

# Gaurav Patel

+91 9173440539 — patelgaurav454@gmail.com — LinkedIn — GitHub — Portfolio Website

## Professional Summary

---

Data & AI Engineer with 3+ years of experience delivering production-grade Generative AI, machine learning, and data platforms end to end. Specialized in LLM/SLM fine-tuning (LoRA, PEFT), agentic and multi-agent systems (LangGraph, LangChain), and RAG pipelines over vector databases, supported by strong data engineering (ETL/ELT, Airflow) and MLOps practice. Builds scalable Python microservices (FastAPI, REST) and deploys across Azure, GCP, and on-prem environments, consistently cutting manual effort by 40–90% through automation, evaluation-driven model development, and cost-efficient inference.

## Education

---

### PG Diploma in Big Data Analytics (PG-DBDA)

Aug 2024 – Feb 2025

CDAC Kharghar, Navi Mumbai

### Bachelor of Engineering in Computer Engineering

Aug 2020 – May 2024

Dr. D.Y. Patil College of Engineering and Innovation, Talegaon (SPPU Pune)

## Experience

---

### Capgemini

Feb 2026 – Present

AI Engineer

- **Test Automation Pipeline:** Architected a Python pipeline orchestrating **LLM agents, Tesseract OCR, and schema-driven JSON/YAML parsing** to convert Word test specifications into device-ready Excel scripts, cutting manual script creation time by **~40%**.
- **Merlin ITSM Assistant:** Shipped an end-to-end IT support agent on a **fine-tuned Mistral model (LoRA)** served via **FastAPI**, with role-based user/admin consoles and full ticket-lifecycle automation (auto-resolution, escalation, confirmation, closure); built a JavaScript frontend with live polling, threaded user-IT chat, and **in-browser predictive autocomplete** (unigram/bigram language model).
- **AI-Powered Resume Screening System:** Built a full-stack platform (**80% accuracy, 70% faster screening**) using a hybrid **TF-IDF, embeddings, and ChromaDB** retrieval pipeline for semantic JD matching, with a multi-page **Streamlit** interface for scoring, keyword search, and re-ranking; adopted internally by the RMG team.
- **Multi-Agent RAG System (LangGraph):** Designed a **5-agent architecture** with query routing, context-aware generation, and optimized chunking/re-ranking strategies to maximize retrieval relevance and reduce hallucination.
- **.NET Analyzer & Upgrader Agent:** Delivered an autonomous agent adopted by **3 teams**, improving legacy upgrade efficiency by **30%** through automated dependency resolution, security scanning, and API modernization workflows.
- **Agentforce (Einstein AI):** Developed conversational AI and ML forecasting features including real-time product retrieval, automated reminder systems, and customer-facing prediction interfaces.

### Capgemini

July 2025 – Jan 2026

Data Scientist

- Engineered custom training pipelines to fine-tune **SLMs, LLMs, and classical ML models** on structured and unstructured data (text, JSON, XML) using **LoRA and parameter-efficient fine-tuning (PEFT)**.
- Optimized training and inference for **resource-constrained hardware** (i5/AMD CPU, 16GB RAM, 8GB VRAM) via quantization and memory-efficient batching, enabling self-contained GenAI applications without GPU clusters.
- Architected a **secure local model registry** to deploy fine-tuned models as autonomous AI agents, achieving **100% offline on-prem inference** and full data privacy compliance.
- Built a **modular Python framework (SOLID, OOP)** automating data preprocessing, training, and evaluation workflows, reducing manual effort by **90%** and accelerating experimentation cycles.
- Implemented **MLOps** practices: automated versioning, experiment tracking, and rollback for model artifacts and LoRA adapters across environments.
- Led a **Proof-of-Concept (PoC)** benchmarking model accuracy, latency, and footprint; documented reproducible workflows and deployment via LM Studio.

*Software Engineer*

- Performed **LCM trend analysis** on 6 months of historical data to surface underperforming projects across India, enabling data-driven remediation decisions.
- Built **interactive Power BI dashboards** with drill-down KPIs to visualize regional performance trends and highlight actionable variances for leadership.
- Automated document and reporting workflows with **Python (python-docx, python-pptx)**, extracting structured data from Word and PowerPoint sources and integrating it into **RPA workflows** for a major travel technology client.

**Freelance****Independent Work**

Jan 2023 – April 2025

*Data Engineering, AI/ML & Web Development*

- Delivered end-to-end solutions across **data engineering, AI/ML, automation, and web development** for clients and independent products over **2 years**.
- **Raghuvanshi Rubber Industries** | **Live Project**: Built and maintain a responsive company website in **HTML, CSS, and JavaScript** with Google Maps and dynamic UI components, improving engagement and search visibility.
- **Enterprise Data Pipeline** | **GitHub**: Developed a scalable **ETL/ELT** pipeline with **Python, dbt, Snowflake, Airflow, and Docker**; improved processing speed by **45%** through model optimisation and warehouse performance tuning.
- **MailSweep Web App** | **GitHub**: Built a secure **Streamlit** app automating Gmail cleanup via **OAuth 2.0**; reduced cleanup time by **70%** using rule-based filtering and **Airflow** scheduling.
- **Sonar Rules Extractor** | **GitHub**: Wrote a concurrent **asyncio** scraper extracting SonarQube rules into a structured JSON database, achieving **3× faster** extraction via semaphore-controlled async requests.

**TECHNICAL SKILLS**

- **Programming:** Python, SQL, JavaScript, Bash
- **AI & Machine Learning:** Machine Learning, Deep Learning, NLP, LLMs, SLMs, Transformers, Fine-tuning (LoRA/PEFT), Quantization, Predictive Modeling, Model Evaluation
- **Generative AI:** LangChain, LangGraph, RAG, Agentic AI, Multi-Agent Systems, Embeddings, Vector Search, Prompt Engineering, MCP
- **Data Engineering:** ETL/ELT, Apache Airflow, dbt, Snowflake, Pandas, NumPy, Data Modeling, Data Pipelines
- **Databases:** ChromaDB, FAISS, PostgreSQL, MySQL, MongoDB
- **Cloud & MLOps:** Azure, Azure OpenAI, GCP, OpenAI API, Docker, Git, GitHub, CI/CD, Model Versioning, Monitoring
- **Frameworks & BI:** FastAPI, Streamlit, Chainlit, REST APIs, PyTorch, Hugging Face, Scikit-learn, Power BI, RPA Integration

**CERTIFICATIONS — Live**

- **GCP Associate Cloud Engineer** — 2026
- **Microsoft Azure AI-900** — 2025