SAIGANESH NELLORE

+1 (773) 822-5301 | nsaiganesh2003@gmail.com | linkedin.com/in/saiganeshn/ | github.com/SAIGANESH02 | saiganesh02.github.io/SaiGanesh

PROFESSIONAL SUMMARY

Lead Machine Learning Engineer experienced in building and optimizing AI systems, including agentic voice automation for sub second patient conversations, LLM analytics, and data pipelines, across cloud native stacks. Strong in NLP and GenAI, LLM evaluation, RAG, and computer vision, with a track record of shipping HIPAA aligned, cost efficient solutions at scale on AWS.

PROFESSIONAL EXPERIENCE

XSELL Technologies **Chicago, IL, USA**
Lead Machine Learning Engineer *June 2025 - January 2026*
- Built and scaled Agentic Voice (STT, TTS, LLM dialog) to enable natural, sub-second patient conversations in prod facing 1000s of customers a day.
- Fine-tuned and compressed domain LLMs (Llama-3.1, GK) to deliver faster, cheaper inference while preserving conversational quality and safety.
- Productionized multi-model inference on AWS (Bedrock, SageMaker, Triton) with autoscaling, observability, and cost controls to uphold latency.
- Implemented HIPAA-aligned data flows with PHI redaction, role-based access, and secure logging to meet privacy, audit, and compliance needs.
- Scaled inference to 100k+ req/day via Ollama+vLLM+TensorRT on AWS EC2/SageMaker and Compressed LLMs with RLHF, GRPO, distillation, etc., reducing OOB model cost from $300 to $40/day

Vanguard **Philadelphia, PA, USA**
Machine Learning Engineer *September 2024 - January 2025*
- Optimized RAG evaluations for a customer representative use case by leveraging LLMs to reduce testing time and lower developer and SME costs.
- Saved $300K+ in SME costs by automating question creation, using robust metrics (accuracy, diversity, relevance) to refine RAG at scale.
- Launched an LLM-driven pipeline for RAG testing, generating chunk-based questions that enhance coverage, variety, and real-time insights.
- Applied advanced chunking (line, para, LLM) and specialized prompts to ensure test-question accuracy, diversity, and relevance across documents.

Paramount **Des Plaines, IL, USA**
AI Team Lead Intern - Conversational AI *June 2024 - December 2024*
- Led a team of 3 Engineers to build a scalable Voice AI Sales Agent that handles simultaneous calls, and dynamic conversations with zero human input.
- Managed tasks, planning, and cross-team collaboration, integrating Azure services and Twillio for robust components and increased system uptime.
- Fine-tuned Open AI realtime model to improve TTS humanization, achieving higher user satisfaction through context-aware, natural responses.
- Elevated system performance by continuously aligning model with feedback, ensuring real-time insights and seamless customer experiences.

zoho **Chennai, Tamil Nadu, India**
Member Technical Staff - AI R&D *December 2021 - July 2023*
- Applied Symbolic AI domain knowledge to enhance DL for financial fraud detection, achieving a 30% gain via gradient-based math refinements.
- Confronted data scarcity in ML, achieving ~80% accuracy in 12 epochs with a single-layer network, significantly reducing training costs.
- Developed OtterTune, an ML-driven Automated Database Management Tuning tool that reuses data from past sessions to improve knob settings.
- Configured sub-minute DB configuration within 94% of expert setups and improved latency/performance by up to 80% over default settings.

EDUCATION

Northwestern University **September 2023 - December 2024**
Master's, Artificial Intelligence *GPA: 3.95*
- Course work: Machine Learning, Intro to AI, Human Computer Interaction, Scalable Software Architectures, Deep Learning, Causal Inference, Data Science Seminar, Natural Language Processing (NLP), Frameworks of AI, Knowledge Representation and Reasoning, AI Industry Capstone.
- Engineering Management (Minor): Technology Venture Capital Investing, Decision Tools for Managers, Product Management.

Amrita School of Engineering **May 2019 - December 2022**
Bachelor's, Computer Science
- Course Work: Data Structures and Algorithms, Linear Algebra, Probability, Statistics, Natural Language Processing, Mathematics for ML, Computer Networks, Financial Time Series Analysis, Signal Image Processing, Operating Systems, Big Data & Database Management.

PROJECTS & OUTSIDE EXPERIENCE

ResumeBoost AI - Cloud-native Application to Optimize Resumes for ATS using AWS - *Link to project*
- Built the tool using AWS Lambda Functions for PDF parsing, job description scraping, and AI analysis with an Open AI GPT-4 API for AI modeling.
- Employed API Gateway to route client requests, linking streamlined Lambda workflows for real-time scraping, ingestion, and GPT-based feedback.
- Leveraged AWS S3 for retrieving PDFs, enabling concurrent execution via orchestrator, and delivering ATS optimization insights via Streamlit.

Same Same Collective - Chatbot tailored for LGBTQI+ youth in South Africa and Zimbabwe - *Link to project*
- Achieved 86% suicidal message detection in the SameSame Collective's chatbot via a fine-tuned BERT, boosting real-time interventions.
- Deployed automated empathetic responses and alert systems for high-risk users, ensuring timely and appropriate mental health support.
- Preprocessed data and created specialized prompts, significantly enhancing emotion classification and empathetic chatbot engagement.

Hand Gesture Detection Based Real-Time Indian Sign Language Recognition Chat System - *Link to project*
- Integrating ISL into e-Governance, leveraging Sign Language Recognition to provide accessible services for deaf individuals and inclusivity.
- Developed Python scripts via OpenCV & MediaPipe Holistic for pose/hand detection, used keyframe selection to generate refined videos.
- Enhanced Computer Vision accuracy by modifying Inception-ResNet, boosting detection precision by 20% and reducing false positives by 15%.

SKILLS

- **Programming Languages:** Python, Java, R, MATLAB, Git, Scala, JavaScript, C/C++
- **Frameworks:** LangChain, Tensorflow, Pytorch, NumPy, Pandas, Python NLTK, Scikit-learn, OpenCV, Keras, ROS, FastAPI
- **ML Models / Architectures:** LLM, RAG, GPT (ChatGPT, GPT-3.5, GPT-4), BERT, RoBERTa, T5, Transformers, CNN, GAN, U-Net, ResNet, YOLO-V4, LSTM, ARIMA, RCNN, Splinter, Q-Learning, DDQN, Flask, Decision Trees, Clustering, Regression, Predictive Models, forecasting
- **DevOps / MLOps:** AWS, Microsoft Azure, Google Cloud Platform, Docker, Kubernetes, CI/CD, GitHub Actions, Linux Bash, Apache Spark, Jenkins, ETL, MLflow, DVC, Dagshub, Apache Airflow, SageMaker, Databricks, LLMOps, Version Control, GCP, containerization, Cloud Computing, Kubeflow, Cloud Technologies, Cloud Infrastructure, Websockets, Web Services, Microservices
- **Databases / Big Data Technologies :** SQL, NoSQL, MongoDB, Hadoop, Spark (MLLlib), PySpark, Power BI, BigQuery, Snowflake, Postgres