Rahul Saini

AI/ML Engineer

Irvine, CA · 205-643-1054 · sainirahul0802@gmail.com · linkedin.com/in/sainir02 · github.com/RahulSaini02

SUMMARY

AI/ML Engineer with 5 years of experience spanning data engineering, cloud architecture, and full-stack development. Specialise in RAG pipelines, multi-agent LLM workflows, and NL2SQL systems using LangChain, AWS Bedrock, and Python. Proven track record of shipping AI that holds up beyond the demo — from a geospatial vehicle onboarding platform processing 500 VINs weekly to document intelligence pipelines that replaced fully manual processes. Actively seeking AI Engineer or Senior ML Engineer roles.

TECHNICAL SKILLS

AI / LLM: LangChain, RAG Pipelines, Multi-Agent Workflows, NL2SQL, Prompt Engineering, AWS Bedrock, OpenAI API

ML & Data Science: LSTM, SVD, Recommender Systems, Classification, Feature Engineering, Predictive Analytics

Cloud & DevOps: AWS (Lambda, Step Functions, Glue, S3, EC2, Textract, SageMaker, SNS, Redshift), Docker, CI/CD, Git

Data & ETL: Python (Pandas, NumPy), PL/SQL, Informatica, Apache Spark, PostgreSQL, Oracle, AWS Redshift

BI & Visualisation: Tableau, TabPy, Excel, Power BI

Languages / Misc: Python, JavaScript (React, Node.js), SQL, C#, R, HTML, Shell Scripting

PROFESSIONAL EXPERIENCE

AI/ML Engineer · Tata Consultancy Services · Irvine, CA *Jul 2025 – Present*

- Architected an end-to-end autonomous vehicle onboarding platform processing 100–500 VINs weekly using AWS Step Functions and Lambda — cutting dealer assignment turnaround from days to hours.
- Built a geospatial dealer matching engine (Haversine + real driving-distance API) that optimally assigns grounded vehicles across 2,000+ dealers by capacity and proximity.
- Developing an NL2SQL AI agent powered by LangChain and AWS Bedrock that converts plain-English questions into database queries, reducing insight delivery from weeks to minutes and eliminating dedicated BI resources.
- Building a dynamic dealer risk monitoring framework that generates real-time risk scores across 2,000+ dealers, surfacing compliance signals for operations teams.
- Automated 4–8 ETL pipelines across Oracle, Redshift, and PostgreSQL RDS using AWS Glue and Python, supporting fleet operations and risk reporting workloads.

Business Intelligence Developer · Tata Consultancy Services · Hyderabad, India *Aug 2021 – Jun 2025*

- Designed serverless orchestration architecture with AWS Step Functions and Lambda, delivering $40K+ in annual cloud cost savings through optimised workflow execution.
- Automated ETL pipelines via AWS Glue, Informatica, and Python — eliminating 90% of manual data handling and accelerating reporting cycles for the FP&A team.
- Delivered a document intelligence PoC using AWS Textract and Lambda to extract transactional data from financial PDFs into structured databases, replacing a fully manual process.
- Built a dynamic RBAC system in Tableau using TabPy, enabling user-level access control without exposing sensitive data or underlying models.

- Engineered compliance-focused Tableau dashboards connecting Oracle and Redshift, giving executives real-time visibility into SLA adherence and regulatory KPIs.
- Mentored junior analysts on automation best practices and secure cloud development; partnered with U.S.-based stakeholders to define KPIs and scope analytics requirements.

Full Stack Developer Intern · TA Digital · Hyderabad, India *Mar 2021 – Aug 2021*

- Built reusable React.js components with REST API integration and Spring Boot microservices handling authentication, CRUD operations, and PostgreSQL persistence with role-based access control.
- Contributed to CI/CD pipeline setup using GitHub and Docker in an Agile team, reducing post-deployment bug rate through JUnit and Postman test suites.

PROJECTS

FinanceOS — AI-Powered Personal Finance Platform

- Shipped a full personal finance platform in 10 days with Next.js, TypeScript, Supabase, and the Anthropic Claude API — writing only ~10% of the code manually.
- Implemented a structured AI development system with persistent project memory files, specialised subagents (frontend, backend, QA), reusable slash-command skills, and a Claude-owned TODO/bug tracker — eliminating context re-explanation across sessions.
- Delivered 20+ pages, 15+ API routes, bank integrations, and cron automation by treating AI as a junior engineer: assigning roles and specs, then reviewing every output rather than accepting snippets.
- Developed a TypeScript AI prompt template suite defining a consistent financial advisor persona ("Maya"), structured output formats, behavioural guardrails, and {{variable}} placeholder syntax for runtime injection.

AI Research Newsletter Agent

- Built a fully automated weekly AI newsletter pipeline using n8n, GPT-4.1, and Tavily live search — eliminating manual tracking of model releases, agent frameworks, open-source drops, and research papers.
- Designed a separation-of-concerns architecture where the AI agent outputs only structured JSON (titles, summaries, briefings, tags, links), while a dedicated JavaScript Code node owns the entire HTML email template — ensuring consistent rendering across all email clients at lower token cost.
- Implemented a signal-quality heuristic filtering out opinion pieces and duplicates using a "Would a senior AI engineer care about this?" standard, covering 7 content categories: model releases, agent frameworks, dev tools, infrastructure, open source, business/funding, and research papers.
- Delivers a styled HTML newsletter to inbox every Monday with zero manual effort — each story includes a factual summary, a 30-second spoken briefing, and a specific engineering takeaway.

EDUCATION

Master of Science, Data Science *Aug 2023 – May 2025*
University of Alabama at Birmingham · GPA 3.8

Bachelor of Technology, Computer Science *Aug 2017 – Sep 2021*
CMR Engineering College (JNTUH) · GPA 3.5

CERTIFICATIONS

- AWS Certified AI Practitioner
- AWS Developer – Edureka (Oct 2022)
- AWS Solutions Architect – In Progress (Expected Jun 2025)
- Data Analytics Essential Certification – Cisco Networking Academy
- AI For Everyone – DeepLearning.AI (Coursera)
- Python Data Structures – University of Michigan (Coursera)