VIVEKANAND SAHU

vsahu@ucsd.edu | +1 8588445141 | LinkedIn | GitHub

EDUCATION

University of California San Diego, *United States* *Sep 2023 - Mar 2025*

Master of Science in Machine Learning and Data Science [GPA: 3.63/4]

Courses: Deep Learning, AI Algorithms, ML Applications, NLP, Computer Vision, Statistics, Python, Data Analytics

EXPERIENCE

AI Engineer | ***UC San Diego,*** *California* *Oct 2024 – Mar 2025*

- Architected a Tutor Chatbot by building an end-to-end multi-agent RAG leveraging 6 LLMs using LangChain and CrewAI
- Obtained 3x precise responses and reduced hallucinations by 60% via LLM Fine-Tuning, RAGAS, and Prompt Engineering
- Ensured data privacy by performing RAG search using on-device Llama 3.2 and ChromaDB, while achieving 1.5x speedup by parallelly performing Web search using GPT-3 and web-scraping tool
- Integrated Streamlit UI, implemented CI/CD and scaled the system for 100 queries/sec by deploying it on AWS [GitHub]

Machine Learning Engineer | ***Alkermes,*** *Massachusetts* *Jun 2024 – Sep 2024*

- Enhanced drug discovery using an LLM foundation model to identify target genes associated with brain disorders
- Classified disease states with 0.9 F1 score by fine-tuning the model on 50k cells data using Distributed Multi-GPU training
- Automated 3 types of gene regulation hypotheses by deploying the model in production using Azure ML and Databricks
- Accelerated research timeline by 5 years and reduced costs by 50% by identifying 6 genes that shift diseased cells to a healthy state by a cosine distance of 0.3 [GitHub]

Senior Data Engineer | ***LTIMindtree,*** *Mumbai, India* *Jul 2021 – Jul 2023*

- Built a fraud detection system by analyzing 20K customer records and raising alerts via FastAPI-backed React dashboard
- Developed an ETL pipeline using Apache Spark and S3 to process 1M+ transactional, financial, and user behavior data
- Obtained true-positive-rate of 0.95 by training an XGBoost model on AWS SageMaker on the unified data warehouse
- Automated CI/CD using GitHub Actions and deployed the system on API Gateway and AWS Lambda
- Detected ~1k fraud transactions a day by scaling and monitoring with AWS Elastic Load Balancer and CloudWatch

PROJECTS

Interpreting & Optimizing Transformers | *Explainable AI, CUDA, Performance Tuning, Kubernetes* [GitHub]

- Identified subatomic particle-interactions by adding 4 physics-informed features and visualizing attention weights
- Uncovered binary dependency of particles by classifying particle jets with 0.92 AUC using Linear Attention mechanism
- Boosted model speed by 40% and cut peak memory by 3x using FlashAttention and Int8 Quantization on Nvidia A100 GPUs

Product Billing and Inventory Management | *Tableau, Time series, MySQL, Computer Vision*

- Automated product billing by identifying products with 0.81 mAP and updating the database using Yolo v4 model
- Enhanced supply-demand forecasting accuracy by 25% by designing and analyzing 4 real-time Tableau dashboards
- Improved operational efficiency by reducing the billing time by 85% leading to 30% increase in customer footfall

Typing Assistant Chatbot | *Auto-regression, Tokenization, Model Profiling, NLP*

- Developed a text-autocomplete system by building a GPT model and training on 220k-conversations Movie dataset
- Achieved a perplexity of 32 through Byte-Pair-Encoding tokenization and experimenting with local and global contexts
- Applied sparse attention to generate 22% faster single/multi-word suggestions, improving customer satisfaction by 15%

SKILLS

Programming:	Python, SQL, C, C++, Java, CUDA, R, PL/SQL, Git, Bash, JavaScript, HTML/CSS
Frameworks:	PyTorch, TensorFlow, CrewAI , LangChain, Hugging Face, Scikit-learn, NumPy, Pandas, ONNX
Generative AI:	Agentic AI, RAG, LLM Finetuning, Model Profiling, Ollama, OpenAI, Distributed Training
Databases:	Big Data – Spark, Hadoop | RDBMS – MySQL | NoSQL – MongoDB | Vector – ChromaDB, Pinecone
Cloud & Web:	AWS, Azure, GCP, Kubernetes, Docker, DevOps, CI/CD, MLflow, Streamlit, React, Fast API

PUBLICATIOS

[1] 'Interpreting Transformers for Jet Tagging' | **NeurIPS** [Link] *Dec 2024*

[2] 'Machine Learning based Automated Product Billing and Inventory Management' | **IEEE Xplore** [Link] *Jan 2023*

[3] 'Autonomous Tomato Harvester Using Robotic Arm and Computer Vision' | **Springer** [Link] *Jul 2021*