

Yamini Srija Koulury

+1 (812) 7785430 | yaminisrija16@gmail.com | LinkedIn | www.yaminisrijakoulury.com

SUMMARY

AI Data Engineer with experience building LLM applications, retrieval-augmented generation (RAG), and reliable cloud data pipelines using Python, SQL, AWS, and Docker. Implemented Llama fine-tuning with LoRA, vector-based retrieval, and quantized model deployment for clinical AI research. Built Gemini-powered receipt extraction and automated workflows, plus semantic course search backed by vector embeddings and persistent agent memory. Optimized containerized ETL workflows to accelerate backend retrieval by 25% and automated data processing to reduce manual effort by 40%. Combines API integration, data validation, and model development to turn structured and unstructured data into practical AI applications and analytics.

EDUCATION

M.S. in Data Science | Indiana University, **GPA:** 3.57/4 | Bloomington, IN

Aug 2024 – May 2026

B.Tech. in Computer Science | Manipal Academy of Higher Education, **GPA:** 8.4/10 | Dubai, UAE

Aug 2019 – Jun 2023

TECH STACK

- **Languages:** Python, SQL (relational databases, query optimization), R, JavaScript, HTML, CSS.
- **Machine Learning & AI:** Llama, Gemini, Ollama, Mistral, RAG, LoRA, supervised fine-tuning, 4-bit quantization, vector embeddings, semantic search, agentic workflows, Scikit-learn, SHAP, NLP, predictive modeling.
- **Data Engineering:** ETL/ELT, REST APIs, data integration, validation, data modeling, pipeline monitoring, Pandas, NumPy.
- **Cloud, Databases & DevOps:** AWS (EC2, S3, App Runner, ECR), Google Cloud Pub/Sub, Firestore, CockroachDB (relational/vector), Docker, Linux, Git, Agile/Scrum, Fusion 360.
- **Applications & Analytics:** Flask, Node.js, Next.js, Power BI, DAX, Power Query, Tableau, Excel (Advanced), Plotly, Matplotlib, Seaborn, PyQt5.

EXPERIENCE

Research Analyst | IU Health | Indianapolis | Remote

Sept 2026 – Present

MV-Ed Project | Mechanical Ventilation Education

- Supported early evaluation of the MV-Ed simulator by reviewing resident ratings and comments across **32 practice cases**.
- Identified **two issues for follow-up**, a hard-to-see indicator and inconsistent case values, by examining reviewer comments.
- Documented initial findings and open questions in review notes by summarizing feedback on case clarity and simulator usability.

Data Engineer & Analyst | Heartland Community Network | Bloomington, Indiana | Remote

Jun 2026 – Present

- **Eliminated weekly manual reconciliation cycles** across **3** distinct reporting platforms by building a canonical metric dictionary to align marketing, operations, and executive teams.
- **Improved downstream reporting accuracy by 30%** by standardizing data definitions and automating SQL validation checks to surface previously undetected anomalies.
- Transitioned executive reporting from reactive to real-time capabilities by designing and deploying Power BI dashboards utilizing custom **DAX measures** and **Power Query transformations**.
- Translated complex operational requirements into a scalable analytics infrastructure by defining foundational KPI frameworks and structuring performance metrics in direct collaboration with key stakeholders.
- Partnered with marketing, operations, and executive stakeholders to gather data requirements and translate them into **reusable SQL transformations** and clearly documented reporting workflows.

Research Assistant | Indiana University Bloomington | Bloomington, Indiana | On-site

Jan 2026 – May 2026

- **Engineered an obstetric triage classification model**, achieving a 94.0% Clinical Safety Accuracy and 93.5% F1 Score, by implementing a two-stage supervised fine-tuning pipeline on a Llama 3 8B model.
- Minimized hallucinated reasoning and anchored predictions to validated guidelines by developing a gated Retrieval-Augmented Generation (RAG) architecture, dynamically retrieving clinical context via vector search when similarity exceeded a **0.65** threshold.
- Streamlined model training by curating a structured dataset of 788 patient scenarios, utilizing knowledge distillation to translate deterministic decision rules into natural-language training data.
- **Eliminated cloud API latency** and guaranteed local patient data privacy by deploying a 4-bit quantized Small Language Model (SLM) on high-performance computing clusters using Low-Rank Adaptation (LoRA).

Data Engineer Intern | Ikkabit Inc. | Folsom, CA | Remote

Jun 2025 – Aug 2025

- **Ensured 100% data integrity and accelerated backend retrieval by 25%**, by engineering a Python UDP pipeline, optimizing SQL queries, and containerizing ETL workflows with Docker and AWS.
- Built **ETL workflows** that extract, transform, and load real-time session data into SQL databases with built-in **validation checks** to preserve data integrity and support reliable reporting.
- **Standardized data visualization layers and enabled real-time monitoring** by transforming 50+ HTML/CSS/JS templates into dynamic PyQt5 layouts to create a unified Data Management Console.
- **Reduced the development cycle by 20% and minimized sync errors by 90%** by leveraging GenAI tools for logic generation and **Git** to version-control integrated Fusion 360 hardware specs.

Data Analytics Intern | Emirates Steel Arkan | Abu Dhabi, UAE | On-site

Mar 2024 – Jul 2024

- Achieved a 20% increase in data precision to facilitate a seamless ERP upgrade by orchestrating the cleaning and optimization of complex Materials Management datasets using Microsoft Excel (Advanced) and SQL.
- Reduced total processing time by 25% and enhanced data availability for stakeholders by streamlining procurement workflows and integrating **Cloud platforms with SAP APIs**.
- Eliminated reporting discrepancies and accelerated feature delivery by collaborating in an **Agile/Scrum** environment to develop and validate internal analytical applications against legacy systems.

Data Science Research Intern | New York University | Abu Dhabi, UAE

Sept 2023 – Dec 2023

- Collaborated in an Agile Environment to develop and test in-house analytical applications, leveraging Data Analytics to ensure robust system interoperability and reporting accuracy.
- Cut manual data processing effort by **40%** and improved actionable insights by building automated scraping and **visualization pipelines** using R and Python.

PROJECTS

ReturnMinder (Return Deadline Tracker) | *Next.js, Node.js, Gemini, Firestore, Google Cloud Pub/Sub*

- Automated purchase and return tracking from Gmail receipts by connecting **Google Cloud Pub/Sub** with a Node.js backend and using **Gemini** to extract product details, return policies, and deadlines.
- Simplified receipt entry and review by using **Gemini** to extract details from uploaded receipt photos, displaying editable results in a **Next.js dashboard**, and storing purchase records in Firestore.
- Enabled timely return reminders at **7 days and 1 day** before deadlines by scheduling background checks on Firestore records and sending automated emails with direct links to the purchase dashboard.

CourSelect (Course Advising Platform) | *Python, AWS, CockroachDB, Agentic workflows*

- Engineered a **full-stack AI academic advising platform**, achieving sub-3-second semantic search responses across an 8,983-course catalog, by implementing a two-level caching system and vector embeddings on **CockroachDB**.
- Guaranteed zero-downtime AI query resolution and 100% response reliability during external API outages by designing a 4-tier **LLM fallback architecture**, dynamically routing requests from **Google Gemini to local Ollama** models and Mistral Cloud.
- Centralized complex application state and strictly enforced major-specific enrollment logic across 242 departments by architecting a 5-tier **agentic memory system** utilizing Cockroach DB Cloud Serverless to manage relational, vector, and contextual data.
- Automated structured auditing and durable recommendation storage, capturing 100% of student-advisor interaction sessions, by integrating **AWS S3** for continuous JSON report generation and deploying the Python Flask backend to an **AWS EC2** instance.
- Integrated multiple external **REST APIs** (Gemini, Mistral Cloud) and a relational/vector database behind a Python Flask backend, with session logging to monitor reliability.

Analyzing and Visualizing Human Reference Atlas (HRA) UI Usage | *Python, Pandas, Plotly, AWS*

- Enabled data-driven documentation improvements by processing **120K+ log records** (665MB) by architecting a scalable analytics pipeline using **AWS (EC2)** to parse CloudFront logs with Pandas and NumPy.
- Built an **ETL-style workflow** to extract raw CloudFront logs, clean and transform them with Pandas, and load structured outputs that feed dashboard reporting.
- Uncovered dominant user workflows and underutilized components by designing an interactive dashboard that visualizes session data using **Matplotlib** and **Seaborn** to map high-demand Knowledge Graph resources.

The Untold Stories of IPL: Advanced ML & Visualization | *Python, Power BI, Tableau*

- Identified key performance drivers beyond title counts by modeling the correlation between competitive win percentages and brand value using Predictive Modeling and Scikit-Learn ensemble models (SVM, Random Forest).
- Mapped player consistency and transformed aggregate statistics into compelling narratives by utilizing **SHAP** for Explainable AI, **NLP** for textual analysis, and creating advanced visualizations (Rose Charts) in **Power BI** and **Tableau**.

CERTIFICATIONS

Microsoft Certified: Power BI Data Analyst Associate