Sai Chaitanya Pachipulusu

Email : siai.chaitanyap@gmail.com
linkedin.com/in/psaichaitanya | GitHub | pachipulusu.vercel.app
Mobile : +1-551-344-5967

Machine Learning Engineer with 5+ years of experience architecting and deploying ML models, end-to-end deep learning systems, and Generative AI solutions (LLMs, RAG systems, AI Agents), building ETL pipelines, crafting data solutions & data strategy across many sectors

EXPERIENCE

- **Community Dreams Foundation** — USA
  Machine Learning Engineer — *Jul 2024 - Present*
  - Engineered multi-stage HR management system (15k+ resumes/month) using **E5-large embeddings & DeBERTa-v3-large cross-encoders**; achieved 94% top-10 candidate relevance (HR-validated) & 18% qualified shortlist uplift with 65% faster screening
  - Formulated contextual **Learning-to-Rank (LTR) algorithms with LambdaMART** reducing manual screening by 90% & hiring time by 40%; deployed models on **Kubernetes with KServe** for auto-scaling (70% resource utilization) & blue-green deployments
  - Pioneered on-premise resume screening by **distilling** Phi-3-vision (teacher) to **quantized** text-based Phi-3-mini student models (INT8, QLoRA for 4-bit); cut model size 85%, reduced inference latency by 5.2x, maintained 97% accuracy with strict sovereignty requirements
  - Optimized candidate communication (rejections & feedback) with **fine-tuned quantized Phi-3-mini (on-premise, LoRA, HITL)**; slashed HR admin work by 90%, lifted rejected candidate response by 40%, earned 96% positive feedback
  - Architected **RAG chatbot** with Llama-3-8B-Instruct, **hybrid search** (BM25+vector), E5-large embeddings, Weaviate **vector DB** on AWS SageMaker; achieved 92% response relevance (RAGAS evaluated), decreased tier-1 support tickets by 35%
- **CGI** — Bengaluru, India
  Associate Software Engineer / Data Engineer (Client: Shell Corporation) — *Sep 2020 - Jun 2022*
  - Engineered real-time sensor **data pipelines** (Kafka Connect, Spark Streaming) for 8 critical streams (400 events/sec), processing with PySpark, landing in BigQuery; mitigated data availability lag from batch to less than 5 mins & lessened false anomaly alerts by 30%
  - Implemented **regression and transfer-learning models** for predictive maintenance using Azure ML for experiment tracking & endpoint creation; shortened model deployment time by 50% through containerization (Docker) and Kubernetes (AKS) orchestration
  - Led migration of 52 legacy servers to AWS EC2 (t3.xlarge) using Terraform for IaC; generated savings on monthly operational costs by $8k and ensured 99.9% uptime over a 6-month period
  - Optimized **Databricks Medallion architecture** with robust schema validation and incremental processing; decreased data pipeline failures by 75% (12 to 3 weekly) and trimmed data recovery time from 4 hours to 45 minutes
  - Awarded '**Best Employee of the Quarter**' (Q4 2021) for leading 30% of the data migration effort, enhancing project efficiency, and fostering cross-functional team collaboration
- **Imbuedesk Pvt. Ltd** — Hyderabad, India
  Machine Learning Engineer — *May 2018 - Aug 2020*
  - Fine-tuned ResNet-50 & custom VGG-16 on AWS (EC2 GPUs, SageMaker) for facial expression recognition (97% FER2013)
  - Designed an **image processing pipeline** with Tesseract OCR for vehicle ID recognition (5k plates/day), orchestrated with Kubernetes
  - Constructed and rolled out **predictive maintenance dashboards** (Python, Flask/Dash, AWS Elastic Beanstalk) from multi-modal sensor data, reducing equipment downtime by an estimated 28%
  - Built a **Kafka-based image processing pipeline** processing 35 MB/hour, reducing latency from 66% per image with a 3-node consumer topology. Implemented back-pressure handling for peak traffic periods (7AM-9AM) when processing volume increased by 300%

TECHNICAL PROJECTS

- **Career Roadmap Generator** | Generative AI, RAG System, Streamlit, ChromaDB | Huggingface
  Integrated GPT 3.5, GPT-4o models with ChromaDB for context-aware recommendations to the job description achieving 0.87 hit@5 score on 200 anonymized job transitions; async implementation diminished API latency from 12 to 4 seconds for 10+ concurrent users
- **SideBuilds.space** | React, Node.js, Chakra UI, CockroachDB, Vercel, Render, Stripe | www.sidebuilds.space
  An end-to-end public platform for showcasing and iterating on side projects; embedding ideas, showcasing MVPs, hosting mini-apps (AI-powered apps, RAG prototypes), and able to sell them on the platform for a small commission
- **AI vs Human: Exploring the limits of Machine Intelligence** | Selenium, SentenceTransformers, SVM, Python | GitHub
  Developed a model to distinguish between human and ChatGPT answers, and identified key linguistic markers distinguishing AI texts while model accomplished a notable accuracy of 90% and F1 score of 89%

EDUCATION

- **Stevens Institute of Technology** — Hoboken, NJ
  Master of Science in Machine Learning; CGPA: 3.9 — *Sep 2022 - May 2024*
- **Sreenidhi Institute of Science and Technology** — Hyderabad, India
  Bachelor of Technology in Information Technology — *Aug 2016 - May 2020*

SKILLS & ACCOMPLISHMENTS

- **Technical**: Python (primary), Java, SQL, Statistical Analysis, Probability, Machine Learning Algorithms, MLOps, Neural Networks, CNN, RNN, LSTM, GANs, VAEs, Autoencoders, Transformers, C++, Selenium, BeautifulSoup, R, Linux/UNIX, Time Series, Predictive Models, Data Science, Natural Language Processing, Data structures and Algorithms
- **Frameworks/Libraries**: PyTorch, TensorFlow, Snowflake, Streamlit, Keras, Scikit-learn, XGBoost, PySpark, NumPy, Pandas, spaCy, NLTK, Matplotlib, Hadoop, Seaborn, LightGBM, Plotly, JAX, RAGAS, CrewAI, Airflow, Scala
- **Generative AI**: LLMs, RAG, LangChain, LangGraph, Agentic AI, Llamaindex, Weaviate, Pinecone, ChromaDB, QLoRA/PEFT, FAISS
- **Database Management Systems**: Relational-Databases (MySQL, PostgreSQL), MariaDB, NoSQL, MongoDB, CockroadDB, Supabase
- **Tools/Platforms**: AWS (S3, EC2, Lambda, SageMaker, Bedrock), Databricks, Tableau, PowerBI, Visual Studio, Excel, Version Control (Git), CI/CD (Docker, Kubernetes), A/B testing, Terraform, BigQuery, Apache Spark, Kafka
- **Soft skills**: Communication skills, Cross-functional teamwork, presentations, leadership, mentoring, Critical thinking, Decision making
- **Certifications**: Getting Started with PowerBI (Coursera), Machine Learning Specialization (Coursera), Python for Everybody (Coursera), SPSS certified professional in Data mining and Warehousing, AWS Certified Machine Learning – Specialty (in progress)