Sriram Chunduri

(619) 436-6443 | csriramsachit@gmail.com | LinkedIn | GitHub

SUMMARY

AI/ML Engineer with 2+ years of experience building production ML systems and LLM applications. MS Data Science (Drexel, GPA 3.8) with hands-on expertise in RAG architectures, transformer fine-tuning, and real-time ML pipelines. Built enterprise RAG systems achieving 94% retrieval accuracy and fraud detection engines processing 10K+ transactions/minute. Strong foundation in PyTorch, LangChain, and MLOps.

EXPERIENCE

RK Financial Management Consultants | *Data Scientist* **Jan 2021 - Dec 2023**

- Conducted machine learning and data analysis using Python, R, and SQL, enhancing predictive accuracy by 30%.
- Generated over 20 reports and visualizations with Plotly and Tableau, improving strategic decision-making and client outcomes by 25%.
- Collaborated with cross-functional teams to produce actionable insights leveraging data science techniques using SQL, Spark, and Tableau, enabling faster decision-making for stakeholders

Vanvi Solutions | *AI Engineer* **Aug 2025 - Present**

- Built invoice processing automation for e-commerce client using Python and GPT-4 API, extracting structured data from 500+ monthly invoices and reducing manual entry time by 60% through rigorous debugging
- Developed customer support chatbot using LangChain and RAG architecture with rapid prototyping, indexing 200+ FAQ documents to handle 70% of routine inquiries without human intervention
- Created sales analytics dashboard integrating Shopify and Google Sheets data via REST APIs, providing real-time KPI tracking for small business owners
- Implemented lead scoring model using XGBoost for marketing agency client, improving qualified lead identification by 35% through behavioral data analysis

PROJECTS

RAG Financial Document Analysis **Jun 2025 - Sep 2025**

- Built production-grade retrieval-augmented generation (RAG) system indexing 500+ SEC filings (10-K, 10-Q) from Apple, NVIDIA, Microsoft spanning 3 years, chunking into 12K+ segments with dense and BM25 keyword embeddings for hybrid retrieval via Qdrant vector database.
- Implemented baseline retrieval pipeline (dense + BM25 hybrid search with cross-encoder reranking) achieving 85%+ retrieval accuracy and F1 0.80+ on 100-question financial benchmark; developed CLaRa variant reducing context token usage by 35% and latency from 1000ms → 500ms p50 while maintaining answer quality.
- Deployed FastAPI microservice with Streamlit demo comparing baseline vs. CLaRa modes side-by-side, with comprehensive evaluation suite (precision, recall, F1, hallucination rate); documented methodology and results in Jupyter notebooks.

MCP-Powered Business Intelligence Agent **Apr 2025 - Jun 2026**

- Developed MCP-powered Business Intelligence agent using Claude API and Anthropic MCP SDK, connecting to multiple SMB data sources with natural language interface for ad-hoc queries.
- Engineered tool-calling orchestration layer that parses user intent, selects appropriate data sources, generates SQL, and formats business-friendly answers with confidence scores; achieved 85%+ query accuracy on 30-question benchmark dataset.
- Deployed FastAPI service with scheduled insight generation, performance monitoring, and tested against realistic SMB scenarios, demonstrating production readiness for SMB analytics automation.

Algorithmic Trading Bot **Jun 2024 - Present**

- Designed and developed a Python-based financial trading bot, boosting back-tested returns from 15% to 47% with XGBoost and advanced indicators.
- Manipulated financial data using Pandas and NumPy, achieving 96% accuracy in price forecasts with machine learning models.
- Took full ownership of development and deployment, integrating APIs and WebSockets for real-time buy/sell predictions and demonstrating strong quantitative analysis skills.

EDUCATION

Drexel University | *Master of Science, Data Science* (GPA: 3.8) **Jan 2024 - Jun 2025**

Vardhaman College of Engineering | *Bachelor of Technology, Electronics and Communication* **Mar 2019 - Jan 2023**

SKILLS

- **Languages:** Python, SQL, R, PySpark, Bash, C++, Java
- **ML Frameworks:** PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, LangChain, LlamaIndex
- **Data Processing:** Pandas, NumPy, PySpark, Spark, Dask, Kafka, Polars
- **Deployment & DevOps:** Docker, Kubernetes, Flask, FastAPI, Git, MLflow, Airflow, CI/CD
- **Cloud Platforms:** GCP, AWS, Azure ML
- **Databases:** SQL, NoSQL
- **ML Techniques:** Regression, Classification, Clustering, Recommendation Systems, Deep Learning, NLP, Time Series Analysis, A/B Testing, Feature Engineering, Statistical Modeling, Machine Learning, Data Science
- **Engineering & Data Engineering:** Apache Spark, Apache Kafka, PySpark, Redis, PostgreSQL, MongoDB, ETL Pipelines, Hadoop, Finite Element Analysis, Rapid Prototyping, Debugging

CERTIFICATIONS

- Python Certification: University of Michigan
- Introduction to Model Context Protocol: Anthropic Academy

ACHIEVEMENTS & INTERESTS

- Published IEEE Conference Paper on QCA Nanotechnology-Based Low-Power Circuits
- 2nd Runner-up, Hyderabad Speed Fest, K6 category