Keshav Garg

Delhi, India | +91 9873501040 | gargkeshav504@gmail.com | Linkedin | Portfolio

Experience

AI Engineer Apr 2025 – Present
Kroolo *Delhi, India*

- Architected and built a complete Enterprise Search system from the ground up, managing the full backend lifecycle—from data ingestion and preprocessing to a search pipeline using traditional machine learning methods—achieving performance comparable to an industry leader.
- Revamped Chat with Anything module using advanced text extraction algorithms, reducing processing latency by 50%.
- Rebuilt AI agent ecosystem with 250+ integrated tools, transforming prompt-engineered agents into sophisticated agentic systems.
- Led technical design reviews and managed full project lifecycle from requirements to production deployment.
- Delivered comprehensive system documentation for enterprise-grade AI agent platform.
- **Skills:** Enterprise Search, Backend Development, Agentic AI , Multi-Agent Systems , Tool Integration , Performance Optimization , Technical Leadership.

AI Engineer Nov 2024 – Apr 2025
GenAI Protos *Delhi, India*

- Engineered multimodal Deep Research Multi-Agent System with 97% accuracy for fact-checking and misinformation detection.
- Developed NLP2SQL system using RAG architecture, achieving 98% query accuracy with automated chart generation, deployed via WhatsApp.
- Optimized system for on-premises deployment using open-source models, meeting enterprise security requirements.
- Built YOLO-based Object Detection Model with custom fine-tuning, eliminating 100% manual CCTV monitoring.
- **Skills:** Multi-Agent Systems, RAG, NLP2SQL, Multimodal AI, YOLO, Fine-tuning, On-premises Deployment

AI/ML Intern May 2024 – July 2024
Agnisys Technology *Noida, India*

- Implemented a Retrieval-Augmented Generation (RAG) pipeline, significantly improving data retrieval accuracy and contextual response generation.
- Created an automated dashboard using user behavior data, eliminating manual data entry and reducing processing time.
- Collaborated with cross-functional teams to deliver scalable, high-impact solutions, improving project outcomes and client satisfaction.
- **Skills:** RAG, NLP, Dashboard Automation, Python, Data Visualization, Collaboration, Streamlit, Pandas

Projects

Ra.One – Human-like AI WhatsApp Assistant | Demo

- Developed and deployed an AI-powered WhatsApp assistant using LangGraph for real-time, context-aware conversations.
- Implemented dual-memory architecture with PostgreSQL (short-term) and Pinecone (long-term) for memory-augmented dialogue.
- Integrated LLMs (Groq), ElevenLabs voice modules, and multimodal inputs.
- **Skills:** Python, LangGraph, LLMs, NLP, Pinecone, PostgreSQL, RAG, FastAPI, WhatsApp

Enterprise Support Automation System | Demo

- Developed autonomous AI support agent reducing manual workload by 80%.
- Engineered web scraping pipeline using Selenium to extract 100+ help articles with summarization and embedding for retrieval.
- Built LLM-powered agent with integrated search, generation, and verification tools.
- **Skills:** Python, LangGraph, Web Scraping, RAG

LegalInsight | Demo

- Developed legal document analysis platform using RAPTOR for hierarchical document understanding and automated PDF summarization.
- Implemented interactive "Chat with PDF" using RAPTOR-based retrieval and LangChain, reducing review time by 70%.
- Architected privacy-first deployment using open-source LLMs and PyMuPDF4LLM for optimized data extraction.
- **Skills:** Python, RAG, LangChain, Ollama, FastAPI

Education

Bachelor of Technology (Computer Science) 2021 – 2025
Amity University Noida *Noida, India*

Technical Skills

Programming: Python, SQL
Machine Learning & AI: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Multi-Agent Systems, Fine-tuning, Deep Learning, Natural Language Processing (NLP)
ML Frameworks: PyTorch, Scikit-learn, Pandas, LangChain, LangGraph, Hugging Face
Backend Development: FastAPI, AWS