Ganesh Agrahari

ganeshagrahari108@gmail.com
+91 9044232872
Lucknow, India
My-Portfolio
LinkedIn
GitHub

PROFILE

AI Engineer with hands-on experience building real-world, **production-grade systems**. Developed serverless **AI microservices, Retrieval-Augmented Generation (RAG) systems using embedding models, and scalable vector search pipelines. Experienced with Azure Functions, GPT-4, Elasticsearch, LangGraph, and n8n-based automation workflows.** Strong foundation in **Python, Machine Learning, NLP, and cloud architecture, with proven experience self-hosting Elasticsearch and n8n on Azure Virtual Machines** to deliver fast, reliable, and cost-efficient AI solutions at scale.

EDUCATION

BCA Data Science & Artificial Intelligence(in collaboration with IBM)
BBD University
08/2023 – 09/2026 | Lucknow, India
Last year SGPA : 8

Intermediate(PCM)
SVM Inter College Ntpc
2022 | Raebareli, India
Percentage : 82%

ACTIVITY

HackerRun Problem Solving
03/2025 – present

GitHub Streak Maintenance
10/2024 – present

Member of GDG(Google's Developer Group)
Lucknow, India

Member of Technical team
BBD University, Lucknow, India

LeetCode Problem Solving

AI/ML Workshops Attendee
Lucknow, India

PROFESSIONAL EXPERIENCE

AI Engineer Intern
Edubuk
08/2025 – Present

- Built an **AI-powered JD–CV matching system** combining **vector similarity for accuracy** and **LLM-based analysis for contextual understanding**, enabling high-precision candidate–job matching at scale.
- Designed and deployed a **serverless architecture on Azure using Azure Functions**, including migration from an earlier **AWS Lambda–based serverless setup**, ensuring improved scalability and operational consistency.
- Implemented a **Retrieval-Augmented pipeline** using **OpenAI text-embedding-3-large** for semantic embeddings and **GPT-4** for deep contextual evaluation between job descriptions and resumes.
- Migrated the search infrastructure from **AWS OpenSearch to self-hosted Elasticsearch on Azure Virtual Machines**, improving system control, flexibility, and long-term cost efficiency.
- Engineered backend services for **resume ingestion, JD processing, and matching orchestration**, delivering fast, reliable responses suitable for production-grade hiring workflows.

AI/GenAi Intern at QTechSolutions
AICTE
03/2025 – 08/2025

- Working on a **real-world healthcare site** to develop and integrate an **AI-powered chatbot for doctor consultations, medicine delivery**, and prescription recommendations.
- Implementing **NLP and machine learning** techniques to enhance chatbot accuracy, ensuring seamless and efficient user interactions in a scalable AI-driven system.
- Leveraging **Retrieval-Augmented Generation (RAG)** to provide dynamic, context-aware responses by combining LLM capabilities with real-time retrieval from healthcare databases.

Data Science Intern
Unified Mentor
11/2024 – 01/2025

- Implemented a **text classification pipeline,** improving document categorization accuracy by 20%
- Developed an NLP-based climate change analysis system using **sentiment analysis and topic modeling** to extract insights from global news articles.

PROJECTS

TruJobs – AI Recruitment System (Edubuk)

- Designed an **AI-driven JD–CV matching workflow** that evaluates both **semantic similarity** and **contextual relevance**, improving hiring signal quality beyond keyword-based matching.
- Implemented an **embedding-based retrieval layer** using **OpenAI text-embedding-3-large**, enabling accurate semantic comparison between resumes and job descriptions.
- Integrated **GPT-4** to perform contextual analysis and reasoning over retrieved candidates, enhancing match quality for complex and non-obvious skill relationships.
- Built **event-driven APIs** for resume ingestion, JD uploads, and match retrieval, optimized for high-throughput processing in a production environment.
- Enabled **real-time schema flexibility and fast search operations** using **self-hosted Elasticsearch**, supporting continuous data evolution without service disruption.

CERTIFICATES

- Data Science Level 1 – IBM
- Analytics in IBM Cognos
- Machine Learning–Udemy
- Cyber Security – Microsoft
- App Development -BBDU
- NoSql & DbaaS 101 -IBM
- Data Science 101 -IBM
- Predictive Modeling Fundamentals I - IBM
- Python 101 for Data Science - IBM

Face Recognition Attendance System :

- Designed and deployed a real-time face recognition attendance system using **Python, OpenCV, and CNN, achieving 95%+ accuracy** and r**educing manual attendance tracking efforts by over 40%. Integrated automated data logging and Excel** export for seamless record-keeping.

Nextjs Portfolio site :

- **Interactive Portfolio Website** – Built using **Next.js and deployed on Vercel, featuring a responsive design** and an **AI-powered chatbot** for seamless user interaction. Showcases my skills, projects, and experiences with a dynamic and engaging UI.

SKILLS

Programming & Scripting:

- Python, JavaScript
- Async Programming, REST API Development

AI/ML & LLM Engineering

- Machine Learning, Deep Learning, NLP, LLMs, RAG
- Vector Search (Elasticsearch), Multi-Vector Embeddings (3072-dim)
- Hybrid Scoring (GPT-4 + vector similarity), Prompt Engineering
- Frameworks: PyTorch, LangChain, LangGraph Scikit-learn, OpenCV
- Aumation Tools : N8N

Cloud & Backend Architecture

- **Azure (current):** Functions, IAAS(VMs), API Management, Blob Storage, App Insights, Azure OpenAI (GPT-4/Embeddings)
- **AWS (previous version of TruJobs):** Lambda, Amazon Bedrock (Claude 3 Haiku, Titan Embeddings), API Gateway, S3, OpenSearch
- Microservices, Serverless Architecture, Async Pipelines
- CI/CD, Git, GitHub

Data Engineering & Analytics

- Data Processing, ETL, Feature Engineering
- Power BI, Jupyter Notebook, Matplotlib, Seaborn

Databases & Storage

- SQL
- Elasticsearch (vector + keyword fields)
- AWS S3, Azure Blob Storage