Sriabhinay (Abi) Kusuma

Dallas, Texas* · sriabhinay1@gmail.com · +1 (256) 468-2929

Objective

AI Engineer with expertise in developing and deploying advanced AI solutions from fine-tuning to deployment. I'm a workhorse. I get things done. I love intense environments where I can wear multiple hats. I don't mind doing grunt work or long hours; all I care about is making an impact.

Masters

University of Alabama in Huntsville — Alabama, US
Master of Science Computer Science - *GPA: 4.0* — Graduated May 2025

Experience

AI Engineer @ Garuna — Aug 2024 - Present

- Designed and deployed Generative AI solutions to automate warranty claims saving 1000+ of man hours.
- Fine-tuned transformer models with Parameter-Efficient Fine-Tuning (PEFT) lowering 80% memory costs.
- Developed Retrieval-Augmented Generation (RAG) using LangChain & FAISS improving response accuracy by 20%.
- Developed Quantized LLMs on SageMaker, achieving 35% lower latency & lower inference costs by optimized scaling.
- Model evaluation with SageMaker Model Monitor, CloudWatch, Prometheus to track model health and latency.
- Developed enterprise chatbots using LangGraph agents with MCP, SQL query handling and compliance driven.

AI - Full Stack Developer @ Saayam For All — San Jose, CA (Remote) July 2025 - Present

- Engineered a high-impact volunteer-matching platform using RAG and DistilBERT, leveraging vector embeddings to deliver highly personalized and accurate help recommendations with hyperparameter tuning on SageMaker.
- Implemented solutions that reduced testing phase time by 50% and improved defect detection accuracy.
- Implemented i18n internationalization, reducing payload, improving page load by 25% for multi-lingual users.
- Developed CI/CD GitHub Actions pipelines to automate build, test, & deployment stages, reducing manual overhead.

Software Developer @ National Institute of Standards and Technology — Huntsville, AL Jan 2025 - May 2025

- Developed Python/Bash scripts for entropy measurements, automating data collection and running test scripts across different environments, analyzing CPU jitter for security compliance with NIST SP 800-90B.

Frontend Developer | **GTA** @ University of Alabama, Huntsville — Jul 2023 - Dec 2023

- Developed hiring system in React and JavaScript. Improved review productivity by 50%. Worked on faculty collaboration tool improving user experience and code maintenance. Build REST APIs with NodeJS.
- Organized Cloud Computing labs for 100+ students, designing cloud assignments, offering support and grading.
- Taught Python and DSA courses to 80+ students, improving average scores by 25% from the previous semester.

Skills

Python, JavaScript, Java, SQL, Bash, TensorFlow, Keras, PyTorch, Scikit-learn, Hugging Face Transformers, LangGraph, LangChain, RAG, BERT, LLaMA, GloVe, Text Classification, Sentiment Analysis, Clustering, Vector Embeddings, Retrieval Pipelines, AWS (Lambda, Bedrock, EC2, S3, DynamoDB), Docker, Kubernetes, GitHub Actions (CI/CD), Jenkins, MLflow, React.js, Next.js, Node.js, Express.js, Django, Spring Boot, REST APIs, MySQL, MongoDB, DynamoDB, ChromaDB (Vector DB), SageMaker Model Monitor, Prometheus, CloudWatch, SageMaker Hyperparameter Tuning, Apache AirFlow

Project Experience

AI-Scraper | *LLM Agents, Optimization, Data Pipelines*
Developed LLM Agents that combined transformers with optimization routines to automate large-scale data acquisition under anti-bot constraints. Saved costs with 60% token reduction and fast inference with caching.

Top Match | *DistilBERT, Cosine Similarity, RAG, AWS*
Built a candidate-to-job matching platform engine that scrapes 1000+ jobs and matches the candidates resume with cosine similarity, both lexical and semantic. Levarge AWS services like S3, Lambda, DynamoDB. Deployed on FastAPI.

Real-Time Facial Emotion Classifier | *CNN, Transfer Learningc, TensorFlow, VGG19*
Trained a transfer learning model in TensorFlow leveraging CNN (VGG19) and optimization algorithms to reach 90% validation accuracy; deployed live inference with Dockerized Flask API and OpenCV integration.

Contextual Sentiment Analysis Model | *RNN, Bi-LSTM, GloVe, Regularization*
Implemented Bi-LSTM with Attention and GloVe embeddings for contextual text sentiment analysis; tackled over-fitting with 88% accuracy using L2 regularization and batch normalization. Securing top place in the course.

Certifications & Publications

AWS Certified Developer – Associate — Amazon Web Services (AWS) - Link
Machine Learning – Deep Learning based Cyberbullying Detection. — ResearchGate Publication - Link
Awarded for Outstanding Technical Leadership and Contributions at NIST — Testimonials available: Link