Saketh Ram Kalavakuntla

+1 317 9516107 | sakethram8524@gmail.com | LinkedIn | GitHub | United States

PROFESSIONAL SUMMARY

Motivated and results-driven Machine Learning Engineer with 3+ years of industry and research experience building scalable AI solutions across GenAI, NLP, computer vision, and MLOps. Proven success in deploying deep learning models and LLM-integrated systems using Python, PyTorch, TensorFlow, and Hugging Face, with strong proficiency in LangChain, OpenAI APIs, and cloud platforms like AWS and GCP. Adept at developing production-grade pipelines, automating model deployment with Docker and CI/CD, and integrating AI into end-user tools with Flask and RESTful APIs. Passionate about leveraging AI-assisted development, agentic architectures, and secure engineering practices to drive innovation and impact in fast-paced, collaborative environments.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer | G-Technologies Jan 2023 - Dec 2024

- Designed a GenAI-enabled ECG interpretation system combining EfficientNet-based image classification with transformer-generated diagnostic summaries, reducing manual report creation time by 60%.
- Achieved 92%+ F1-score on a 4-class ECG dataset using data augmentation techniques (elastic transforms, contrast shifts), improving minority class recall by 23%.
- Built a modular PyTorch and Hugging Face pipeline to handle preprocessing, classification, and narrative generation, improving processing efficiency by 45%.
- Managed 100+ experiments using MLflow, ensuring reproducibility and structured model versioning across the development lifecycle.
- Integrated OpenAI GPT-3.5 APIs for text summarization, achieving an 18% improvement in clarity (ROUGE/BLEU) vs. rule-based baselines.
- Contributed to internal GenAI experimentation (Fine-tuned LLMs), evaluating prompt tuning strategies and vector search components for future diagnostic Q&A tools.

Machine Learning Engineer | The Cigna Group Feb 2020 - Mar 2023

- Developed and deployed ML models using TensorFlow, PyTorch, and Scikit-learn to enhance predictive accuracy by 28% across critical healthcare workflows.
- Built scalable ML pipelines encompassing data ingestion, training, validation, and deployment, reducing end-to-end model delivery time by 35% through automation.
- Designed and deployed real-time inference services using REST APIs on AWS, enabling sub-200ms response latency for production applications..
- Optimized GPU workloads via profiling, cutting training/inference time by **40%** and accelerating iteration cycles.
- Containerized ML solutions using Docker and deployed via AWS, enabling reproducible and resilient model delivery in live clinical systems.
- Instrumented models with **MLflow** and custom monitoring to reduce drift incidents by **30%**, while collaborating cross-functionally to align ML outputs with business logic, improving project delivery velocity by **25%**.

TECHNICAL SKILLSET

- **Technologies:** Machine Learning | Deep Learning | Natural Language Processing (NLP) | Computer Vision | MLOps | Agentic AI | Data Science | Data Engineering | Big Data | Cloud Computing (AWS, GCP, Azure) | Statistical Modeling | A/B Testing | Model Optimization | ETL Pipelines | Data Modeling | Data Manipulation | Agile | SaaS | Linux | Containerization (Docker, Kubernetes)
- **Programming Languages:** Python | C++ | R | SQL | LaTeX | JavaScript | HTML | SAS| Scala |Go
- **Databases:** PostgreSQL | MongoDB | Cassandra | Neo4j | Google BigQuery | SQL Server | Apache Spark | Ray | Apache Flink
- **Tools & Frameworks:** TensorFlow TensorFlow | PyTorch | Keras | Hugging Face Transformers | MLflow | LangChain | LangGraph | OpenAI API | ChromaDB | FAISS | Gradio | Streamlit | Pandas | NumPy | Flask | Jupyter | Git | Power BI | Tableau | Databricks | Hadoop | Kafka | NGINX | Visual Studio Code | Google Data Studio

PERSONAL PROJECTS (SELECTED)

- **DocuQuery – Internal Document Q&A Chatbot** *(Flask | LangChain | FAISS | OpenAI | RAG | AWS)***:** Built a PDF-based RAG system using LangChain and OpenAI to enable semantic search and contextual Q&A from internal documents.
- **SmartCover AI** *(Flask | LLaMA 3 | LangChain | ChromaDB | OpenAI | AWS)***:** Developed a GenAI tool to generate personalized cover letters by aligning resumes with scraped job descriptions using LangChain, LLaMA 3 and ChromaDB.
- **Photo Memories App – CLIP-Powered Image-to-Text Retrieval** *(CLIP | Gradio | Python | NumPy | OpenAI)*: Created a memory slideshow generator that retrieves images from user uploads based on text prompts using CLIP embeddings.
- **Multi-Agent Book Genre Classification Bot** *(Flask | BERT | Hugging Face | REST APIs | AWS)*: Fine-tuned a BERT model to classify book summaries by genre; built RESTful Flask APIs and deployed on AWS.
- **Smart Retail Forecasting Engine (Walmart) (***Python | Prophet | Pandas | Power BI)*: Built a Prophet-based time series model to forecast Walmart sales trends. Integrated results into Power BI dashboards for actionable retail insights.
- **Metaphor Detection – Linguistic Pattern Classifier** *(PyTorch | DistilBERT | NLP | Hugging Face)***:** Trained a DistilBERT model to classify metaphorical expressions in text for improved linguistic understanding and pattern detection.
- **Dynamic Pricing Experimentation Engine** *(Python | Statsmodels | A/B Testing | Pandas):* Built a statistical modeling pipeline to evaluate pricing experiments across customer segments using regression and lift analysis.

EDUCATIONAL BACKGROUND

Purdue University (Dean's Scholarship Recipient) Dec 2024
Master of Science in Computational Data Science (Computer Science, Mathematics, ML, AI)