glarus Technology

Delhi | Mar 2024 - Present

- Architected autonomous multi-agent framework using Llama 3.2B & Groq API for NL-to-code generation, achieving 95% accuracy in HTML5/Tailwind output.
- Engineered parallel agent pipelines for UI generation, reducing content latency by 40% vs sequential baselines.
- Built production virtual try-on system integrating SCHP, MODNet & PIFuHD for photorealistic garment rendering.

### AI Engineer | Softage.ai

Delhi | Jun 2023 - Mar 2024

- Designed high-performance RAG pipelines with LangChain & Pinecone, improving response accuracy by 25% for enterprise chatbots.
- Built Large Action Models (LAMs) for RL-based autonomous decision-making in predictive modeling scenarios.
- Optimized ML inference pipelines across NLP/CV domains, reducing latency by 35% for recommendation systems.

### AI/ML Research Intern | Gilbert Research Center

Remote | Jan 2023 - May 2023

- Transitioned 3 academic prototypes (Sentiment Analysis, Diabetes Prediction, Text Classification) to production demos for healthcare & ed-tech.

## KEY PROJECTS

### Agentic Resume Optimization System (LangGraph, FastAPI, Groq, Tavily, Streamlit)

- Built multi-agent system with 4 specialist agents for automated resume optimization; improved ATS match rates by 45% (33% to 78%).
- Implemented real-time company research & JD gap analysis using Groq (Qwen 32B) and Tavily API.

### Enterprise Conversational AI (LangChain, GPT-4, Pinecone, FastAPI)

- Developed autonomous chatbot handling multi-domain queries with sub-200ms latency using optimized RAG pipelines.

### MCQ Generator (PyTorch, Hugging Face, FastAPI)

- Fine-tuned LLM for auto-generating exam-grade MCQs, deployed for ed-tech platforms with 92% quality rating.

## TECHNICAL SKILLS

Languages:	Python, SQL, JavaScript, Java, R, HTML5/CSS3
AI/ML:	LangGraph, LangChain, PyTorch, TensorFlow, Hugging Face, scikit-learn, FastAPI
GenAI/LLMs:	Multi-Agent Systems, RAG, Prompt Engineering, GPT-4, Llama 3, Claude, Qwen
Cloud/Tools:	AWS, GCP, Docker, Git, Pinecone, Weaviate, Streamlit, MLflow

## EDUCATION

### Bachelor of Technology (B.Tech) in Computer Science

2019 - 2023

## CERTIFICATIONS

AI Engineer Certificate (Pro5.ai) | ML & Data Science (GeeksforGeeks) | Advanced Python (RKCL, Grade A)