Venkatesh Shanmugam

Virginia US | +1 (703) 216-2540 | svenkatesh.js@gmail.com | LinkedIn | GitHub | Portfolio

SUMMARY

Founding Machine Learning Engineer with 4+ years building production-scale ML systems serving 1K+ RPS. Led end-to-end ML lifecycle from data architecture to model deployment, achieving 78% ROC-AUC in predictive models and $6K annual cost savings through infrastructure optimization. Proven expertise in distributed training, MLOps, and Agentic AI systems with experience scaling ML operations from inception to production.

TECHNICAL SKILLS

- **Programming**: Python, SQL, C++, Java, R, Bash
- **ML/AI Frameworks**: TensorFlow, PyTorch, scikit-learn, Keras, LangChain, Transformers, DeepSpeed
- **Cloud & MLOps**: AWS (S3, Glue, Athena, EC2, SageMaker), GCP (Vertex AI), Docker, Kubernetes, MLflow, CI/CD
- **Data Engineering**: Apache Spark, Kafka, Airflow, ETL Pipelines, Distributed Training, Model Monitoring, A/B Testing
- **AI Specializations**: Deep Learning, NLP, Computer Vision, Federated Learning, Agentic AI, Graph-RAG, Model Optimization (LoRA, PEFT), Neo4j (Graph DB), Elasticsearch

WORK EXPERIENCE

ScriptChain Health **May 2024 - Present**
Founding Machine Learning Engineer *Washington DC*

- Architected and deployed a LLM based deep learning model for 30-day hospital readmission prediction, achieving 78% ROC-AUC, enabling timely clinical interventions and reducing readmission risks at scale
- Designed and implemented a distributed training pipeline with DeepSpeed and multi-GPU clusters, accelerating model training time by 67% (from 72 to 24 hours) to support scalable healthcare workloads
- Engineered CI/CD workflows using Cloud Build, reducing model deployment time by 94% (4 hours to 15 minutes) while maintaining sub-200ms inference latency at 1,000+ requests per second, ensuring robust real-time performance
- Automated model promotion and monitoring in Vertex AI with MLflow A/B testing, improving precision-recall AUC by 6 percentage points and reducing false positive rates by 8%, facilitating continuous model quality enhancement
- Developed and deployed an agentic AI Graph-RAG recommendation system leveraging autonomous AI agents for adaptive information retrieval, boosting user engagement by 25% and recommendation relevance by 18%

Tata Consultancy Services **Apr 2021 - Aug 2023**
Machine Learning Engineer *Chennai, India*

- Engineered and deployed scalable automation frameworks across 14+ client teams, reducing manual operational effort and cutting infrastructure costs by 40% by optimizing resource allocation and automating complex workflows, directly supporting future ML model integration
- Implemented CI/CD pipelines for automation scripts, reducing release cycles by 50% and improving code quality through automated testing, directly transferable to accelerating ML model iteration and deployment.
- Developed 12+ software automation bots with seamless system integrations, improving process turnaround times by 30% and enabling scalable operations, demonstrating expertise in building robust, integrated software solutions for automated ML deployment and monitoring pipelines

EDUCATION

George Washington University **Aug 2023 - May 2025**
Master of Science, Computer Science

- **GPA:** 3.88/4.0

SRM University **Aug 2016 - May 2020**
Bachelor of Technology, Computer Science

- **GPA:** 3.5/4.0

PROJECTS

Agentic Graph RAG for Building codes | https://vabuildingcode.netlify.app/

- Architected production-scale multi-agent AI system with 5 specialized agents using LangGraph state machines and conditional routing, achieving 90% query accuracy improvement and 40% cost reduction
- Implemented comprehensive observability with LangSmith tracing, Prometheus metrics, and real-time monitoring dashboards, ensuring 99.9% system uptime

AI-Text Discriminator | https://github.com/Venkat-Git98/AI-Text-Discriminator

- Pioneered advanced parameter-efficient fine-tuning techniques (PEFT) with Low-Rank Adaptation, reducing trainable parameters by 90% while maintaining state-of-the-art classification performance
- Built scalable NLP pipeline using LoRA-based fine-tuning of transformer models on 1.2M+ text dataset, achieving 97% accuracy with 3x faster training and 50% reduced memory usage