Sriabhinay Kusuma

+1 (256) 468-2929 | sriabhinay1@gmail.com | linkedin/sriabhinay556 | **Dallas, Texas***

SUMMARY

- **AI Engineer** with expertise in developing and deploying production-grade **AI/ML, full-stack** software solutions across Healthcare, Retail and HRTech.
- Specialized in **Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) and Agentic systems** for regulated and trust-critical enterprise systems.
- Proven record of delivering measurable business outcomes with saving **1000+** manual hours with automation and saving **~80–90%** in cost while **maintaining strict safety, correctness, and reliability** guarantees.
- Experienced in designing **low-hallucination systems** through schema-grounded agents, retrieval evaluation, and deterministic orchestration layers.
- Experience in **Python**-based ML engineering with **PyTorch, Keras**, transformer-based models using **HuggingFace**, agentic systems via **LangChain/LangGraph**, PEFT-based fine-tuning (**LoRA/QLoRA**), data engineering and production-grade **MLOps/LLMOps** on **AWS SageMaker** with **OpenTelemetry, Prometheus and Grafana**-driven observability.
- Strong ML and software engineering foundation, supported by a **Master's degree in Computer Science (4.0 GPA)** and **AWS Certified Developer – Associate**

WORK EXPERIENCE

Garuna

AI Engineer *Aug 2024 – Present*

- Designed and deployed a **LLM**-powered HIPAA-compliant, PHI-safe healthcare **retrieval and automation platform**, consolidating clinical policies, compliance documents, **reducing compliance audit preparation time by ~80%** while ensuring zero PHI leakage across inference, logging, and observability layers.
- Built and optimized **end-to-end RAG pipelines** using **LangChain**, OpenSearch, and FAISS, enabling contextual Q&A across **2M+** healthcare documents, enabling **persistent-memory** with retrieval confidence scoring and ranking controls to **minimize hallucinations** and maintain compliance-aligned response accuracy.
- Implemented adaptive **chunking** and retrieval strategies across data formats (PDFs, tables, and policies), to improve semantic recall and **increase retrieval relevance** while **reducing hallucination**-driven re-queries by **~30%**
- Developed a **text-to-SQL** agent for compliance officers and executives to query structured data in natural language, significantly reducing SQL hallucinations and improving query reliability and trustworthiness by **~80%**
- Built secure AI chatbot services shared across multiple departments, using **AWS Cognito** and **API Gateway** to enforce tenant-level access control and prevent cross-department data leakage.
- Operationalized **AIOps/LLMOps** with **OpenTelemetry**-driven stage tracing, SLO validation using A/B tests, and EKS blue-green plus canary deployments, enabling safe rollouts while maintaining p95 latency, inference cost, and output quality within defined SLOs before enterprise release. Implemented monitoring with **Grafana**.
- Developed a **Generative AI** solution for high-volume warranty claims processing, leveraging **NLP model fine-tuning** on a proprietary, multi-format dataset (claims, receipts, images) to automate classification and response generation.
- Conducted feature engineering with **Python, Pandas, SQL and NumPy** from rule-based heuristics and behavioral patterns to improve model training for **50,000+** complex data.
- Optimized a state-of-the-art **NLP** model with Parameter-Efficient Fine-Tuning (**PEFT/QLoRA**) techniques, iteratively optimizing hyperparameters to achieve high accuracy in domain-specific tasks and **saving ~80% training costs**.
- Engineered a **Retrieval-Augmented Generation (RAG)** system to enhance the precision of information retrieval on domain-specific knowledge, improving **retrieval accuracy by 20%** for more accurate claim processing.
- Architected and deployed a scalable **MLOps** pipeline via **AWS AI services (SageMaker, Bedrock)**, integrating the model with a **FastAPI-based REST API** and **Docker** for secure, low-latency, real-time handling of thousands of weekly claims.
- Developed a robust full-stack data pipeline including an automated web scraping component that processed **over 10,000+** articles monthly, reducing manual data collection time by **1000+** hours/month.

Archeed

AI Software Engineer *Jan 2024 – Aug 2024*

- Developed and deployed end-to-end **Retrieval-Augmented Generation (RAG)** systems for intelligent recommendation & **semantic search** for over **5,000+** professionals significantly enhancing job-to-profile recommendation accuracy.
- Architected and optimized real-time **vector search** pipelines using **LangChain** and **pre-trained models** (e.g., **DistilBERT** embeddings), resulting in a **40%** improvement in content matching and retrieval speed across domain-specific data.
- Engineered and managed the scalable **MLOps** deployment lifecycle: containerizing model-backed microservices (**Docker**) and leveraging **Amazon SageMaker** for high-throughput, low-latency inference in the production environment.
- Implemented a multi-role architecture for intelligent job curation, integrating **continuous integration/continuous deployment (CI/CD)** workflows via **GitHub Actions** to ensure system reliability and rapid feature iteration.
- Developed the foundational full-stack components, including **RESTful APIs** and front-end integration **(Next.js/React),** while managing specialized data storage using highly available **Vector Databases** and MongoDB.

Saayam For All San Jose, CA (Remote)

AI Full Stack Developer *July 2025 – Present*

- Implemented an automated UI testing framework using **Playwright**, reducing overall **testing cycle time by ~50%** and improving defect detection across critical user flows.
- Leveraged Playwright with **Model Context Protocol (MCP**) to generate and bootstrap UI test scripts via agent-driven workflows, accelerating test authoring and **reducing manual QA effort**.
- Used **AI-assisted test generation** to create initial Page Object Models (POMs) and regression scenarios, refining them into deterministic Playwright tests for CI stability.
- Executed Playwright test suites in CI using **distributed test sharding**, parallelizing execution across runners and cutting end-to-end test time by **~75% (4× speedup)**.
- Implemented i18n internationalization, reducing payload size and improving page load times by **~25%,** enhancing global user experience.
- Developed **CI/CD pipelines** using **GitHub Actions** to automate build, test, and deployment stages, significantly reducing manual overhead and release friction.

National Institute of Standards and Technology Huntsville, AL

Software Developer *Jan 2025 – May 2025*

- Developed an automated pipeline to collect, pre-process, analyze CPU-generated jitter with NIST SP 800-90B standards.
- Automation with Python/Bash scripts, evaluation across Docker, VM, BareMetal to assess security vulnerabilities.
- Contributed the tool and research that helps NIST to assess the security trade-offs using CPU-jitter for their cryptography.

University of Alabama, Huntsville Huntsville, AL

Frontend Developer *Jul 2023 – Dec 2023*

- Collaborated on a microservice-based recruitment system to optimize hiring workflows for university departments. Improved review productivity by 50%.
- Worked on faculty collaboration tool improving user experience and code maintenance. Implemented RESTful APIs and microservices architecture using Django and Python, ensuring scalable system design.

Graduate Teaching Assistant *Aug 2022 – Apr 2023*

- CS454/554 Cloud Computing - Lab Setup and Support: serving 50+ students, creating a practical cloud environment that increased successful project completions by 20%. Worked with professors to build impactful assignments for students.
- CS104 Python - Student Assistance and Grading: Guided 40+ students through in-person labs, offering individualized support that elevated assignment completion rates and improved average scores by 10%.
- CS317 Design and Analysis of Algorithms - Responsible for grading and in-person academic support for 60+ students. Assisted the professor in preparing coursework, assignments, and exams which enhancing class efficiency.

EDUCATION

University of Alabama in Huntsville Huntsville, Alabama

Master of Science – Computer Science *(Machine Learning & Cybersecurity)* ***4.0 GPA*** *Graduated – May 2025*

SKILLS & INTERESTS

- **Languages**: Python, JavaScript, Java, SQL, Bash
- **AI & Machine Learning**: TensorFlow, Keras, PyTorch, Scikit-learn, Hugging Face Transformers, LangChain, LangGraph, Retrieval-Augmented Generation (RAG), Agentic Systems, Text-to-SQL, BERT, LLaMA, DistilBERT, GloVe, Generative AI, Model Evaluation, Hallucination Mitigation, Confidence Scoring, Re-ranking
- **NLP & Data Science**: Text Classification, Sentiment Analysis, Semantic Search, Clustering, Vector Embeddings, Retrieval Pipelines, Feature Engineering, Information Extraction, Document Intelligence, Schema-Grounded Querying
- **Computer Vision**: OpenCV, Image Preprocessing, Object Detection (Foundational)
- **MLOps & Cloud**: AWS (SageMaker, Bedrock, EC2, Lambda, S3, DynamoDB), Docker, Kubernetes (EKS), GitHub Actions (CI/CD), A/B Testing, Canary Deployments, Blue-Green Deployments, OpenTelemetry, Prometheus, Grafana, MLflow, Nginx
- **Web & Backend**: React.js, Next.js, Node.js, Express.js, Django, Spring Boot, FastAPI, REST APIs, Full-Stack
- **Databases**: PostgreSQL, MySQL, MongoDB, DynamoDB, OpenSearch, FAISS, ChromaDB, Vector Databases
- **Tools & Platforms**: Git, Linux, JIRA, Postman, Vercel, Figma, Agile/Scrum, Playwright, Model Context Protocol (MCP)
- **Certifications:** AWS Certified Developer – Associate - Link
- **Publications:** Machine Learning – Deep Learning based Cyberbullying Detection. ResearchGate - Link

PROJECT EXPERIENCE

AI-Scraper

- Engineered a prompt-driven AI scraping pipeline using LLM agents to automate data collection across anti-bot websites.
- Reduced token usage by 80% and increased scraping efficiency, enabling scalable use in data-driven apps like market research.

Real-Time Facial Emotion Classifier

- Developed a Transfer Learning VGG19-based Convolutional Neural Network (CNN) for facial emotion classification, achieving a real-time validation accuracy of 90%.
- Packaged the model into a Dockerized Flask API and integrated it with OpenCV to enable live, webcam-based inference for monitoring applications.

Contextual Sentiment Analysis Model

- Built and trained an advanced sentiment model using a Bi-LSTM architecture with Attention Mechanisms and GloVe embeddings for contextual richness.
- Achieved 88% validation accuracy by implementing regularization techniques to mitigate severe overfitting, resulting in top performance in the project class.