SAIGANESH NELLORE

+1 (773) 822-5301 | nsaiganesh2003@gmail.com | linkedin.com/in/saiganeshn/ | github.com/SAIGANESH02 | saiganeshn-ai-portfolio.vercel.app/

PROFESSIONAL SUMMARY

Lead Machine Learning Engineer with experience building production-grade GenAI systems across healthcare, fintech, and enterprise AI environments. Shipped agentic voice platforms, RAG evaluation pipelines, and LLM inference stacks supporting 100K+ daily requests while cutting model costs by up to 85%. Deep expertise across NLP, LLM evaluation, RAG, and MLOps, with a consistent focus on latency, safety, compliance, and cost efficiency in cloud-native systems. Seeking a Senior or Lead ML Engineer role focused on deploying scalable GenAI products in regulated or high-impact domains.

SKILLS

- **ML and GenAI:** LLM, RAG, LLM Evaluation, NLP, Computer Vision, Prompt Engineering, Model Compression, Model Distillation, RLHF/GPRO
- **Frameworks:** Pytorch, Tensorflow, Transformers, LangChain, LangGraph, Scikit-learn, NumPy, Pandas
- **Serving and MLOps:** vLLM, Sagemaker, AWS Bedrock, CI/CD, Docker, Kubernetes, Jenkins, AWS, Apache Spark, Microsoft Azure, Google Cloud Platform, GitHub Actions, ETL, MLflow, Apache Airflow, Databricks, LLMOps, Version Control, GCP
- **Data and Databases:** SQL, Postgres, MongoDB, Spark, Snowflake, BigQuery, Power BI
- **Programming Languages:** Python, Java, Git, JavaScript, C/C++

PROFESSIONAL EXPERIENCE

XSELL Technologies **Chicago, IL, USA**
Lead Machine Learning Engineer *June 2025 - January 2026*
- Built and scaled a production agentic voice platform (STT, TTS, LLM dialog) enabling sub-second patient conversations across thousands of daily users in regulated healthcare environments.
- Scaled LLM inference to 100K+ daily requests using vLLM, and model compression techniques, reducing out-of-box inference costs from $300/day to $40/day.
- Fine-tuned and distilled domain LLMs (Llama 3.1, Qwen 3) to deliver faster, cheaper inference while preserving conversational quality and safety with HIPAA-aligned data flows with PHI redaction
- Productionized multi-model inference on AWS (Bedrock, SageMaker, Triton) with autoscaling, observability, efficient GPU usage, and cost controls to uphold latency.
- Built evaluation framework and CI/CD with github actions for voice agents with auto-labeling, drift detection, and guardrails to improve accuracy, robustness, and trust.

Vanguard **Philadelphia, PA, USA**
Machine Learning Engineer *September 2024 - January 2025*
- Led a team of five as project manager, developing a scalable evaluation framework to generate high-quality test questions for optimizing RAG performance for the financial customer representative use case.
- Reduced RAG evaluation costs by $300K+ by automating SME-dependent question generation using LLM-driven pipelines and quality metrics.
- Built an LLM-driven pipeline for RAG testing with advanced chunking strategies (sentence, paragraph, LLM), and specialized prompts to enhance coverage, ensure test-question accuracy, diversity, and relevance across documents.

Paramount **Des Plaines, IL, USA**
AI Team Lead - Conversational AI *June 2024 - December 2024*
- Led a 3-engineer team to deliver a fully automated Voice AI sales agent for Car dealership sales and service appointment automation which is capable of handling concurrent calls and dynamic conversations without human intervention.
- Customized OpenAI realtime model to improve TTS humanization, achieving higher user satisfaction by continuously aligning model with feedback conducting A/B testing, improving LLM accuracy to 98% and sub-100ms latency, ensuring real-time insights and seamless customer experiences reducing fallback rate by 46%.
- Managed tasks, planning, and cross-team collaboration, integrating Azure services, Whisper, Twilio, Replit, and LangChain with CRM and ML infrastructure for robust components and increased system uptime.

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.
- Built an ML-driven automated database tuning system (OtterTune) that reused historical session data to optimize configuration knobs, achieving up to 80% performance gains over default setups.

OTHER EXPERIENCE

Data Scientist @ Why of AI - RaceGPT: NextLap's Race Radio Intelligence Copilot - *Link to project*
- Built an end-to-end race comms intelligence pipeline (denoise, diarization, alignment, transcription, embeddings, RAG) to convert multi-channel radio into citation-grounded NL search and strategy answers.
- Developed competitor-analysis tooling over driver/crew transcripts to surface key calls (cautions, tire, fuel, issues) and reduce manual review for strategists and analysts.
- Architected PoC blueprints and data/LLM workflows to standardize delivery and decision support across race programs, accelerating repeatable AI deployments.

Workflow-Orchestrated RAG Chatbot - *Link to project*
- Built an end-to-end RAG agent (n8n, PostgreSQL + PGVector) to automate article ingestion and deliver document-grounded conversational Q&A.
- Designed modular ingestion and chat workflows with clear schemas (metadata, embeddings, chat histories) to improve maintainability, traceability, and extensibility.
- Containerized n8n, Postgres, and Adminer with Docker Compose to simplify deployment and debugging, enabling simplest local setup for rapid iteration.

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.

EDUCATION

Northwestern University **September 2023 - December 2024**
Master's, Artificial Intelligence *GPA: 3.95*

Amrita School of Engineering **May 2019 - December 2022**
Bachelor's, Computer Science *GPA: 3.65*