Sai Sri Swetha Battu

New York , US | sabattu@my.bridgeport.edu | 740-602-7120 | yourwebsite.com

linkedin.com/in/saisri-swetha-battu-889a0649 | github.com/Saisriswetha

Professional Summary

GenAI Engineer with 3+ years of experience and specializied in generative models and natural language processing, with hands-on expertise in LLMs (GPT, LLaMA, Falcon), MLOps, cloud deployment (AWS, Azure OpenAI), and data pipeline development. Demonstrated expertise in building Model Context Protocol (MCP) and Agent-to-Agent (A2A) integrations, as well as CrewAI frameworks for advanced automation. Adept at driving end-to-end AI initiatives, from research and prompt engineering to model optimization, project management, and stakeholder communication. Passionate about ethical AI and delivering robust, production-ready solutions that support business goals. Strong technical foundation in Python, JavaScript, SQL, and cloud platforms (AWS, GCP, Azure), alongside containerization (Docker, Kubernetes) and CI/CD best practices. Proven track record of boosting system efficiency by 25%, optimizing real-time data processing, and establishing end-to-end model observability.

Projects/Applied Experience

Cognitive HOI Reasoning – Research Project Apr 2023 – Mar 2024

- Developed and fine-tuned deep learning models for human-object interaction understanding (Bongard-HOI, HAKE datasets), addressing gaps between synthetic and real-world image cognition.
- Raised model accuracy from 55% to 80% on >2,000 annotated image pairs by integrating DVRL, advanced attention, and componential analysis.
- Engineered custom data augmentation, self-supervised representation learning, and interpretability workflows to analyze reasoning failures and bridge cognitive paradigms.

Personal Attorney – Legal RAG Assistant Apr 2024 – Nov 2024

- Designed a Retrieval-Augmented Generation (RAG) platform with ChromaDB, LangChain, and DeepSeek 8B (LoRA fine-tuning) to deliver state-specific DMV/legal support, reducing retraining costs by 70%.
- Deployed scalable REST APIs using FastAPI on AWS, handling 1,500+ user queries/month and maintaining 98% information retrieval accuracy across 50+ workflows.
- Led integration of prompt engineering, vector database optimization, API documentation (OpenAPI/Swagger), and MLOps (CI/CD for rapid releases).

SmartMed AI – Pill Detection App Jan 2025 – May 2025

- Engineered a computer vision pipeline using YOLO (v5-v12), ResNet, multi-scale detection, and attention modules for accurate identification and tracking of overlapping pills in pharmacy workflows.
- Achieved 91% test accuracy (up from 76%), improving operational reliability by 25% and automating processing of 10,000+ images/month.
- Implemented containerized API deployment (Docker, FastAPI, AWS), advanced tracking, and model explainability (GradCAM, albumentations) for resilient production inference.

Custom Sanskrit Transformer Jun 2025 – Aug 2025

- Built and trained a Transformer-based NLP model on ancient Sanskrit Vedic texts with a custom 25,000-token vocabulary, deploying advanced tokenization and architecture modifications to capture linguistic complexity.
- Optimized inference and training using NVIDIA GPUs, CUDA, cuDNN, and integrated TensorRT—reducing inference latency by 46% and boosting throughput by 39% over baseline; profiled and tuned matrix multiplications for efficient attention.
- Engineered kernel-level performance tuning, system profiling, and benchmarking for scalable, production-grade deployment; delivered robust real-time serving pipeline with <75ms inference times.

Time-Series Forecasting – Stock Market Analysis Sep 2025 – Running

- Built forecasting solutions for highly volatile financial data, comparing ARIMA/SARIMA, Meta Prophet, and XGBoost ensembles with engineered lag and window features.
- Increased accuracy from 68% (baseline) to 84%, reduced error by 31% during market swings, and built MLflow-tracked pipelines handling >100,000 data points/month.
- Delivered model reliability insights with prediction intervals, and automated forecasting process for scalable analysis.

Technologies

Programming Languages: Python | Java | C/C++ | JavaScript | SQL | NoSQL | Data Structures & Algorithms | Object-Oriented Programming | Clean Code | Design Patterns

Machine Learning & AI: Model Training & Deployment | Fine-Tuning (LoRA, transfer learning) | Optimization | Attention Mechanisms | Deep Learning | Neural Networks (CNNs, Transformers, LLMs) | Generative AI (GenAI) | Large Language Models (GPT, LLaMA, Falcon, DeepSeek) | NLP | Computer Vision | Prompt Engineering | Text Generation | Sentiment Analysis | Named Entity Recognition | Retrieval-Augmented Generation (RAG) | Synthetic Data Generation | Model Evaluation & Interpretability | Bias Control | Anomaly Detection | Ethical AI | Feature Engineering | Statistical Modeling | Time-Series Forecasting

Frameworks & Cloud Platforms: PyTorch | TensorFlow | Keras | scikit-learn | Hugging Face Transformers | LangChain | FastAPI | Flask | Pandas | NumPy | SQLAlchemy | AWS (S3, EC2, Lambda, SageMaker, Bedrock, Fargate) | Azure ML | GCP AI/ML | OpenAI API | Docker | Kubernetes | MLflow | Kubeflow | MLOps | Cloud Deployment | Microservices | REST API Development & Documentation | ChromaDB | DVC | TensorRT | CUDA | cuDNN

Strengths & Tools: Git | GitHub | CI/CD Automation | Docker | Kubernetes | MLflow Tracking | Postman | PyCharm IDE | ITSM Tools | Performance Profiling | Benchmarking | System Optimization | Code Reviews | Agile/Scrum | Project Ownership | Cross-functional Collaboration | Backlog Management | Production Monitoring | Problem Solving | Debugging | Attention to Detail

Education

University of Bridgeport, M.S. in Computer Science Aug 2023 – Dec 2025

- Focus on Artificial Intelligence, Full-Stack Development, and Cloud Computing
- Relevant Coursework: Advanced Machine Learning, Cloud Computing, Software Engineering

Adikavi Nannaya University (AKNU), BCom with Computer Science Jun 2019 – Jun 2023

- Completed coursework in Data Structures, Algorithms, Operating Systems, and Databases
- Academic Projects: Open Banking API Aggregator for Seamless Financial Data Integration

Experience

Starrez Developer, University of Bridgeport – Bridgeport, CT Aug 2024 – Apr 2025

- Managed and customized the StarRez housing portal to support student housing operations across semesters.
- Developed and deployed StarRez Wizards to streamline housing applications, automate deposit processing, and support semester-specific procedures, reducing manual workload and errors.
- Integrated housing workflows with AwareManager and UpKeep portals, accelerating maintenance coordination and facilities management response times.
- Built automated tools for dynamic reporting and portal updates using Python and AWS, which boosted operational efficiency and enabled real-time insights.
- Collaborated with cross-functional campus departments to optimize data integration and elevate user experience across multiple platforms.