Gaurav Chaudhary

+91 8650119836 | gaurav.chaudhary.865022@gmail.com | linkedin.com/in/gaurav-chaudhary | github.com/rider315 | gauravports.onrender.com | Bengaluru, India

Summary

AI-focused **Full Stack Engineer** with 1+ year of experience building and shipping production **LLM** and web applications end to end. Hands-on with **Retrieval-Augmented Generation (RAG)**, **LLM evaluation**, and **prompt engineering**, alongside **React**, **Node.js**, and **REST API** development deployed on **AWS**. Comfortable operating **customer-facing** as a **forward deployed engineer** – scoping requirements with stakeholders, prototyping rapidly, and taking **Generative AI** workflows from proof of concept into real user workflows.

Technical Skills

Languages: Python, JavaScript, C/C++, Java, SQL, HTML/CSS

AI & Machine Learning: LLM Evaluation, Retrieval-Augmented Generation (RAG), Prompt Engineering, LLM Fine-Tuning, Question Answering, Transformers, Vector Databases, Embeddings, Generative AI, NLP, Unstructured Data, AI Quality Assurance, Model Benchmarking, Pandas, OpenCV, AI/ML

Full Stack & Frontend: React.js, Node.js, Express.js, REST APIs, WebSockets (Socket.io), Redux, HTML/CSS, Responsive UI

Databases & Caching: MongoDB (NoSQL), MySQL, Redis, Firebase, Mongoose, Vector DB, DBMS

Cloud & DevOps: AWS (EC2, S3), Docker, CI/CD, Git, GitHub, Jira, Confluence, Postman, Linux, Operating Systems

Other: Power BI, Android Development (Java, Gradle, Firebase), Embedded Systems (ESP32, AUTOSAR, dSPACE), Data Structures & Algorithms, OOP

Experience

Innodata Jun 2026 – Present

Senior Associate Engineer *Noida, India*

- Evaluate and **fine-tune Large Language Models (LLMs)** across diverse task suites, scoring **response accuracy**, reasoning quality, factual consistency, and instruction adherence to drive measurable model quality gains.
- Assess and improve **Retrieval-Augmented Generation (RAG)** workflows by measuring **retrieval relevance**, context grounding, and hallucination rate, raising the quality of grounded **Generative AI** responses in production.
- Own **AI quality assurance** end to end: reproduce model failures, define **evaluation rubrics**, validate fixes, and partner with engineering teams to harden **workflow reliability** and output quality.
- Automate evaluation and **data validation** pipelines in **Python**, standardizing how results are collected and reported and cutting repetitive manual review effort.
- **Tech:** LLM Evaluation, RAG, Prompt Engineering, Python, Pandas, Generative AI, AI Quality Assurance, Data Validation.

KPIT Technologies Ltd. Apr 2025 – Apr 2026

Associate Software Engineer *Pune, India*

- Developed and validated **ECU software** and control systems for the EVCU2 program, covering **ECU virtualization (vECU)**, system integration, and timing analysis across the delivery cycle.
- Drove virtual ECU workflows with **dSPACE SystemDesk, ControlDesk, and VEOS Player**, enabling validation to run without hardware-in-the-loop dependencies.
- Built **Python automation** scripts for environment setup, run orchestration, and log/result collection, improving repeatability and cutting turnaround time on repetitive tool-driven tasks.
- **Tech:** ECU Virtualization (vECU), AUTOSAR, C, Python, Control Systems, Timing Diagrams, Embedded Systems.

KuppiSmart Solutions (Livestockify) Sep 2023 – Mar 2024

Web Development Head, Internship *Vellore, India*

- Led **full-stack** integration of real-time **IoT** sensor data into a live monitoring dashboard, delivering actionable insights that improved livestock health monitoring by **25%**.
- Designed and maintained the **database schema** for streaming sensor data, ensuring efficient storage, indexing, and low-latency access for live queries.
- Built interactive **data visualizations** for temperature, humidity, air quality, and ventilation monitoring. **Tech:** Python, ThingSpeak API, SQL, Power BI.

InsideFPV May 2023 – Oct 2023

Web Developer, Internship *Vellore, India*

- Optimized data handling and rendering on the production storefront, reducing page load times by **40%** and lifting user engagement by **25%**.
- Developed server-side logic and **REST APIs** with **Node.js** and **Express**, tuning **MongoDB** queries for efficient data processing.
- Integrated secure **payment gateways** and enforced industry security standards. **Tech:** Node.js, Express, MongoDB, AWS EC2, AWS S3, WordPress.

Projects

Chills – AI Job Application Platform | *React, Node.js, MongoDB, LLM, Tesseract OCR* 2026

- Built and deployed a live AI platform that parses **unstructured resume PDFs** into schema-valid **JSON**, with a **Tesseract OCR** fallback for scanned documents and retry logic that recovers malformed model output.
- Engineered a **multi-provider LLM router** with automatic failover across **Gemini 2.5 Flash, Llama 3.3 70B, gpt-oss-120b, and Claude**, keeping generation available through provider rate limits and outages.
- Implemented **prompt engineering** that classifies each target company by sector and branches the prompt to match, shipped over **REST APIs** and **MongoDB** with JWT authentication and a Chrome extension.

Tokins – Live Queue & AI Admin Platform | *React, Node.js, MongoDB, Redis, AWS, Socket.io, LLM* 2025

- Built and deployed a **full-stack B2B SaaS** queue-management platform now serving **100+ businesses**, with a custom **LLM pipeline** driving AI marketing across Email, Telegram, Instagram, and voice alerts.
- Engineered real-time queue tracking over **WebSockets (Socket.io)**, cutting average customer wait time by **40%**, and load-tested against a **1M+ user database** with zero crashes using **Redis** caching and indexed **MongoDB** queries.
- Shipped a **React** admin panel with client-side routing and Recharts analytics, plus **Razorpay** payments and secure **Firebase authentication**.

Smart Surveillance – Edge AI on ESP32-CAM | *Python, OpenCV, Deep Learning* 2024

- Engineered **OpenCV** deep-learning **object detection** models tuned to run within the ESP32-CAM's constrained embedded hardware budget, analyzing live video streams at up to **30 FPS**.
- Published the resulting system at **IEEE AIIoT 2024**, contributing to IoT-based surveillance and smart-city research.

Lastone – Academic Management Android App | *Java, Firebase, AWS S3, Gradle* 2023

- Shipped an Android app with **10K+ Play Store downloads**, with an admin interface uploading exam papers to **AWS S3** and syncing links through **Firebase** for instant in-app access.
- Managed dependencies with **Gradle**, integrating RecyclerView, Firebase Authentication, Glide, and Lottie to extend app functionality and protect user data.

Education

Vellore Institute of Technology (VIT) Jul 2021 – Jul 2025

B.Tech. in Electronics and Communication Engineering – CGPA: 8.36/10 *Vellore, Tamil Nadu, India*

- **Relevant Coursework:** Data Structures & Algorithms, DBMS, Databases, Operating Systems, Object-Oriented Programming.

Achievements & Publications

IEEE Publication: *Smart Surveillance Object Detection System* – 2024 3rd International Conference on Artificial Intelligence for Internet of Things (AIIoT).

10K+ Downloads: Lastone, an Android application for managing academics, published on the Google Play Store.

Soft Skills

Communication | Teamwork & Collaboration | Problem-Solving & Analytical Thinking | Continuous Learning | Attention to Detail