Sai Kalyan Veera

Dallas, TX | +1 (716) 465-8781 | kalyanveera66@gmail.com

Summary

Generative AI Engineer specializing in large language models (LLMs) and scalable machine learning systems on cloud platforms like AWS and GCP. Currently, I have successfully designed and implemented a robust Retrieval-Augmented Generation (RAG) platform, achieving 99.99% uptime and substantially reducing inference latency. Proven ability to develop and deploy high-performance microservices and analytics solutions, enhancing pipeline throughput and user engagement. Committed to leveraging expertise in generative AI and cloud engineering to drive innovative solutions and performance improvements in future projects.

Education

University at Buffalo **Jan 2024 - Jun 2025**
Master of Science, Data Science

KI University **Jul 2019 - May 2023**
Bachelor of Technology, Electronics and Communication Engineering

Work Experience

Alorasoft Inc **Aug 2025 - Present**
Generative AI Engineer *Dallas, TX*

- Designed a production RAG platform using Python, LangChain agents with MCP tools, ChromaDB/Qdrant retrieval, AWS S3/EKS/Elastic Beanstalk, and CloudWatch, launched with 99.99% uptime and 60% lower inference latency
- Developed LLM microservices in Java Spring Boot (REST), orchestrating retrieval, tool-use, inference, deployed on EC2+ALB, lifting pipeline throughput 45% and cutting p95 response time 30%
- Tuned Amazon RDS/Aurora (MySQL/PostgreSQL/Oracle) for RAG workloads, 6 schema/indexing for geospatial & time-series, achieving sub-second queries, 30% storage cost reduction, and reliability at 50k+ daily requests
- Shipped real-time analytics UIs (React, Node/WebSockets, AWS IoT Core/API Gateway) and hardened delivery with Docker, Jenkins, GitHub/GitLab, CodePipeline/CloudFormation, achieving 9%+ test coverage, 3x developer productivity, 50% faster refresh, and 35% higher engagement

Foundever **Jul 2023 - Jan 2024**
Technical Analyst *Hyderabad, India*

- Streamlined and executed Python (Pandas) ETL pipelines processing over 1 million records per day, reducing batch time by 60% and improving reliability with validation and logging
- Led AWS migration (S3, EC2, RDS) of legacy workloads, reducing maintenance overhead 50% and enabling scalable, automated backups
- Architected an ML-based recommendation system using behavioral signals, which increased enrollments by over 30% and enhanced user engagement
- Executed 7+ Tableau dashboards (churn, sales forecasting, performance KPIs), standardizing metrics and accelerating decision-making across teams

Academic Projects

Customer Support Chatbot

- Launched a customer service chatbot powered by LLaMA-3.1-8B, used LoRA fine-tuning and 4/8/16-bit quantization to cut response time by 40
- implemented a scalable ML pipeline with Flask APIs and Ollama, containerized in Docker and deployed on AWS ECS Fargate, supporting up to 500 concurrent queries per day
- Automated CI/CD with AWS CodePipeline, reducing deployment errors by 35% and increasing iteration speed by 30%
- Instrumented performance with AWS CloudWatch, trimming inference latency by 22% and improving reliability via real-time alerts and logs

RAG Chatbot

- Architected a Retrieval-Augmented Generation system over 60,000+ Wikipedia docs, achieved 89% document-match accuracy with an evaluation harness.
- Implemented hybrid retrieval: Sentence Transformers embeddings + TF-IDF with cosine re-ranking, delivering 3× faster retrieval and higher precision.
- Orchestrated GPT API generation with context windows and prompt templates, reached 92% response relevance in Q&A tests

New York Airbnb Market Analysis

- Executed end-to-end EDA on NYC Airbnb (20k+ rows, 12+ features) in Python (Pandas/NumPy/Matplotlib/Seaborn, Jupyter/VS Code), standardized ID dtypes, removed nulls/duplicates, and capped price outliers (<$1.5k) to deliver an analysis-ready dataset
- Engineered price_per_bed and 10+ visuals, released neighborhood-group × room-type pricing benchmarks, geo clusters (lat/long), and correlation insights (price higher with beds, price lower with minimum nights) to guide pricing and supply decisions.

Technical Skills

- **Programming Languages**: Python, C, SQL, HTML, CSS
- **Frameworks & Libraries**: Numpy, Pandas, Seaborn, Scikit-Learn, PyTorch, TensorFlow, Keras, OpenCV, Hadoop, Flask, OpenCV, FastAPI
- **Databases & Cloud**: PostgreSQL, MySQL, Snowflake, AWS, Docker, GitHub, Tableau
- **AI/ML & GenAI**: Machine Learning, Deep Learning, Computer Vision, NLP, LLMs, LangChain, Quantization, Fine-tuning, Distributed Computing, Inference, RAG, MCP, CI/CD, Hugging Face, LangChain, LlamaIndex, CrewAI
- **Other:**: REST APIs, Agile/Scrum (Basics), MLOps