

GARIMA KUMARI

Delhi, India (IST, flexible overlap with US time zones) | +91-7061868848 | garimakumari265@gmail.com | github.com/garimakumari44 | linkedin.com/in/garima-singh-4122331bb

SUMMARY

AI engineer building production LLM systems — RAG pipelines, multi-agent workflows, and document intelligence — end to end from FastAPI backends to Next.js frontends. Four deployed AI platforms with live demos. Open to remote roles at US startups; available for contract-to-hire.

EXPERIENCE

AI/ML Engineer Intern — ITJOBXS (AI career-prep platform)

May 2026 – Jul 2026

- Built LLM-powered features with RAG and REST APIs for an AI-driven career preparation platform; automated information-processing and personalization workflows in production.

Freelance Full-Stack Developer — Independent

Nov 2024 – Jun 2025

- Designed and deployed client applications end to end: backend services, PostgreSQL databases, auth, and REST APIs.

PROJECTS

AI Research Intelligence Platform — [Live Demo](#) · [GitHub](#)

2026

- Production-grade research assistant that ingests, retrieves, and synthesizes AI/ML papers using dense + BM25 hybrid retrieval with RRF fusion, metadata filtering, and neural reranking.
- Adaptive RAG with evidence attribution and traceable citations, automated report generation, and knowledge-graph-based research-gap detection. LangGraph orchestration, FastAPI, PostgreSQL, FAISS, Redis, Docker, Next.js/TypeScript.

Orion — Multi-Agent AI Research System — [Live Demo](#) · [GitHub](#)

2026

- Multi-agent system coordinating specialized analyst agents through a planner-driven DAG execution engine — tool-augmented reasoning, hybrid retrieval, evidence validation, and critic evaluation.
- Applied to equity research (company, financial, macro, valuation, risk analysts) with MCP-enabled tools and full execution observability. FastAPI, LangGraph, PostgreSQL, Redis, Docker, Next.js.

Enterprise Document Intelligence — [Live Demo](#) · [GitHub](#)

2025

- B2B platform for document understanding: OCR, layout-aware extraction, semantic retrieval, and citation-backed Q&A over unstructured business documents.
- Built a reusable AI Workflow Orchestration engine — multi-stage pipelines, approvals, retries, versioning, execution states, MinIO object storage.

RLHF Self-Improving Q&A — [GitHub](#)

2025

- Implemented a preference-learning pipeline in PyTorch: reward model trained on human preference data, policy optimized with PPO to iteratively improve answer quality.

SKILLS

Languages: Python, TypeScript, JavaScript, SQL · **LLM/AI:** RAG (hybrid, adaptive), reranking, multi-agent systems (LangGraph, MCP), LLM evaluation, prompt engineering, RLHF/PPO

ML: PyTorch, Transformers, Scikit-learn, NLP · **Systems:** FastAPI, PostgreSQL, Redis, FAISS, BM25, Docker, MinIO, React/Next.js, Tailwind, Git/Linux

EDUCATION

University of Delhi — Miranda House

2023 – 2026

B.A. (Hons.) Economics, Minor in Computer Science — Statistics, Econometrics, Mathematical Economics, Data Structures & Algorithms, DBMS, AI

WRITING & COMMUNICATION

Published poet (Poet's Choice, American Jewish Magazine; multiple award wins; publishing-house offer for original poems). Strong written communication for async, remote-first teams.