Sasi Pretham Nakka

+1(469-638-3071) | sasipreetham.nakk@gmail.com | LinkedIn:sasipretham

Professional Summary

- **AI/ML Engineer** with 4+ years of experience building and deploying **ML** and **deep learning** solutions across **finance**, **healthcare**, and **retail** using **Python**, **PyTorch**, and **TensorFlow** for real-time, scalable pipelines.
- Created **GenAI** pipelines using **LangChain**, **OpenAI APIs**, and **FAISS/Pinecone**. Built AI assistants for summarization and support, reducing manual workload by 30% in **education** and internal **knowledge tools**.
- Built **NLP** workflows with **Transformers**, **spaCy**, and **NLTK** for classification, summarization, and entity extraction. Improved audit workflows by automating document parsing in **compliance** and **risk analysis** pipelines.
- Applied **time-series forecasting** models like **ARIMA**, **LSTM**, and **Prophet**. Integrated predictions with **Snowflake**, **BigQuery**, and **Kafka/Kinesis** for real-time dashboards and live business **decision engines**.
- Designed **CI/CD** workflows using **GitHub Actions**, **Docker**, and **Kubernetes**, cutting deployment time by 40%. Deployed models on **AWS SageMaker** and **Azure ML** with **MLflow** and **Prometheus** monitoring.
- Built **explainable AI** tools using **SHAP** and **LIME**. Collaborated with cross-functional teams to align model KPIs with business goals and led benchmarking to standardize experimentation and **model evaluation** processes.

Skills

- **Languages**: Python, SQL, C++, Bash, Java
- **Frameworks**: PyTorch, TensorFlow, Scikit-learn, Keras, Transformers (HuggingFace)
- **GenAI & NLP**: LangChain, OpenAI APIs, RAG, BERT/GPT models, Prompt Engineering
- **MLOps & Deployment**: AWS SageMaker, Airflow, MLflow, Docker, Kubernetes, GitHub Actions, FastAPI
- **Cloud Platforms**: AWS (Lambda, S3, ECS), GCP (BigQuery, Vertex AI), Azure (Functions, DevOps)
- **Monitoring & Compliance**: Model Drift Detection, Explainability (SHAP/LIME), HIPAA, AWS KMS
- **Data Engineering**: Apache Spark, Trino, Snowflake, PostgreSQL, MongoDB, NoSQL
- **Visualization**: Streamlit, Tableau, Seaborn, Matplotlib
- **Tools**: Git, JIRA, VS Code, Notion, Slack, Linux

Publications

- **Deep Learning for Underwater Condition Monitoring of Offshore Energy Installations**, 2025 IEEE Green Technologies Conference – *Awarded 2nd Best Presentation.*
- **Underwater Robot Design for Monitoring of Offshore Energy Structures**, IEEE SoutheastCon 2025.

Certifications

- Generative AI with Large Language Models – DeepLearning.AI
- AWS Machine Learning Learning Plan – AWS Skill Builder
- Google Cloud: Generative AI Fundamentals
- Microsoft Certified: Azure Administrator Associate (AZ-104)
- Microsoft Certified: DevOps Engineer Expert (AZ-400)

Experience

AI/ML Engineer **Jan 2025 – Present**

Target *Chicago, IL*

- Built an internal **LLM-based chatbot** using **LangChain**, **OpenAI GPT-4**, and **Pinecone**, reducing internal support tickets by 40% and accelerating onboarding processes with document-aware response generation.
- Developed robust **Retrieval-Augmented Generation (RAG)** pipelines using **FAISS**, enabling knowledge-grounded answers in document-heavy workflows, significantly improving information retrieval accuracy and reducing manual search overhead.
- Fine-tuned **Transformer models** (**BERT**, **RoBERTa**) for domain-specific classification in financial text, achieving an 18% improvement in F1-score and reducing manual effort for fraud flagging and compliance alerts.
- Led end-to-end **MLOps pipeline development** using **Docker**, **Kubernetes**, and **MLflow**, enabling automated deployment, versioning, and monitoring for over 10 production models with zero-downtime releases.
- Designed cost-efficient model inference endpoints using **FastAPI** and **AWS ECS**, implementing autoscaling and caching to maintain sub-250ms latency and cut infrastructure costs by 22%.
- Implemented **drift detection** using statistical hypothesis tests and **Prometheus**-based alerting, ensuring timely interventions and minimizing model degradation across high-impact scoring pipelines.
- Built **time-series forecasting models** using **LSTM** and **Prophet**, improving sales and inventory planning accuracy for retail product categories, resulting in measurable supply chain efficiency.
- Engineered reusable **feature pipelines** with **pandas**, **NumPy**, and **featuretools**, reducing model training time by 30% while supporting both batch and real-time inference systems.
- Orchestrated ETL and training pipelines using **Airflow** and **Apache Spark**, automating model retraining and nightly inference cycles, while maintaining high throughput for production-scale data.

ML Engineer **Aug 2023 – May 2025**
UNC Charlotte *Charlotte, NC*

- Engineered underwater robots using **YOLOv5 (CSPDarknet, PANet)** for real-time visual inspection, reducing infrastructure monitoring costs by 93% and improving fault detection in harsh environments via robust **computer vision**.
- Partnered with **Duke Energy Innovation Lab** to build **LSTM-based anomaly detection models**, forecasting failures 7 days in advance and minimizing downtime in high-risk energy infrastructure.
- Built and deployed **GPT-powered educational assistants** fine-tuned on lecture transcripts and curricula, improving student engagement by 35% through real-time feedback and personalized learning in AI-integrated classrooms.
- Developed **recommendation systems** using matrix factorization for adaptive STEM platforms, enhancing content retention and reducing mastery time through data-driven personalization.
- Co-authored and presented a peer-reviewed paper on explainable forecasting at **IEEE SoutheastCon 2025**, earning a Best Poster nomination among 200+ research submissions.
- Utilized **on-chip debug tools**, oscilloscopes, and lab instrumentation to optimize inference latency on **embedded ARM prototypes**; led hands-on workshops on **LLMs**, **prompt engineering**, and **GenAI tools** for 30+ graduate students.
- Benchmarked open-source **LLMs** (LLaMA, Mistral, Falcon), analyzing hallucination rates, latency, and response quality to recommend cost-efficient, academically aligned chatbot deployment strategies.

Machine Learning Engineer – Client: Earn In **Mar 2022 – Jul 2023**
Harman International *Bangalore, India*

- Built computer vision models using **PyTorch** for product detection on retail shelves, improving audit automation by 25% and enabling scalable **edge deployment** across field sites with low-latency inference capabilities.
- Integrated **YOLOv5** object detection pipeline with **Raspberry Pi** for real-time warehouse monitoring; deployed **on-device models** that reduced manual tracking effort and increased response accuracy in operational audits.
- Developed **NLP pipelines** using **BART** and **T5** models for report summarization and question answering, automating compliance report generation and increasing throughput for healthcare document processing.
- Engineered **voice-to-text** NLP system using **Whisper** and fine-tuned **BERT**, enabling voice-activated form submission in enterprise apps and reducing entry friction for mobile field users.
- Deployed ML models using **Azure ML Pipelines** and scheduled workflows via **Function Apps**, automating batch scoring with seamless integration into **Azure Blob Storage** pipelines for large-scale data persistence.
- Built operational dashboards in **Power BI** using **Snowflake** as a backend, empowering regional managers to track live KPIs such as inventory mismatch, out-of-stock events, and shelf compliance metrics.
- Designed a **recommendation engine** using **collaborative filtering** and matrix factorization techniques to analyze user clickstream data, improving personalization and increasing upsell conversion by 12%.
- Optimized training pipelines with **Optuna** and **Ray Tune**, reducing hyperparameter tuning cycles while improving model performance metrics for both **computer vision** and **NLP tasks**.
- Established test coverage for ML pipelines using **pytest** and **GitHub Actions**, while validating data integrity using **Great Expectations** for robust and reproducible ML system behavior.

Data Scientist – Client: Deloitte **Jan 2021 – Mar 2022**
Hyderabad, India

- Developed **churn prediction models** using **XGBoost** and **LightGBM** on large-scale telecom datasets, improving customer retention strategies and enabling proactive outreach campaigns for at-risk segments.
- Automated end-to-end **data preprocessing** workflows in **Python**, reducing manual overhead by 35% and accelerating model development cycles across multiple analytics use cases.
- Designed interactive dashboards in **Tableau** to visualize customer **KPI trends**, enabling leadership to make data-driven decisions for marketing, operations, and user engagement initiatives.
- Built **sentiment analysis pipelines** using **TextBlob** and **VADER** on social media datasets, helping brand teams assess customer perception and adapt communication strategies in near real-time.
- Constructed complex **SQL queries** and automated **ETL pipelines** to consolidate engagement metrics, standardize reporting, and support compliance with internal data quality standards.
- Deployed machine learning models via **Flask APIs** with **cron-based** scoring jobs on on-premise servers, ensuring predictable performance and scheduled inference across client applications.
- Collaborated with **UI/UX teams** to integrate ML insights into user-facing dashboards, improving client adoption and surfacing relevant actions based on predictive model outcomes.
- Contributed to **data governance reviews** and GDPR compliance assessments by ensuring anonymized outputs, secure data processing, and full transparency of model usage in reporting workflows.
- Supported client workshops by translating business problems into **machine learning solutions**, presenting insights to executive stakeholders, and iterating on feedback for real-world deployment.

Education

University of North Carolina at Charlotte **Aug. 2023 – May 2025**
Master of Science in Computer Science - Minors: Artificial Intelligence, Robotics and Data Science *Charlotte, North Carolina*

- Related Coursework: Machine Learning, Deep Learning, Natural Language Processing, Big Data Analytics, Cloud Computing, Knowledge Discovery, Advanced Algorithms, Software Systems Design, Database Systems, Visual Analytics, Information Security