Durga Katreddi

+1 917-400-7205 | katreddisrisaidurga@gmail.com | linkedin.com/in/sri-sai-durga-katreddi- | github.com/KATREDDIDURGA

SUMMARY

- AI/ML Engineer with 5+ years of experience in solution architecture, GenAI development, and large-scale financial AI systems across banking and fintech. Proven expertise in building compliance-driven solutions for KYC, Suspicious Activity Reports (SAR), credit risk, and fraud detection, delivering $6M+ in measurable business impact. Skilled in Python, SQL, Java, and cloud platforms (AWS, GCP, Azure) with hands-on experience in microservices architecture, APIs, and data integration. Advanced proficiency in LLMs, multi-agent systems, RAG, PEFT/LoRA fine-tuning, and frameworks such as PyTorch, TensorFlow, LangChain, and Hugging Face. Adept at statistical modeling, dynamic model tuning, and AI/ML pipelines, with strong MLOps, observability, and governance practices. Recognized as a trusted advisor, providing technical guidance, evaluating tools/frameworks, and leading cross-functional teams to design and implement scalable, efficient GenAI solutions.

SKILLS

- **Programming & Scripting:** Python, SQL, Java, C++, Go, JavaScript, TypeScript, HTML, Node.js
- **AI/ML Frameworks & Libraries:** PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, LightGBM, LangChain, LangGraph, Hugging Face Transformers, OpenAI APIs, Vertex AI
- **GenAI & Advanced AI:** Retrieval-Augmented Generation (RAG), Multi-Agent Systems, LLMOps, Prompt Engineering, PEFT/LoRA Fine-Tuning, Stable Diffusion, Whisper, DALL·E, CLIP
- **Solution Architecture & Analytics:** GenAI Solution Design, Microservices Architecture, API Development (REST, GraphQL), Data Integration, Statistical Modeling, Dynamic Model Tuning
- **MLOps & Observability:** CI/CD Pipelines, Model Deployment, Drift Detection, A/B Testing, Performance Monitoring, Weights & Biases, Guardrails, Risk Mitigation
- **Cloud Platforms & DevOps:** AWS (EC2, S3, Lambda, EMR), GCP (Vertex AI, BigQuery), Azure, Docker, Kubernetes, Terraform, Infrastructure as Code
- **Databases & Data Processing:** Snowflake, MongoDB, Hadoop, Spark, Hive, SQL/NoSQL Databases, ChromaDB, FAISS, Pinecone, Data Lakes, ETL Pipelines
- **Visualization & Reporting:** Tableau, Power BI, Streamlit, Matplotlib, Seaborn, Interactive Dashboards
- **Specialized Domain Expertise:** Credit Risk Modeling, Fraud Detection, Suspicious Activity Reports (SAR), KYC Workflow Automation, Compliance Analytics, Document Processing
- **Collaboration & Leadership:** Stakeholder Communication, Technical Guidance, Cross-Functional Team Leadership, Trusted Advisor Role, Executive Presentations

EXPERIENCE

Bank of America **January 2024 – Present**
AI Software Developer

- Architected GenAI-powered multi-agent systems (LangChain, LangGraph) to automate KYC workflows and Suspicious Activity Reports (SAR), reducing review time by 70% while maintaining compliance.
- Designed and deployed retrieval-augmented generation (RAG) frameworks with FAISS and ChromaDB, serving 10,000+ regulatory queries daily at 99.9% uptime.
- Evaluated and recommended frameworks/tools (LangChain, Hugging Face, Weights & Biases), ensuring alignment with compliance and business priorities.
- Built solution architectures for AI observability, capturing decision traceability, confidence scores, and reasoning chains to support auditability.
- Implemented biometric verification systems with JWT/Bcrypt, reducing onboarding time by 65% while meeting global data protection requirements.
- Deployed LLMs with PEFT/LoRA fine-tuning for document understanding, achieving 92% accuracy over baseline.
- Led cross-functional teams through POC to production phases, providing technical guidance and acting as a trusted advisor on compliance AI use cases.

- Developed cloud-native microservices (FastAPI, Flask, Docker, Kubernetes, AWS/GCP Vertex AI) enabling 99.95% availability and auto-scaling of AI workloads.
- Created full-stack AI-powered dashboards (React, Tableau, Power BI) to monitor KYC and fraud risk metrics in real-time, improving executive visibility.
- Built LLMOps pipelines with risk mitigation strategies, detecting model drift and hallucinations in real-time, cutting manual interventions by 70%.

University of North Texas **August 2023 – December 2023**
Student Assistant

- Developed a demonstration project leveraging GPT-3.5 and CLIP to create an interactive learning tool that helped students visualize relationships between text prompts and generated images.
- Built and maintained a benchmark testing environment for comparing LLaMA 2, Mistral 7B, and Falcon model performances using PyTorch and Hugging Face Transformers, which became a core resource for the department's AI curriculum.
- Created and presented a technical seminar on SAS programming for analytics, demonstrating practical applications for financial data analysis to students.
- Developed Snowflake data pipelines and SQL optimization techniques for handling large datasets, improving query performance by 40% for student research projects.
- Assisted students in troubleshooting and optimizing their implementations of Python and SQL code for data analysis projects, improving project completion rate from 70% to 95%.
- Collaborated with professor to develop a comparative analysis of credit scoring methodologies using XGBoost and LightGBM, which was incorporated into the department's financial analytics curriculum.
- Implemented advanced data visualization techniques using Matplotlib and Seaborn to create compelling presentations for academic conferences and student projects.
- Mentored junior developers and students in modern development practices, including Git version control, CI/CD pipelines, and containerization with Docker.
- Developed automated grading systems using Python and machine learning algorithms, reducing manual grading time by 60% while maintaining accuracy.
- Created comprehensive documentation and tutorials for AI/ML model development, serving as reference materials for over 200 students in the program.
- Implemented A/B testing frameworks for evaluating different teaching methodologies and learning outcomes in data science courses.
- Built RESTful APIs using FastAPI and Flask for student project submissions and automated evaluation systems.
- Integrated cloud services (AWS, GCP) into the curriculum, providing hands-on experience with modern deployment practices for machine learning models.
- Developed cross-platform applications using JavaScript and TypeScript for educational tools, enhancing the learning experience for data science students.
- Conducted workshops on prompt engineering and LLM fine-tuning techniques, preparing students for careers in AI and machine learning.

Cognizant(American Express) **December 2020 – December 2022**
Software Engineer with Machine Learning

- Designed and implemented solution architectures for credit risk models, enabling real-time scoring for 1.5M+ cardholders using AWS, Hadoop, and Snowflake.
- Built and deployed default risk models (XGBoost, LightGBM, Random Forest), delivering $6M+ in pre-tax income gains and a 150-basis-point Gini lift.
- Applied advanced statistical methods and dynamic model tuning, improving prediction accuracy by 5% across ensemble models.
- Developed fraud detection pipelines with Apache Spark for real-time transaction monitoring, reducing false positives by 25%.
- Created customer segmentation models (K-Means, Hierarchical Clustering) that improved targeted marketing and increased retention by 15%.
- Leveraged NLP and sentiment analysis on customer feedback, driving a 15% uplift in satisfaction scores.
- Provided technical guidance to cross-functional risk teams, aligning AI modeling frameworks with compliance and business objectives.
- Implemented MLOps best practices with Docker and Kubernetes, ensuring scalable, secure deployments across production environments.

- Built interactive dashboards (Tableau, Power BI) for risk and churn insights, empowering executive decision-making.
- Developed feature engineering pipelines (200+ variables) and integrated SQL/NoSQL sources into unified data warehouses for advanced analytics.

UJR Corporate Solutions Pvt. Ltd. **August 2019 – November 2020**
Data/Decision Scientist

- Engineered a 100-day historical data processing pipeline in MongoDB, implementing profile-based imputation that averaged three years of historical patterns to ensure 30% improvement in data completeness.
- Developed sophisticated anomaly detection methods using statistical analysis and machine learning to identify outliers across large-scale financial datasets, reducing forecasting errors by 25%.
- Created specialized machine learning models including ARIMA, LSTM, CNN, and XGBoost using TensorFlow and PyTorch to capture different market behaviors across 3 market segments.
- Wrote complex SQL queries and stored procedures for data extraction and transformation from multiple sources, creating a unified analytical framework for financial modeling.
- Implemented an ensemble model using advanced stacking techniques to aggregate predictions from 24 models, increasing forecast accuracy by 15% for financial and market predictions.
- Optimized MongoDB query performance through compound indexing and aggregation pipelines, reducing data retrieval times by 60% and enabling real-time dashboard visualizations.
- Designed a RESTful Flask API with stateless architecture for seamless dashboard integration, reducing manual intervention by 40% and supporting horizontal scaling.
- Implemented comprehensive error handling and logging systems using Python that increased pipeline reliability from 92% to 99.7% completion rate for daily processing tasks.
- Developed automated data quality assessment tools using statistical methods and machine learning algorithms, ensuring data integrity across all analytical processes.
- Built interactive visualization dashboards using Tableau and Matplotlib for stakeholder reporting, enabling data-driven decision making across multiple business units.
- Implemented version control and CI/CD practices using Git and automated testing frameworks, ensuring code quality and deployment reliability.
- Created comprehensive documentation and knowledge transfer materials for data science workflows, facilitating team collaboration and onboarding of new team members.
- Utilized cloud computing resources (AWS EC2, S3) for scalable data processing and model training, optimizing costs while maintaining performance standards.
- Developed time series forecasting models using advanced statistical techniques and deep learning approaches, achieving superior prediction accuracy for financial markets.
- Implemented A/B testing frameworks for model evaluation and business strategy optimization, providing statistical significance testing for decision-making processes.

PROJECTS

FinGuard Agents : Multi-Agent GenAI System for Real-Time Fraud Prediction & Explanation

- Architected an agentic AI system using open-source LLMs, multi-agent workflows, and custom rule engines to predict and explain fraud in financial transactions in real-time.
- Integrated retrieval-augmented generation (RAG) and narrative generation agents to produce compliant, human-readable fraud justifications, improving auditability and user trust.
- Designed agent coordination logic with tool usage control, intent routing, and fallback behavior logging, supporting extensibility across retail, fintech, and regulatory domains.
- Built an event-driven pipeline with modular hooks for LLMOps, graph-based tracing, and domain-specific knowledge bases using Python, LangGraph, and SQLite.

AgentScope : AI Agent Traceability & Decision Debugging Framework

- Designed a general-purpose decision traceability system for AI agents using FastAPI, SQLAlchemy, SQLite, and Streamlit to visualize multi-step decision paths.
- Implemented a backend API to fetch agent run metadata and step-wise execution (thoughts, tools, actions, final response), improving model interpretability, debuggability, and fallback analysis.
- Created a professional UI for agent trace search using Streamlit, enhancing trace audits for LLM, rule-based, and intent-based AI agents across fraud, legal, and retail domains.
- Engineered the system with extensible data models and modular hooks for LangChain, Rasa, and multi-agent orchestration, supporting future LLMOps, compliance, and real-time agent monitoring needs.

VisionAlign: Feedback-Driven Image Generation via Low-Resolution Preview

- Developed a two-stage GenAI image pipeline that first generates low-resolution previews on CPU to align with user intent before committing to full-resolution GPU rendering.
- Minimized compute costs by up to 80% using conditional generation feedback loops and progressive refinement, avoiding redundant high-resolution model calls.
- Implemented a natural language feedback module to capture semantic mismatches (e.g., object pose, context) between user expectations and generated previews.
- Tech stack included Stable Diffusion, CPU-inference U-Net, custom feedback handler, and prompt alteration agent for interactive fine-tuning.

AI Insurance Claims Processing System: NLP & Fraud Detection

- Built an end-to-end AI-powered insurance claims processing system using GPT-4 and advanced NLP techniques, achieving 95% reduction in processing time and 92% accuracy in fraud detection, resulting in $500K annual savings per 1000 claims processed.
- Developed natural language claim processing, automated fraud analysis with confidence scoring, and professional report generation capabilities, creating a modular architecture with real-time analytics dashboards for executive decision-making.